



INSTITUTO TECNOLÓGICO
SUPERIOR DE MISANTLA.

“DISCRIMINACIÓN DE PROVEEDORES CON BASE
EN SU HISTORIAL DE ABASTECIMIENTO
MEDIANTE LA SELECCIÓN COMBINADA DE
CLASIFICADORES E HIPERPARAMETRIZACIÓN”

TESIS

PARA OBTENER EL GRADO DE MAESTRA
EN SISTEMAS COMPUTACIONALES.

P R E S E N T A

L.I. DANIELA LIZBETH HERNÁNDEZ SÁNCHEZ

DIRECTOR

DR. EDDY SÁNCHEZ DE LA CRUZ.

Autorización de impresión



INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA

AUTORIZACIÓN DE IMPRESIÓN DE TRABAJO DE TITULACIÓN MAESTRÍA

13 DE NOVIEMBRE DEL 2023

ASUNTO: AUTORIZACIÓN DE IMPRESIÓN:
TESIS PROFESIONAL

A QUIEN CORRESPONDA:

Por medio de la presente se hace constar que el (la) C:

DANIELA LIZBETH HERNÁNDEZ SÁNCHEZ

Estudiante de la MAESTRIA EN SISTEMAS COMPUTACIONALES con No. de Control 122T0554 ha cumplido satisfactoriamente con lo estipulado por el Lineamiento de Posgrado para la obtención del grado de Maestría mediante **TESIS PROFESIONAL**

Por tal motivo se Autoriza la impresión del Tema titulado:

"DISCRIMINACIÓN DE PROVEEDORES CON BASE EN SU HISTORIAL DE ABASTECIMIENTO MEDIANTE LA SELECCIÓN COMBINADA DE CLASIFICADORES E HIPERPARAMETRIZACIÓN."

Dándose un plazo no mayor de un mes de la expedición de la presente a la solicitud del examen para la obtención del grado de maestría.

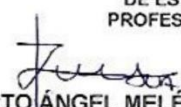
ATENTAMENTE


DR. EDDY SANCHEZ DE LA CRUZ
Presidente

COORDINACIÓN
DE ESTUDIOS
PROFESIONALES


DR. IRAHAN OTONIEL JOSÉ GUZMÁN

Secretario


DR. ROBERTO ÁNGEL MELÉNDEZ ARMENTA

Vocal

Archivo.

VER. 01/03/09

F-SA-39

Agradecimientos

Aprovecho este espacio para agradecer a las entidades y personas que directa o indirectamente contribuyeron en la realización del presente trabajo.

Agradezco al Consejo Nacional de Ciencia y Tecnología (Conacyt) el cual impulsó esta investigación mediante el incentivo económico brindado durante los últimos meses de realización del programa de posgrado.

Agradezco al Tecnológico Nacional de México Campus Misantla por brindarme la oportunidad de combinar mis actividades laborales con programas de superación profesional.

Externo con orgullo mi aprecio y máxima gratitud al Dr. Eddy Sánchez de la Cruz, científico y escritor mexicano, quien confió en mi persona para el desarrollo de este proyecto; por brindarme la fuente de datos necesaria para llevar a cabo la investigación, por su atinada dirección y guía en el proceso, por el apoyo y la motivación brindada en momentos de crisis, por sus valiosos consejos a nivel personal y académico, por su confianza y gran calidad humana, gracias.

Gracias a los sinodales del jurado, el Dr. Irahan Otoniel José Guzmán y al Dr. Roberto Ángel Meléndez Armenta, por ser partícipes de la revisión del escrito y emitir su juicio a favor de acuerdo con lo estipulado en los lineamientos de posgrado.

Dedicatoria

A mi familia.

A todos los que fueron un apoyo insustituible e imprescindible en el transcurso de mi formación personal y académica.

En especial a mis hijas, Daniella, Ella, Hellen, quienes forman parte de la fuente de inspiración en cada proyecto planteado.

Resumen

En el presente trabajo se aborda una de las problemáticas que concierne a la cadena de suministros, específicamente dentro del eslabón de proveeduría en el proceso de selección de proveedores que se presenta en la exportadora de Lima persa “Mónica, S.A. de C.V”, de Martínez de La Torre, Veracruz, donde radica el interés que da origen a este trabajo de investigación y se refiere a la posibilidad de detectar patrones que permitan identificar al mejor proveedor o proveedores bajo un modelo que brinde un alto grado de confianza y permita abastecer un pedido que cumpla con los requerimientos solicitados. El estudio se define como un proyecto de investigación y evaluación a nivel exploratorio dado que analiza información secundaria disponible, el proyecto se aplica a datos históricos de proveeduría en la empacadora Mónica de Martínez de la Torre, Veracruz, donde el conjunto de datos incorpora la recolección de información registrada durante el periodo 2014 a 2018, se implementa por una parte, un diseño de investigación experimental pura, donde para este caso se aplican técnicas de minería de datos que permiten controlar dos grupos de datos (entrenamiento y prueba) y por otra parte, se establece la investigación aplicada la cual se orienta a demostrar la validez de la técnica de *selección combinada de clasificadores* mediante el método de *hiperparametrización* en el proceso de discriminación de proveedores para la empacadora en mención donde los resultados generados de la investigación permiten ser utilizados como apoyo a los decisores para evaluar y seleccionar la mejor de sus alternativas.

El principal instrumento del presente trabajo se centra en una base de datos sobre producción citrícola derivada de estudios anteriores la cual se analiza y se separa en conjuntos de datos clasificados por temporalidad y zona de producción. La población que forma parte de este estudio se compone de los datos históricos sobre los proveedores que atiende las necesidades de requerimiento de dicha empacadora, la muestra se construye a partir de la técnica implementada al conjunto de datos (*cross-validation*, *percentage split* y muestra representativa) y el muestreo se realiza atendiendo la definición propia de cada técnica.

Las técnicas de minería de datos se aplican en la plataforma de software para el aprendizaje automático *WEKA* donde se define el origen de cada conjunto de datos para implementar el clasificador : *selección combinada de clasificadores e hiperparametrización*, por último, se verifican las técnicas de clasificación mediante los indicadores de rendimiento, en el cual se analizan los datos del clasificador y la matriz de confusión obtenida para cada conjunto de datos según el modo de entrenamiento elegido.

Los resultados indican, el mejor clasificador encontrado, el número de instancias correctamente e incorrectamente clasificadas y el porcentaje de exactitud del clasificador según la matriz de confusión resultante, asimismo para cada conjunto de datos se obtienen los indicadores para la técnica con mayor rendimiento, conduciendo así a la posibilidad de discriminar proveedores de lima persa con base en su historial de abastecimiento mediante la *selección combinada de clasificadores e hiperparametrización* para la exportadora de Lima persa “Mónica, S.A. de C.V”, de Martínez de La Torre, Veracruz, lo cual representa una herramienta útil centrada en el aprendizaje automático que proporciona soporte a la toma de decisiones bajo un alto grado de confianza para el decisor.

Abstract

This paper addresses one of the problems of interest within the supply chain, specifically within the supply link in the supplier selection process presented at the Persian Lima exporter “Mónica, S.A. de C.V”, of Martínez de La Torre, Veracruz, where lies the interest that gives rise to this research work and refers to the possibility of detecting patterns that allow identifying the best supplier or suppliers under a model that provides a high degree of confidence and allows to supply an order that complies with the request.

The study is defined as an exploratory level research and evaluation project as it analyses available secondary information, the project applies to historical supply data at the Mónica de Martínez de la Torre baler, Veracruz, where the dataset contains information acquired from the year 2014 to 2018, is implemented, on the one hand, a pure experimental research design, where data mining techniques are implemented to classify two sets of data (training and testing), and otherwise, established the applied research which is oriented to demonstrate the validity of the technique of *combined selection of classifiers* through the method of *hyperparameter optimization* in the discrimination process of suppliers for the baler in mention where the results generated from the research allow to be used as support to decision-makers to evaluate and select the best of their alternatives. The main instrument of this work focuses on a database on citrus production derived from previous studies. The database is analyzed and separated into datasets classified by time and production area. The population of this study is composed of historical data on suppliers that meet the requirements of such baler. The sample is constructed from the technique implemented in the dataset (cross-validation, percentage split, and representative sample).

Data mining techniques are applied in the WEKA machine learning software platform where the source of each dataset is defined to implement the classifier: *combined selection of classifiers and hyperparameter optimization*. Finally, classification techniques are verified using performance indicators, in which the data of the classifier and the confusion matrix obtained for each dataset are analyzed according to the chosen training mode.

The results indicate the best classifier found, the number of instances correctly and incorrectly classified, and the accuracy percentage of the classifier according to the resulting confusion matrix. Also, for each dataset, the indicators for the most efficient technique are obtained, thus leading to the possibility of discriminating Persian lime suppliers based on their supply history by the *combined selection of classifiers and hyperparameter optimization* for the Persian Lima exporter “Mónica, S.A. de C.V”, of Martínez de La Torre, Veracruz, which represents a useful tool focused on machine learning providing support for decision making under a high degree of confidence for the decision-maker.

Índice general

Autorización de impresión	2
Agradecimientos	3
Dedicatoria	4
Resumen	5
Abstract	7
Índice general.....	9
Índice de imágenes	12
Índice de tablas	13
Índice de ecuaciones	14
Índice de anexos	15
Introducción.....	16
1 Generalidades.....	17
1.1 Descripción del problema.....	17
1.2 Justificación.....	18
1.3 Objetivos.....	20
1.3.1 Objetivo general.....	20
1.3.2 Objetivos específicos.....	20
1.4 Hipótesis	20
1.5 Alcances y limitaciones	21
1.5.1 Alcances	21
1.5.2 Limitaciones	21
1.6 Propuesta de solución.....	22
1.7 Estructura de la tesis.....	22
2 Marco Teórico.....	24
2.1 Aprendizaje automático (Machine learning)	24
2.1.1 Aprendizaje supervisado	27
2.1.2 Aprendizaje no supervisado	29
2.1.3 Aprendizaje mediante refuerzo	30
2.1.4 Selección de clasificadores e hiperparametrización	31
2.2 Preprocesamiento de datos	36
2.2.1 Integración y recopilación	37
2.2.2 Limpieza y transformación	37

2.2.3	Data mining	38
2.2.4	Interpretación y evaluación	39
2.3	<i>Métricas de validación (Indicadores del rendimiento del clasificador)</i>	39
2.3.1	Matriz de confusión	41
2.3.2	Espacio ROC.....	43
2.4	<i>Técnicas de estimación</i>	44
2.4.1	Error aparente	45
2.4.2	Entrenamiento y prueba o holdout.....	45
2.4.3	Validación cruzada	46
2.4.4	Bootstrap.....	47
2.5	<i>Descripción de la empresa</i>	47
2.6	<i>Estado del arte</i>	48
2.6.1	Resultados de la revisión literaria	49
2.6.1.1	Investigaciones previas –Algoritmos de clasificación e hiperparametrización:.....	50
2.6.1.2	Investigaciones previas – Evaluación y selección de proveedores basada en técnicas de IA.....	59
3	Metodología de la Investigación	64
3.1	<i>Desarrollo metodológico</i>	64
3.1.1	Diseño de investigación.....	64
3.1.2	Definición de las variables.....	65
3.1.3	Población y muestra.....	65
3.1.4	Instrumentos	65
3.1.5	Procedimientos	66
3.1.6	Análisis de datos.....	69
3.2	<i>Metodología para la selección de proveedores</i>	69
3.2.1	Definición del problema	70
3.2.2	Descripción del método	70
3.2.3	Identificación de atributos	71
3.2.4	Evaluación del proveedor mediante una técnica	73
4	Experimentos y Resultados	74
4.1	<i>Análisis e interpretación de los resultados</i>	74
4.2	<i>Análisis de los indicadores de rendimiento del clasificador</i>	75
4.2.1	Resultados de clasificación bajo el criterio de muestreo: Cross-validation (10 folds)	75
4.2.2	Resultados de clasificación bajo el criterio de muestreo: Percentage split(66%).....	75
4.2.3	Resultados de clasificación bajo el criterio de muestreo: Muestra representativa.	76
4.2.4	Comparación de resultados de clasificación para cada conjunto de datos bajo los criterios de muestreo establecidos en el estudio.	78
4.3	<i>Análisis de la matriz de confusión y del resumen detallado por clase</i>	79
4.4	<i>Análisis de las métricas de validación aplicadas</i>	88
4.4.1	Resultados obtenidos mediante el desarrollo de fórmulas para el cálculo de: <i>overall accuracy</i> , precisión, sensibilidad, especificidad y <i>FP Rate</i>	88
4.4.2	Resultados generados en weka para las principales métricas de desempeño en clasificadores: <i>overall accuracy</i> , precisión, sensibilidad, <i>FP Rate</i> y curva roc.	89
4.5	<i>Análisis del espacio ROC</i>	90
4.5.1	Gráfico ROC para el conjunto de datos: <i>alta_norte</i>	92
4.5.2	Gráfico ROC para el conjunto de datos: <i>alta_centro</i>	92

4.5.3	Gráfico ROC para el conjunto de datos: alta_sur	93
4.5.4	Gráfico ROC para el conjunto de datos: baja_norte	94
4.5.5	Gráfico ROC para el conjunto de datos: baja_centro.....	95
4.5.6	Gráfico ROC para el conjunto de datos: baja_sur	95
4.6	<i>Indicadores utilizados para la discriminación de clases (proveedores).....</i>	96
4.7	<i>Indicadores de confiabilidad para el total de las clases.....</i>	98
4.8	<i>Comparación de resultados con el estado del arte</i>	101
5	Conclusión y Trabajo Futuro.....	104
5.1	<i>Conclusión.....</i>	104
5.2	<i>Trabajo futuro.....</i>	106
	Referencias	108
	Anexos	110
	<i>Experimentos realizados en weka implementando meta.AutoWEKAClassifier.....</i>	110
	Anexo 1 Conjunto de datos:alta_norte bajo el criterio de muestreo: cross-validation (10 folds).	110
	Anexo 2 Conjunto de datos: alta_centro bajo el criterio de muestreo: <i>cross-validation</i> (10 folds).....	110
	Anexo 3 Conjunto de datos: alta_sur bajo el criterio de muestreo: <i>cross-validation</i> (10 folds).	112
	Anexo 4 Conjunto de datos: baja_norte bajo el criterio de muestreo: cross-validation (10 folds).	112
	Anexo 5 Conjunto de datos: baja_centro bajo el criterio de muestreo: cross-validation (10 folds).....	113
	Anexo 6 Conjunto de datos: baja_sur bajo el criterio de muestreo: cross-validation (10 folds).	114
	Anexo 7 Conjunto de datos: alta_norte bajo el criterio de muestreo: percentage_spit (66%).	115
	Anexo 8 Conjunto de datos: alta_centro bajo el criterio de muestreo: percentage_spit (66%).	115
	Anexo 9 Conjunto de datos: alta_sur bajo el criterio de muestreo: percentage_spit (66%).	117
	Anexo 10 Conjunto de datos: baja_norte bajo el criterio de muestreo: percentage_spit (66%).	118
	Anexo 11 Conjunto de datos: baja_centro bajo el criterio de muestreo: percentage_spit (66%).	118
	Anexo 12 Conjunto de datos: baja_sur el criterio de muestreo: percentage_spit (66%).	120
	Anexo 13 Conjunto de datos: alta_norte bajo el criterio de muestreo: muestra representativa.	120
	Anexo 14 Conjunto de datos: alta_centro bajo el criterio de muestreo: muestra representativa.	121
	Anexo 15 Conjunto de datos: alta_sur bajo el criterio de muestreo: muestra representativa.....	122
	Anexo 16 Conjunto de datos: baja_norte bajo el criterio de muestreo: muestra representativa	123
	Anexo 17 Conjunto de datos: baja_centro bajo el criterio de muestreo: muestra representativa	123
	Anexo 18 Conjunto de datos: baja_sur bajo el criterio de muestreo: muestra representativa.....	125
	<i>Cálculo de las métricas de desempeño en clasificadores mediante el desarrollo de fórmulas</i>	125
	Anexo 19 Conjunto de datos: alta _norte bajo el criterio de muestreo con mayor rendimiento: muestra representativa	125
	Anexo 20 Conjunto de datos: alta _centro bajo el criterio de muestreo con mayor rendimiento: <i>cross-validation</i>	126
	Anexo 21 Conjunto de datos: alta _sur bajo el criterio de muestreo con mayor rendimiento: <i>cross-validation</i>	127
	Anexo 22 Conjunto de datos: baja_norte bajo el criterio de muestreo con mayor rendimiento: <i>percentage-split</i>	127
	Anexo 23 Conjunto de datos: baja_centro bajo el criterio de muestreo con mayor rendimiento: cross-validation.	128
	Anexo 24 Conjunto de datos: baja_sur bajo el criterio de muestreo con mayor rendimiento: percentage-split.....	128
	<i>Instalación de paquetes en weka de la categoría classification.....</i>	129
	Anexo 25 Vista del proceso	129

Índice de imágenes

Imagen 2.1 Organización jerárquica de las técnicas de aprendizaje automático. (Javier & Navarro, 2001a)	28
Imagen 2.2 Proceso KDD. Fases de gestión, análisis y explotación de datos (Fayyad, 1996).	37
Imagen 2.3 Representación esquemática de un clasificador (Javier & Navarro, 2001a).	40
Imagen 2.4 Área bajo la curva (AUC) del punto (0.4, 0.6) (Rodrigo, Peñas, & Verdejo, 2009)	44
Imagen 4.1 Gráfico ROC para la clase: proveedor 2.....	92
Imagen 4.2 Gráfico ROC para la clase: proveedor 7.....	93
Imagen 4.3 Gráfico ROC para la clase: proveedor 10.....	93
Imagen 4.4 Gráfico ROC para la clase: proveedor 4.....	94
Imagen 4.5 Gráfico ROC para la clase: proveedor 2.....	94
Imagen 4.6 Gráfico ROC para la clase: proveedor 7.....	95
Imagen 4.7 Gráfico ROC para la clase: proveedor 4.....	96

Índice de tablas

Tabla 2.1 Matriz de confusión de un problema de dos clases (Abreu, 2015).	41
Tabla 2.2 Medidas de evaluación de Clasificadores. (Duval, Vega, & Ruiz, n.d.)	42
Tabla 3.1 Valores frecuentemente utilizados de la distribución normal para Z.	67
Tabla 3.2 Tabla de apoyo para el cálculo del tamaño de la muestra con nivel de confianza = 95%.	68
Tabla 3.3 Valores de la población y muestra obtenida para cada conjunto de datos.	68
Tabla 3.4 Operacionalización de las variables del caso de estudio.	72
Tabla 4.1 Mejor algoritmo clasificador y porcentaje de exactitud para cada conjunto de datos.	75
Tabla 4.2 Mejor algoritmo clasificador y porcentaje de exactitud para cada conjunto de datos.	76
Tabla 4.3 Mejor algoritmo clasificador y porcentaje de exactitud para cada conjunto de datos.	76
Tabla 4.4 Cálculo del tamaño de la muestra para cada conjunto de datos.	77
Tabla 4.5 Comparación de resultados mediante diferentes criterios de muestreo.	78
Tabla 4.6 Mejor algoritmo clasificador y porcentaje de exactitud para el criterio de muestreo con mayor rendimiento sobre cada conjunto de datos.	79
Tabla 4.7 Matriz de confusión para el criterio de muestreo con mayor rendimiento: muestra representativa.	80
Tabla 4.8 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:	81
Tabla 4.9 Matriz de confusión para el criterio de muestreo con mayor rendimiento: cross-validation.	82
Tabla 4.10 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:	83
Tabla 4.11 Matriz de confusión para el criterio de muestreo con mayor rendimiento:	84
Tabla 4.12 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:	84
Tabla 4.13 Matriz de confusión para el criterio de muestreo con mayor rendimiento:	85
Tabla 4.14 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:	85
Tabla 4.15 Matriz de confusión para el criterio de muestreo con mayor rendimiento:	86
Tabla 4.16 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:	87
Tabla 4.17 Matriz de confusión para el criterio de muestreo con mayor rendimiento:	87
Tabla 4.18 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:	88
Tabla 4.19 Valor de los indicadores de rendimiento bajo el desarrollo de fórmulas	89
Tabla 4.20 Valor de los indicadores de rendimiento generados en weka para cada conjunto de datos.	90
Tabla 4.21 Valores de los ejes (x,y) para construir los gráficos ROC.	91
Tabla 4.22 Valor de los indicadores con mayor rendimiento para la discriminación de clases	97
Tabla 4.23 Porcentaje de confiabilidad de los proveedores.	98
Tabla 4.24 Porcentaje de confiabilidad de los proveedores.	100
Tabla 4.25 Comparación de selección combinada de algoritmos con los resultados reportados en (Fernández, 2018)	102
Tabla 4.26 Comparación del mejor criterio de muestreo con los resultados reportados en (Fernández, 2018)	102
Tabla 4.27 Comparación de la mejor clase (proveedor) con los resultados reportados en (Fernández, 2018)	103
Tabla 4.28 Comparación de los mayores porcentajes de confiabilidad de los proveedores con los resultados reportados en (Fernández, 2018)	103

Índice de ecuaciones

Ecuación 2.1 Cálculo del error.....	40
Ecuación 2.2 Error	42
Ecuación 2.3 Exactitud o precisión.....	42
Ecuación 2.4 Sensibilidad	42
Ecuación 2.5 Razón de FP.....	42
Ecuación 2.6 Especificidad	42
Ecuación 2.7 Razón de FN	42
Ecuación 3.1 Cálculo de la muestra de población finita.....	68

Índice de anexos

Anexo 1 Conjunto de datos:alta_norte bajo el criterio de muestreo: cross-validation (10 folds).....	110
Anexo 2 Conjunto de datos: alta_centro bajo el criterio de muestreo: cross-validation (10 folds).	110
Anexo 3 Conjunto de datos: alta_sur bajo el criterio de muestreo: cross-validation (10 folds).....	112
Anexo 4 Conjunto de datos: baja_norte bajo el criterio de muestreo: cross-validation (10 folds).	112
Anexo 5 Conjunto de datos: baja_centro bajo el criterio de muestreo: cross-validation (10 folds).	113
Anexo 6 Conjunto de datos: baja_sur bajo el criterio de muestreo: cross-validation (10 folds).....	114
Anexo 7 Conjunto de datos: alta_norte bajo el criterio de muestreo: percentage_spit (66%).	115
Anexo 8 Conjunto de datos: alta_centro bajo el criterio de muestreo: percentage_spit (66%).	115
Anexo 9 Conjunto de datos: alta_sur bajo el criterio de muestreo: percentage_spit (66%).	117
Anexo 10 Conjunto de datos: baja_norte bajo el criterio de muestreo: percentage_spit (66%).	118
Anexo 11 Conjunto de datos: baja_centro bajo el criterio de muestreo: percentage_spit (66%).	118
Anexo 12 Conjunto de datos: baja_sur bajo el criterio de muestreo: percentage_spit (66%).	120
Anexo 13 Conjunto de datos: alta_norte bajo el criterio de muestreo: muestra representativa.	120
Anexo 14 Conjunto de datos: alta_centro bajo el criterio de muestreo: muestra representativa.	121
Anexo 15 Conjunto de datos: alta_sur bajo el criterio de muestreo: muestra representativa	122
Anexo 16 Conjunto de datos: baja_norte bajo el criterio de muestreo: muestra representativa.....	123
Anexo 17 Conjunto de datos: baja_centro bajo el criterio de muestreo: muestra representativa	123
Anexo 18 Conjunto de datos: baja_sur bajo el criterio de muestreo: muestra representativa	125
Anexo 19 Conjunto de datos: alta _norte bajo el criterio de muestreo con mayor rendimiento: muestra representativa.....	125
Anexo 20 Conjunto de datos: alta _centro bajo el criterio de muestreo con mayor rendimiento: cross-validation	126
Anexo 21 Conjunto de datos: alta _sur bajo el criterio de muestreo con mayor rendimiento: cross-validation.....	127
Anexo 22 Conjunto de datos: baja_norte bajo el criterio de muestreo con mayor rendimiento: percentage-split .	127
Anexo 23 Conjunto de datos: baja_centro bajo el criterio de muestreo con mayor rendimiento: cross-validation.	128
Anexo 24 Conjunto de datos: baja_sur bajo el criterio de muestreo con mayor rendimiento: percentage-split.	128
Anexo 25 Vista del proceso	129

Introducción

En la actualidad las empresas que desean mantenerse en un marco competitivo se ven obligadas a implementar estrategias en áreas que consideren claves en el desarrollo de sus funciones. Para la empackadora de limones Mónica S.A. de C.V resulta necesario optimizar las actividades de la cadena de suministro, específicamente en el eslabón de proveedores.

La motivación del presente trabajo surge del interés de atender a la problemática de la empackadora al momento de identificar en un conjunto de alternativas al mejor proveedor para atender un pedido específico cumpliendo con los parámetros establecidos por la empackadora, ya que el proceso de selección de proveedores se torna complejo al momento de realizar una evaluación multicriterio, es por ello, que resulta importante implementar técnicas computacionales que permitan convertir los datos históricos en información útil para la empackadora al momento de seleccionar a sus proveedores.

Dentro de la propuesta de solución se considera la implementación de metodologías computacionales respaldadas en técnicas de preprocesamiento de datos, aprendizaje automático (ML¹), selección combinada de algoritmos para clasificación y proceso de descubrimiento o extracción de conocimiento.

El utilizar la información como un recurso estratégico, permitirá a los decisores de la empackadora (directivos) discriminar proveedores con un alto grado de confianza, ya que actualmente en el proceso de selección no se implementa ningún modelo computacional de clasificación que lo fundamente.

¹ ML por sus siglas en Inglés. *Machine Learning*

Capítulo 1

Generalidades

1.1 Descripción del problema

Para cualquier empresa es importante lograr competitividad en el mercado con la finalidad de mantener o aumentar las utilidades, alineada con el cumplimiento de los objetivos que se establecen en su filosofía organizacional, por ejemplo: calidad y satisfacción del cliente, tal es el caso de la empacadora de limones Mónica S.A.de C.V, la cual es objeto de estudio del presente trabajo. En esta empacadora, como en cualquier otra empresa, es conocido que, para poder comercializar un producto, se debe seguir una serie de procesos o etapas que se refieren desde la preparación del producto hasta la distribución y venta del mismo, es precisamente en el eslabón de proveeduría o proveedores de la cadena de suministro de la empacadora de limones Mónica S.A. de C.V donde radica la problemática que da origen a este trabajo de investigación y se refiere a que *actualmente en no es posible detectar patrones que permitan identificar al mejor proveedor o proveedores bajo un modelo que brinde un alto grado de confianza, tomando como referencia los datos históricos registrados sobre pedidos y entregas, y que permitan abastecer un pedido que cumpla con los requerimientos solicitados.*

Actualmente los datos a los que se refiere la base histórica están contenidos en archivos con formato Microsoft Excel (libros), clasificados por temporada de abastecimiento y tablas de datos (hojas) que están constituidas por zonas de

abastecimiento, dando como resultado 6 tablas de datos, cada una con 12 atributos y con diferente número de registros, lo cual resulta manualmente complejo el seleccionar al mejor proveedor con un alto grado de confianza, ya que a simple vista los datos no poseen relevancia, sin embargo, los valores en los campos son el resultado de la selección y evaluación de atributos que los expertos de la empacadora consideran relevantes. Por lo tanto, resulta importante implementar técnicas computacionales que permitan convertir los datos históricos en información útil para la empacadora al momento de seleccionar a sus proveedores.

1.2 Justificación

Para la empacadora de limones Mónica S.A.C.V resulta de vital importancia obtener el mayor grado de satisfacción del mercado para mantenerse en un margen competitivo, por lo que es necesario optimizar las actividades de la cadena de suministro, específicamente en el eslabón de proveedores para este caso de estudio. Por tal motivo, el desarrollo de este trabajo responde directamente al interés para identificar en un conjunto de alternativas al mejor proveedor para atender un pedido específico cumpliendo con los parámetros establecidos por la empacadora. Es importante destacar el papel que juega la adecuada selección de proveedores ya que este proceso puede resultar complejo al momento de realizar una evaluación multicriterio, es decir, considerando diferentes atributos que influyen en la discriminación de proveedores a evaluar.

Lo que se expresa con anterioridad y atendiendo a la problemática que se sustenta en el documento, dan como resultado los principales aportes de la investigación, los cuales están implícitos en el aspecto metodológico de la misma. Por tal motivo, se justifica la importancia de cada una de las etapas de desarrollo del presente trabajo mediante la implementación de técnicas que abarcan desde el preprocesamiento de datos hasta su clasificación.

El aporte principal a las ciencias computacionales se centra en el desarrollo de cada una de las etapas de la investigación permitiendo generar un conjunto de *datasets* a partir de una base de datos histórica de abastecimiento, que mediante la implementación de técnicas de preprocesamiento de datos garanticen un conjunto de datos de calidad para

ser tratados mediante herramientas computacionales implementando metodologías respaldadas en técnicas de ML^2 , selección combinada de algoritmos para clasificación y proceso de descubrimiento o extracción de conocimiento para establecer correlaciones entre los atributos o descubrimiento de patrones en los datos que permitan identificar al mejor proveedor. Con lo anterior, se ofrece una metodología para la realización de una búsqueda optimizada en el proceso de selección de proveedores dentro de un conjunto de alternativas evaluadas bajo un enfoque multicriterio, argumentada en el rendimiento de los algoritmos implementados, mediante un análisis de los resultados del proceso de clasificación que permita validar la efectividad de los mismos, es decir, esto permitirá a la empackadora de limones Mónica S.A. de C.V. identificar al mejor proveedor tomando como referencia los datos históricos de abastecimiento ya que no existe la implementación de un modelo que brinde un alto grado de confianza para realizar lo anterior.

Dentro del marco empresarial cabe mencionar que el principal aporte radica en que la propuesta de solución surge de la necesidad de implementar un modelo de clasificación que permita la discriminación de proveedores en la empackadora de limones Mónica S.A de C.V. así mismo, se prevé que el resultado de la investigación pudiera sugerirse como un mecanismo dentro de la empackadora para:

- Evaluar diferentes criterios ante el proceso de selección de proveedores de lima persa.
- Mejorar el servicio en dirección de la selección adecuada de proveedores con un alto grado de confianza.
- Atender un pedido específico cumpliendo con los parámetros establecidos por la empackadora.
- Contribuir en la logística de la empackadora para lograr competitividad de la misma.
- Y de ser posible, agilizar procesos de empaque y distribución del producto.

² ML por sus siglas en Inglés. *Machine Learning*

1.3 Objetivos

1.3.1 Objetivo general

Discriminar proveedores de lima persa con base en su historial de abastecimiento mediante la selección combinada de clasificadores e hiperparametrización.

1.3.2 Objetivos específicos

- Obtener y analizar la base de datos histórica de abastecimiento con información sobre proveedores de lima persa para determinar datos faltantes o incompletos, inconsistentes y datos con ruido (pérdida de información).
- Generar un conjunto de *datasets* con base en la temporada y zona de abastecimiento a partir de la base de datos histórica y transformarlos en archivos de formato estándar para su posterior procesamiento.
- Analizar la información de los *datasets* mediante la técnica de correlación de variables para identificar atributos relevantes que permitan generar nuevos atributos o bien la eliminación de atributos.
- Implementar el método de clasificación *selección combinada de algoritmos e hiperparametrización*, para obtener con grado de confianza el mejor proveedor para un pedido determinado.
- Analizar los resultados obtenidos en la clasificación y validar la efectividad del método usando métricas de evaluación afines.

1.4 Hipótesis

Es posible discriminar proveedores de lima persa con base en su historial de abastecimiento mediante la selección combinada de clasificadores e hiperparametrización con un alto grado de confianza.

1.5 Alcances y limitaciones

1.5.1 Alcances

La implementación del método de selección combinada de clasificadores e hiperparametrización, permitirá: discriminar proveedores de lima persa requeridos en el eslabón de proveeduría dentro de la cadena de suministro de la empacadora de limones Mónica S.A. de C.V. Dicho método es aplicable a cualquier abastecedor en las regiones identificadas (norte, centro, sur) de los estados de Veracruz y Tabasco que suministran a la empacadora, permitiendo cumplir con los siguientes alcances establecidos:

- Generar un conjunto de *datasets* a partir de una base de datos histórica de abastecimiento.
- Analizar los porcentajes de rendimiento para identificar al mejor proveedor y atender una demanda específica (de lima persa) que cumpla con los parámetros de calidad requeridos por la empacadora.
- Ofrecer una metodología para la realización de una búsqueda optimizada en el proceso de selección de proveedores dentro de un conjunto de alternativas evaluadas bajo un enfoque multicriterio.

1.5.2 Limitaciones

- Dado el supuesto, que existe información adicional interna de la empresa, que no se expone por políticas de la misma o derechos de privacidad, se puede considerar como una limitante, ya que esta información pudiera en determinado momento ser contemplada para enriquecer el contenido del *dataset* original.
- Las personas que participan en la implementación del modelo pueden realizar una interpretación errónea de los resultados de la clasificación, o bien, el proceso de decisión puede ser influenciado por aspectos subjetivos que quedan fuera del alcance de este proyecto.
- La metodología de clasificación es aplicable solo abastecedores que suministran a la empacadora de limones Mónica S.A de C.V. ya que debe cumplir con los parámetros específicos establecidos por la misma.

-
- Los resultados de las fases de desarrollo del presente trabajo están diseñados para fines de clasificación, respondiendo a la problemática expuesta con anterioridad, por lo cual, la presente investigación se limita a concluir la propuesta del modelo de clasificación sin tener que realizar una aplicación o sistema informático necesario para lograr lo anterior.

1.6 Propuesta de solución

Para dar solución a la problemática que se presenta en el eslabón de proveeduría dentro de la cadena de suministro de la empacadora de limones Mónica S.A. de C.V. la cual expone el no contar con un mecanismo para optimizar el proceso de selección de proveedores bajo un enfoque multicriterio, se presenta la propuesta de solución que consiste en hacer un análisis exploratorio de la base de datos histórica a fin de identificar datos faltantes e inconsistentes, así mismo, realizar una correlación entre variables a fin de determinar criterios relevantes que permitan generar un conjunto de datos de calidad.

Derivado de la recolección y análisis de datos se implementan técnicas computacionales que permiten optimizar el proceso de selección de proveedores, logrando una clasificación tomando en cuenta los criterios definidos por los expertos con base en valores de atributos como: calidad del producto, temporada de abastecimiento, zona de abastecimiento, cantidad de abastecimiento, entre otros.

Para la empacadora en mención, la evaluación de los proveedores representa una oportunidad de competitividad en el ámbito industrial, por lo tanto, en la propuesta de solución también se considera la implementación del método de selección combinada de algoritmos e hiperparametrización generando un porcentaje de rendimiento, lo cual permite identificar la clasificación con mayor grado de confianza, dando como resultado la solución a la problemática que se presenta actualmente en la empacadora.

1.7 Estructura de la tesis

El desarrollo del trabajo de investigación, se encuentra documentado en la presente tesis bajo una estructura de cuatro capítulos, los cuales se describen de manera general en los siguientes párrafos:

Capítulo 1: En este apartado se detallan las generalidades de la tesis, haciendo mención a la problemática actual que se intenta resolver con el desarrollo de la investigación, así mismo, se exponen los objetivos específicos que conducen al logro del objetivo general mediante la implementación de técnicas y metodologías que permiten cumplir con los alcances que se mencionan en este capítulo.

Capítulo 2: Se presenta el estado del arte, el cual se integra por las referencias derivadas de la investigación documental sobre consultas a trabajos o estudios realizados que incluye información, conocimiento, técnicas, metodologías, algoritmos, etc. relacionados con el tópico de investigación a tratar y que sirven como punto de partida para orientar de manera fundamentada el proceso de la investigación propuesta en el presente trabajo.

Capítulo 3: Se describen los aspectos metodológicos implementados en la investigación, identificando procesos que abarcan desde técnicas de preprocesamiento de datos, minería de datos, selección combinada de algoritmos para clasificación hasta el proceso de descubrimiento o extracción de conocimiento.

Capítulo 4: Se muestra la fase experimental de la investigación aplicando técnicas y procesos alineados con la metodología propuesta en función de los objetivos planteados.

Capítulo 5: Por último, se presentan los principales hallazgos y conclusiones de la investigación, producto del análisis de los resultados obtenidos en la clasificación y la validación de la funcionalidad del modelo bajo los algoritmos implementados.

Capítulo 2

Marco Teórico

Dentro del desarrollo del presente trabajo se hace referencia a diversos términos, los cuales sirven como punto de partida para abordar la problemática propuesta en el capítulo uno, por lo cual se exponen definiciones de conceptos encontrados en la literatura.

2.1 Aprendizaje automático (Machine learning)

La aplicación del aprendizaje automático es de carácter multidisciplinario, sin embargo, como resultado de la investigación documental, se exponen algunas implementaciones de algoritmos de aprendizaje automático para tener una referencia del impacto que genera su aplicación en el ámbito de las industrias, destacando la importancia de los datos, ya que éstos, representan además de un recurso, una herramienta importante dentro de las empresas, tal como se expresa a continuación.

En la recogida masiva de los datos se encuentran las claves para hacer que el negocio crezca. [...] en los datos, se encuentran, en la mayoría de casos, el conocimiento implícito necesario para optimizar los procesos del negocio, el servicio al cliente y la rentabilidad, pero para obtener este conocimiento y para operarlo, no basta con guardar masivamente todos los datos que sean factibles de obtener. Las posibilidades que se abren ahora con la

Inteligencia Artificial³, es poder manejar cantidades ingentes de información para tener una primera aproximación a la descripción de los fenómenos experimentales observados. (Pavón Pérez & Vega Sánchez Sebastián Dormido Canto, n.d.)

El valor de estos datos estará en función de las relaciones que pudieran existir entre ellos y que se traducen en hechos observables o experiencias pasadas que se hayan registrado, tomando como referencia este enfoque se puede inducir que el registro de ciertas correlaciones entre variables o criterios que permitan la toma de decisiones bajo un conjunto de alternativas debería resolver de la mejor manera posible un problema en específico , entonces, partiendo de la idea que el aprendizaje se basa en experiencias pasadas para la solución de un problema o el logro de un objetivo, podemos inferir, que el aprendizaje automático puede ser visto como un proceso de aprendizaje donde el participante (en este caso una computadora) aprende de la experiencia a fin de incrementar su base de conocimientos mediante la combinación de entradas que le permitan en determinado momento predecir salidas.

Existen diferentes formas de adquirir conocimiento, en el ámbito computacional, el ML⁴, hace referencia a dos tipos de aprendizaje: supervisado y no supervisado. La definición y objetivo del ML se fundamentan en las aportaciones que se enuncian a continuación:

Se dice que es un sistema que aprende de forma automatizada (o aprendiz) en un artefacto (o conjunto de algoritmos) que, para resolver problemas, toma decisiones basadas en la experiencia acumulada -en los casos resueltos anteriormente- para mejorar su actuación. Estos sistemas deben ser capaces de trabajar con un rango muy amplio de datos de entrada, que pueden incluir datos incompletos, inciertos, ruido, inconsistencias, etc.(Javier & Navarro, 2001b)

³ El término Inteligencia artificial fue sugerido como la técnica para simular la inteligencia humana.

⁴ ML por sus siglas en Inglés. *Machine Learning*

El aprendizaje automático tiene como objetivo generar expresiones clasificatorias lo suficientemente simples como para ser entendidas fácilmente por el humano (Michie, Spiegelhalter, & Taylor, 1994). Cada problema de aprendizaje automático o aprendizaje máquina puede ser definido como el problema de mejorar alguna medida de rendimiento P al ejecutar alguna tarea T , a través de algún tipo de experiencia de entrenamiento E . (Mitchell, 2017)

Según el tipo de selección y adaptación (transformación) que un sistema realiza sobre la información disponible es posible identificar varios paradigmas del ML⁵, (Moreno Ribas, 1994), tales como se enuncian a continuación:

- Aprendizaje deductivo
- Aprendizaje analítico
- Aprendizaje analógico
- Aprendizaje inductivo
- Aprendizaje mediante descubrimiento
- Algoritmos genéticos
- Conexionismo

Según los diferentes tipos de algoritmos utilizados para llevar a cabo la tarea de aprendizaje, se establece otra clasificación en la que se distinguen los siguientes paradigmas de ML: (Javier & Navarro, 2001b).

- Redes neuronales
- Aprendizaje basado en casos
- Algoritmos genéticos
- Inducción de reglas
- Aprendizaje analítico

⁵ ML por sus siglas en Inglés. *Machine Learning*

Otra posible clasificación de los métodos de aprendizaje explorados en IA⁶, según (Moreno Ribas, 1994) considerando el *tipo de estrategia y las ayudas* que recibe un sistema de aprendizaje, es:

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje mediante refuerzo

2.1.1 Aprendizaje supervisado

La suposición fundamental de este tipo de método es que los ejemplos proporcionados como entrada son necesarios para cumplir las metas del aprendizaje (Moreno Ribas, 1994). El aprendizaje supervisado persigue la inducción de descripciones generales a partir de casos particulares de un problema, cuya solución se incluye como elemento en el aprendizaje. Un tipo de problemas en los que se aplica el aprendizaje supervisado, es el aprendizaje de conceptos, donde el objetivo consiste en obtener un mecanismo clasificador que permita generalizar e indicar si un nuevo caso, no utilizado en el proceso de inducción pertenece o no al concepto o clase aprendido. Los casos utilizados en el proceso de inducción en muchos problemas vienen expresados como muestras o tuplas que incluyen medidas o los atributos obtenidos del caso en estudio además de la solución o clase a la que pertenece. (Javier & Navarro, 2001a)

En la imagen 2.1 se representan mediante una organización jerárquica las técnicas de aprendizaje automático, mencionadas con anterioridad.

⁶ IA. Acrónimo de Inteligencia Artificial.

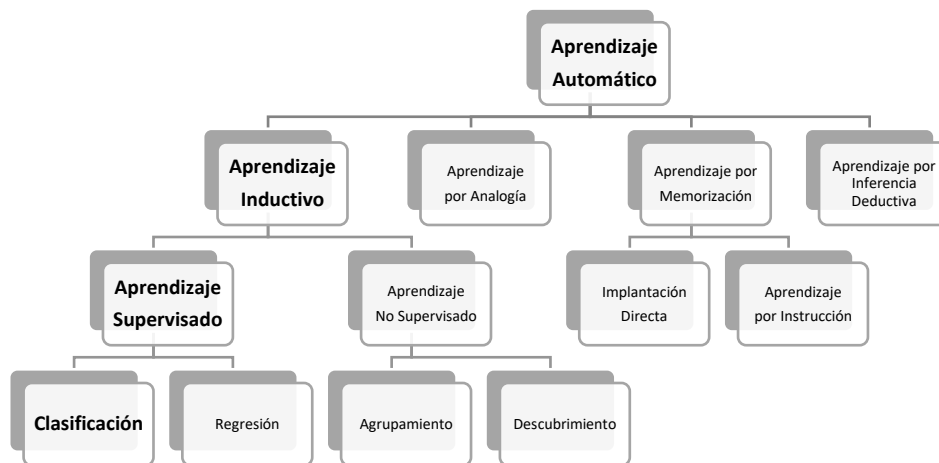


Imagen 2.1 Organización jerárquica de las técnicas de aprendizaje automático. (Javier & Navarro, 2001a)

A continuación se exponen la definiciones de algunos conceptos tratados en el contexto de aprendizaje supervisado, tomando como referencia las aportaciones de (Kohavi & Provost, 1998).

Atributo: También denominado característica, es la descripción de alguna medida de una muestra. Los atributos tienen un dominio definido por el tipo de atributo, determinando dicho dominio los valores que puede tomar el atributo.

Muestra: También llamado ejemplo o caso, es una lista de valores de atributos que se corresponde con la definición de las entidades tratadas en el problema de aprendizaje automático en estudio.

Etiqueta: Atributo (nominal o continuo) que indica la clase a la que pertenece la muestra.

Conjunto de aprendizaje: En el marco del aprendizaje supervisado, es necesario disponer de un conjunto de muestras etiquetadas, donde en general no se supone ninguna ordenación entre las mismas, el conjunto de aprendizaje es dicho conjunto a partir del cual se llevará a cabo el proceso de aprendizaje. El aprendizaje supervisado se parte del conjunto D , y se pretende obtener un mecanismo que realice la clasificación correcta de las muestras contenidas en dicho conjunto, así como de nuevas muestras no utilizadas en el proceso de aprendizaje y para las cuales no se conoce su clase (Javier & Navarro, 2001b)

Clasificador: Función que asigna a una muestra no etiquetada una etiqueta. Todos los clasificadores poseen una estructura de datos interna para realizar la asignación de una etiqueta a un ejemplo.

Algoritmo de inducción: Tiene como objetivo la construcción de un clasificador a partir de un conjunto de aprendizaje.

Un clasificador C asocia a una muestra no etiquetada x un grado de pertenencia a una clase $y \in Y$, mientras que un algoritmo de inducción I , efectúa la correspondencia de un conjunto de aprendizaje D en un clasificador C . Para indicar la etiqueta que asigna el clasificador C que ha sido obtenido a partir del conjunto de aprendizaje D a la muestra x , se formaliza la expresión $I(D, x)$.

Otro aspecto importante que caracteriza a los sistemas de ML⁷ es cómo el conocimiento adquirido, mejora el rendimiento de ese sistema en un determinado entorno. La respuesta a la anterior cuestión equivale a determinar qué conocimiento puede adquirir dicho sistema. El conocimiento que puede adquirir un sistema de ML se puede englobar en los siguientes grandes grupos: (Javier & Navarro, 2001b).

- Aprendizaje de conceptos
- Clasificación
- Resolución de problemas
- Descripción de conceptos

2.1.2 Aprendizaje no supervisado

En el aprendizaje inductivo no supervisado los casos utilizados en el proceso de aprendizaje no se encuentran etiquetados como pertenecientes a un concepto concreto. El objetivo en este tipo de sistemas es la obtención de descripciones que especifiquen las propiedades comunes a los objetivos que pertenecen a una misma clase, esto se puede conseguir mediante la formulación de una teoría que caracterice un conjunto de

⁷ ML por sus siglas en Inglés. *Machine Learning*

entidades, el descubrimiento de patrones, la regularidad en los casos presentados al sistema o la realización de una descripción taxonómica de un conjunto de entidades. (Javier & Navarro, 2001b). A continuación se enuncian los tipos de aprendizaje no supervisado:

- Agrupamiento conceptual
- Aprendizaje por refuerzo

2.1.3 Aprendizaje mediante refuerzo

Este tipo de aprendizaje se realiza mediante la asignación de recompensa o castigo sin necesidad de especificar de qué manera se tiene que llevar a cabo la tarea a aprender. La existencia de esta recompensa en el aprendizaje por refuerzo implica que algunos autores no lo consideren como aprendizaje no supervisado sino como un tipo de aprendizaje que se encuentra entre el supervisado y no supervisado (Javier & Navarro, 2001b). Dentro del aprendizaje mediante refuerzo podemos encontrar las siguientes categorías:

- Aprendizaje asociativo
- Aprendizaje no asociativo

Para el cumplimiento de los objetivos del presente documento, el enfoque del desarrollo de este trabajo de investigación se centra en el aprendizaje inductivo supervisado.

Aprendizaje inductivo: El objetivo de la inferencia inductiva o aprendizaje inductivo es la formulación de hipótesis generales plausibles que expliquen unos hechos conocidos y que además sea capaz de predecir nuevos hechos. La diferencia con la deducción se encuentra que en lugar de partir de axiomas generales, se parte de hechos concretos para llegar a los axiomas generales. (Michalsky,1983), menciona que desde el punto de vista computacional, el objetivo de la inducción es encontrar un conjunto de

reglas, considerando reglas en sentido general, no solo lógicas que puedan explicar unos hechos. (Javier & Navarro, 2001b)

Clasificación: Es una generalización del aprendizaje de conceptos, donde se tiene un número arbitrario de conceptos y por tanto la función obtenida en el proceso de aprendizaje realizará la asignación de un elemento del dominio a uno o más conceptos. Si el problema es mutuamente exclusivo, es decir que un caso no puede pertenecer a más de un concepto, el problema de clasificación se puede convertir en un número de problemas de aprendizaje de conceptos igual al número de conceptos en los que se clasifican los casos. (Javier & Navarro, 2001b)

Lista de técnicas de ML⁸

- Redes neuronales artificiales
- Bayes
- Árboles de decisión
- Reglas de asociación
- Lógica difusa
- Clasificadores ensamblados
- Hiperparametrización

Este último, por ser el de nuestro interés y el aplicado en esta investigación, es el que a continuación se describe.

2.1.4 Selección de clasificadores e hiperparametrización

El método de selección combinada de algoritmos e hiperparametrización propuesto por (Thornton, Hutter, Hoos, & Leyton-Brown, 2012a) es considerado como un enfoque completamente automatizado para abordar problemas de selección simultánea de algoritmos de aprendizaje que involucren la configuración o ajuste de diversos parámetros. Los autores consideran una amplia gama de técnicas de selección de características, combinando 3 métodos de búsqueda, 8 métodos de evaluación y todos

⁸ ML por sus siglas en Inglés. *Machine Learning*

los enfoques de clasificación implementados en el paquete de software libre o de código abierto WEKA⁹, abarcando 2 métodos ensamblados, 10 meta-métodos y 27 clasificadores base, contemplando la configuración o ajuste para cada clasificador. Como consecuencia de algunas pruebas aplicadas a repositorios de datos en línea demostraron que el desempeño de la clasificación es mucho mejor que si se usara el método de selección convencional. Como contribución de los autores cabe destacar que este enfoque permitirá a usuarios no tan expertos en materia de ML a identificar más eficientemente algoritmos y la configuración o ajuste de parámetros apropiados a las aplicaciones o casos de estudio en particular y por lo tanto mejorar el desempeño de los mismos. El paquete de software de código abierto WEKA, permite a los usuarios hacer dos tipos de elecciones: en primera instancia, permitiendo seleccionar un algoritmo de aprendizaje y en segunda personalizando el algoritmo mediante la configuración o ajuste de sus parámetros.

El hecho de dar libertad al usuario al momento de elegir los algoritmos bajo sus propios criterios implica que esta elección no sea precisamente la más adecuada, esto supone un reto para el ML¹⁰ al momento de elegir un algoritmo de aprendizaje de manera automática y simultánea configurando los parámetros que permitan mejorar su desempeño. La práctica de este enfoque demuestra la importancia de su aplicación, sin embargo, no existe gran evidencia de que este enfoque se haya considerado en la literatura mediante investigaciones anteriores. Se afirma que se han abordado de forma separada el problema de selección y la optimización de hiperparámetros, por lo cual los autores suponen que el método propuesto puede ser visto como un problema de optimización jerárquica de hiperparámetros, donde también se afirma que métodos recientes de optimización Bayesiana puede generar resultados de alta calidad en un tiempo razonable con el mínimo esfuerzo humano. Con esta propuesta se define la solución a una problemática abarcando el rango completo de clasificadores y selector de características en el paquete de software de código abierto WEKA, mediante la búsqueda de una combinación de algoritmos e hiperparametrización se demuestra que el

⁹ WEKA por sus siglas en Inglés. *Waikato Environment for Knowledge Analysis*, en español «entorno para análisis del conocimiento de la Universidad de Waikato»

¹⁰ ML por sus siglas en Inglés. *Machine Learning*

rendimiento del modelo es mejor que el desempeño en la selección estándar de algoritmos.

El método de optimización de parámetros propuesto, es un modelo que se enfoca en un problema de clasificación, aprendiendo una función $f: X \leftrightarrow Y$, donde un algoritmo de aprendizaje A mapea un conjunto de puntos $\{d_1 \dots d_n\}$ como datos de entrenamiento, $d_i = (x_i, y_i) \in X \times Y$ a tal función, la cual es expresada comúnmente mediante un vector de parámetros del modelo. En la mayoría de los algoritmos de aprendizaje se exponen los hiperparámetros $\lambda \in \Lambda$, lo cual cambia la forma en que el algoritmo trabaja por sí solo, donde los hiperparámetros son usados para definir, por ejemplo, el número de neuronas en una capa oculta, el número de datos que debe contener una hoja en un árbol de decisión, etc, estos hiperparámetros son normalmente optimizados mediante un proceso externo que evalúa el desempeño de cada configuración de hiperparámetros usando la validación cruzada (*cross-validation*).

Básicamente el modelo propuesto por (Thornton et al., 2012a) se enfoca en los siguientes puntos los cuales se describen a continuación:

1. Selección del modelo.
2. Optimización de hiperparámetros.
3. Selección combinada de algoritmos y optimización de hiperparámetros.
 - a. Configuración de secuencia de algoritmos.
 - b. Estimador *parzen* (árbol estructurado).

- **Selección del modelo:** Dado un conjunto de algoritmos de aprendizaje A y una cantidad de datos de entrenamiento $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ el objetivo de la selección del modelo consiste en determinar el algoritmo $A^* \in \mathcal{A}$ con desempeño óptimo. El desempeño se estima dividiendo D en conjuntos de datos separados de entrenamiento $D_{train}^{(i)}$ y validación $D_{valid}^{(i)}$, aprendiendo funciones f_i mediante la aplicación de A^* a $D_{train}^{(i)}$ y evaluando el rendimiento predictivo de dichas funciones sobre $D_{valid}^{(i)}$. Se usa la validación cruzada en *K-fold*, la cual divide los datos de

entrenamiento en k particiones de igual tamaño $D_{valid}^{(1)}, \dots, D_{valid}^{(k)}$, y conjuntos $D_{train}^{(i)} = D \setminus D_{valid}^{(i)}$ for $i = 1, \dots, k$.

- **Optimización de hiperparámetros:** Comúnmente los hiperparámetros son continuos lo que genera espacios de alta dimensión, para lo cual, en esta fase se puede explotar la estructura de correlación entre diferentes configuraciones de hiperparámetros $\lambda_1, \lambda_2 \in \Lambda$. Dados n hiperparámetros $\lambda_1, \dots, \lambda_n$ con dominios $\Lambda_1, \dots, \Lambda_n$ el espacio de hiperparámetros Λ es un subconjunto del producto cruzado de esos dominios $\Lambda \subset \Lambda_1 \times \dots \times \Lambda_n$, donde un hiperparámetro Λ_1 es condicional sobre otro hiperparámetro Λ_j , Λ_i está activo solo si el hiperparámetro Λ_j toma valores de un conjunto dado $V_1(j) \subsetneq \Lambda_j$, en este caso se denomina a Λ_j un padre de Λ_i . Los hiperparámetros condicionales pueden ser padres de otros hiperparámetros condicionales generando un espacio de árbol estructurado.
- **Selección combinada de algoritmos y optimización de hiperparámetros:** Dado un conjunto de algoritmos $\mathcal{A} = \{A^{(1)}, \dots, A^{(k)}\}$ con espacios de hiperparámetros asociados $\Lambda^{(1)}, \dots, \Lambda^{(k)}$, se define una selección combinada de algoritmos. El problema puede ser abordado de diferentes formas, un enfoque es la optimización Bayesiana, en particular el modelo secuencial basado en optimización (SMBO¹¹), visto como un marco de optimización de trabajo con hiperparámetros categóricos y continuos, lo cual permite explotar la estructura jerárquica proveniente de parámetros condicionales. El SMBO es representado en el algoritmo 1.

¹¹ SMBO por sus siglas en Inglés. *Sequential Model-Based Optimization*,

```

1: initialise model  $M_L$ ;  $H \leftarrow \emptyset$ 
2: while time budget for optimization has not been exhausted do
3:  $\lambda \leftarrow$  candidate configuration from  $M_L$ 
4: Compute  $c = L(A_\lambda), D_{train}^{(i)}, D_{valid}^{(i)}$ 
5:  $H \leftarrow H \cup \{(\lambda, c)\}$ 
6: Update  $M_L$  given  $H$ 
7: end while
8: return  $\lambda$  from  $H$  with minimal  $c$ 

```

Algoritmo 1 Modelo secuencial basado en optimización (SMBO) (Thornton et al., 2012)

Línea 1: construye un modelo M_L que captura las dependencias de la función de pérdida L sobre la configuración de hiperparámetros λ .

Línea 3: usa M_L para determinar la configuración de hiperparámetros λ .

Línea 4: evalúa la pérdida c de λ .

Línea 5: actualiza el modelo M_L con el nuevo dato (λ, c) .

Para seleccionar la próxima combinación de hiperparámetros λ usando el modelo M_L se utiliza una función de adquisición $aM_L: \Lambda \rightarrow R$ el cual usa la distribución estimada del modelo en una configuración arbitraria de parámetros para cuantificar que tan útil sería el conocimiento acerca de λ . El SMBO simplemente maximiza esta función sobre Λ para seleccionar la configuración más útil λ para evaluar la siguiente.

Los autores usan una de las funciones más prominentes de adquisición, el mejoramiento de casos positivos esperados sobre un existente margen de error dado c_{min} . $c(\lambda)$ permite denotar la tasa de error de la configuración de hiperparámetros λ . El mejor rendimiento esperado sobre la función c_{min} está definido como:

$$I_{c_{min}}(\lambda) := \max\{(c_{min} - c(\lambda), 0)\}.$$

Donde el valor de λ es desconocido, por lo tanto, es necesario computar o calcular el resultado respecto al modelo M_L . Una gran diferencia entre los algoritmos de SMBO radica en la clase de modelo que emplean, la configuración de algoritmos basados en modelos secuenciales (SMAC) y el estimador *parzen* de árbol estructurado (TPE) son dos algoritmos SMBO

cuyos modelos pueden manejar parámetros jerárquicos y que son adecuados para la tarea de selección combinada de algoritmos y optimización de hiperparámetros.

Debido a que existen muchos algoritmos diferentes de aprendizaje automático y tomando en cuenta que al seleccionar cada conjunto de hiperparámetros de los algoritmos se genera un número extenso de posibles opciones, la presente investigación sugiere apegarse al enfoque propuesto por (Thornton et al., 2012a) para considerar sus aportaciones en el problema de selección simultánea de algoritmos de aprendizaje y configuración de hiperparámetros. Usando este enfoque integrado propuesto por los autores, les permite a los usuarios identificar más efectivamente no solo el mejor algoritmo de aprendizaje automático sino también la configuración de hiperparámetros correspondiente y el método de selección de características apropiado para tal algoritmo, permitiendo por lo tanto mejorar el desempeño en la tarea de clasificación o regresión.

2.2 Preprocesamiento de datos

El descubrimiento de conocimiento en bases de datos (*KDD*¹²) se define como “el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de datos” (Fayyad, 1996).

El *KDD* (Imagen 2.2) es un proceso complejo que incluye varias fases, además de la obtención de patrones que es el objetivo específico del *data mining*, el *KDD* contempla la preparación de los datos, la evaluación de los datos para comprobar la calidad, la evaluación de los patrones y la interpretación de los mismos.

¹² KDD por sus siglas en Inglés. *Knowledge Discovery in Databases*



Imagen 2.2 Proceso KDD. Fases de gestión, análisis y explotación de datos (Fayyad, 1996).

La imagen anterior establece la relación entre el *KDD* y la minería de datos, identificando que el *KDD* es el proceso que engloba los procesos para descubrir conocimiento útil en bases de datos mientras que la minería de datos se encarga específicamente de generar o encontrar patrones o modelos (Silva, 2015). A continuación se definen las diferentes fases que integran el proceso *KDD*.

2.2.1 Integración y recopilación

Se determinan las fuentes de información útiles. Seguido se unifican los datos dándoles un formato común, transformándolos y normalizándolos para un mejor rendimiento en los procesos de memoria, que facilita la navegación y visualización entre datos.

2.2.2 Limpieza y transformación

Se determina la calidad de los datos a analizar, en esta fase se seleccionan y preparan los datos para su posterior clasificación, ya que no todos los datos son realmente relevantes o necesarios. Esta fase puede involucrar otras tareas como generar nuevos atributos aplicando alguna función a los atributos originales, modificando los mismos atributos o discretizándolos, aplicando un análisis de componentes principales, según sea el caso, la relevancia de los atributos es importante en la obtención de buenos resultados.

2.2.3 Data mining

Es la fase más característica del proceso de *KDD*, por lo que es común utilizar este término para referirse a todo el proceso. Esta fase tiene como objetivo producir y proporcionar nuevo conocimiento a través de un modelo basado en los datos proporcionados. El modelo es una descripción de relaciones entre los datos, que puede usarse para hacer predicciones o entender mejor los datos. Para obtener el modelo, es necesario tomar las siguientes decisiones:

- Determinar la tarea de minería (Clasificación, *clustering* u otros).
- Elegir el tipo de modelo: (Evaluación de similaridad, árboles de decisión, redes neuronales).
- Elegir el algoritmo de minería. (k-nn, J48, Id3, *NBTree*, *Perceptron* Multi-cap, Redes Neuronales Autoorganizadas).

Las tareas que se pueden encontrar en la fase de Data Mining se pueden separar en dos clases: predictivas y descriptivas. En el grupo de las predictivas están la clasificación y la regresión. En las descriptivas se encuentran el agrupamiento (*clustering*), reglas de asociación, reglas de asociación secuenciales y correlaciones.

Clasificación: Cada instancia o registro de la base de datos pertenece a una clase que se le asigna a través del atributo discreto clase de instancia, el objetivo es predecir la clase de nuevas instancias de las que se desconoce la clase.

Regresión: La principal diferencia con la clasificación es que, en la regresión, se asigna un valor real a cada instancia a través del aprendizaje de una función. El objetivo es predecir el número que hay que asignar, minimizando el error entre el valor predicho y el valor real.

Agrupamiento (Clustering): Se trata de una tarea descriptiva, que consiste en obtener grupos naturales a partir de los datos. La diferencia con la clasificación y la regresión es que analiza los datos para crear la clase. Se forman grupos entre datos similares entre sí, y diferentes con los datos de los otros grupos.

Correlaciones: Buscan el grado de similitud entre diferentes valores de instancias.

Reglas de Asociación: Similar a las correlaciones, tiene como objetivo buscar relaciones no explícitas entre atributos categóricos.

2.2.4 Interpretación y evaluación

En esta fase se evalúan y analizan los patrones generados. En caso de ser necesario, puede volverse hacia atrás para realizar algún ajuste en alguna fase anterior. Los patrones obtenidos deben ser, idealmente, ajustados a los requerimientos de precisión, robustez, claridad y rapidez, establecidos para caso de uso.

Validación: Para realizar la validación de los modelos generados, las técnicas básicas son la validación simple y la validación cruzada con n pliegues, (*n-fold cross validation*).

Uso y difusión: Una vez construido y validado el modelo, solo queda usarlo, bien para que un analista experto recomiende acciones según los resultados obtenidos, o bien para aplicar diferentes conjuntos de datos al modelo.

2.3 Métricas de validación (Indicadores del rendimiento del clasificador)

El objetivo de los clasificadores es lograr que sean cada vez más eficientes en su tarea de asociar muestras no etiquetadas a muestras etiquetadas. La forma más común de medir la eficiencia de un clasificador es la tasa de error (clasificación incorrecta). Cada vez que se le presente un nuevo caso debe tomar una decisión sobre la clase correcta para este caso, por lo tanto, la tasa de error se define como el error real que cometerá el clasificador. Cuando al clasificador, que da como salida un valor discreto perteneciente al conjunto Y , se le presenta una muestra, y si el resultado no es el mismo que el esperado se considera un error.

Si todos los errores tienen igual importancia entonces el error cometido en el clasificador es el que sustenta la siguiente expresión (Ecuación 2.1 Cálculo del error). (Javier & Navarro, 2001b)

$$error_D = \frac{\text{número de errores}}{\text{número de muestras}}$$

Ecuación 2.1 Cálculo del error

El proceso de obtención del clasificador puede caracterizarse como una búsqueda en el espacio de parámetros del clasificador, como es el caso de redes neuronales, o el espacio de posibles clasificadores, como ocurre en los árboles de decisión. El objetivo de la mayoría de los algoritmos de inducción es obtener un clasificador con una tasa de error baja. Debido a que no se dispone de todas las posibles muestras que definen el problema a tratar sino de un conjunto limitado, es necesario estimar la tasa de error que cometerá el clasificador sobre las muestras no utilizadas en el proceso de aprendizaje y que toma como una medida de la capacidad de generalización. Un factor importante en el ML¹³ son las muestras utilizadas en el proceso de aprendizaje. Éstas junto al tipo de clasificador a obtener establecen el sesgo de partida y son las únicas fuentes de información disponible por el algoritmo de inducción para general el clasificador. Así el resultado del proceso de inducción, va a depender básicamente por un lado de la capacidad del clasificador seleccionado para separar las diferentes clases y por otro de la calidad de las muestras del conjunto de aprendizaje. Las muestras utilizadas en aprendizaje supervisado constan de n-tuplas donde cada elemento se corresponde a un atributo o característica del individuo; junto con un indicativo de la clase a la que pertenece dicha muestra.

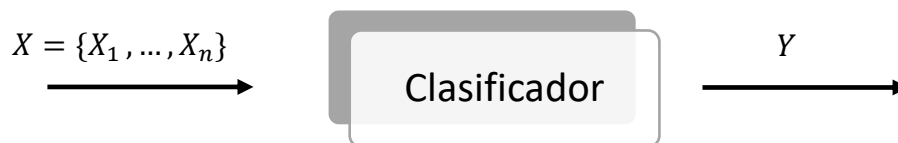


Imagen 2.3 Representación esquemática de un clasificador (Javier & Navarro, 2001a).

¹³ ML por sus siglas en Inglés. *Machine Learning*

Los elementos comprendidos en los algoritmos de inducción o clasificadores se representan en la Imagen 2.3, donde se expone que el clasificador recibe como entrada una muestra no etiquetada $X = \{X_1, \dots, X_n\}$ definida por un vector de características, y como salida, la clasificación o asignación a una clase $Y \in Y$ de la muestra de entrada, este sistema permite estudiar su funcionamiento considerando la relación existente entre las entradas a dicho clasificador y las salidas del mismo. (Javier & Navarro, 2001b)

Las medidas más conocidas para evaluar el rendimiento de un clasificador están basadas en:

- Matriz de confusión
- Análisis ROC (Receiver Operator Curve, Curva de operación del receptor)

2.3.1 Matriz de confusión

La matriz de confusión es una medida que se obtiene cuando se prueba el clasificador en el conjunto de datos de entrenamiento. En la Tabla 2.1 las siglas VP Y VN representan los elementos bien clasificados de la clase positiva y negativa respectivamente y FP Y FN identifican los elementos negativos y positivos mal clasificados respectivamente. Suponiendo que N es el número del conjunto de datos de entrenamiento, entonces $N = VP + FN + FP + VN$, donde el número de instancias clasificadas correctamente es la suma de la diagonal de la matriz y el resto están clasificadas de forma incorrecta.

Tabla 2.1 Matriz de confusión de un problema de dos clases (Abreu, 2015).

Matriz de Confusión		Clase verdadera	
		Pos	Neg
Clase predicha	pos	VP	FP
	neg	FN	VN
Total columna		P	N

Con base en la matriz de confusión se pueden calcular las siguientes medidas de calidad en los clasificadores, mismas que se describen en la Tabla 2.2.

El error ($FP + FN/N$)
(Missclassification Rate)

Ecuación 2.2 Error

La exactitud o precisión ($VP + VN/N$)
(Accuracy)

Ecuación 2.3 Exactitud o precisión

La razón de VP ($rVP = VP/P$), sensibilidad o exhaustividad
(Sensitivity, Recall, True Positive Rate)

Ecuación 2.4 Sensibilidad

La razón de FP ($FP = FP/N$)

Ecuación 2.5 Razón de FP

La razón de VN ($rVN = VN/N$) **o especificidad**
(Specificity, The Negative Rate)

Ecuación 2.6 Especificidad

La razón de FN ($FN = FN/P$)

Ecuación 2.7 Razón de FN

Tabla 2.2 Medidas de evaluación de Clasificadores. (Duval, Vega, & Ruiz, n.d.)

<i>Nombre de la medida</i>	<i>Definición</i>	<i>Descripción</i>
Precisión	$\frac{TP + TN}{N}$	Mide el porciento de muestras correctamente clasificadas.
Especificidad	$\frac{TN}{TN + FP}$	Mide la proporción de muestras negativas correctamente clasificadas.
Sensibilidad (recobrado)	$\frac{TP}{TP + FN}$	Mide la proporción de muestras positivas correctamente clasificadas.
Valor de predicción positiva	$\frac{TP}{TP + FP}$	Da una medida del porciento de los verdaderos positivos del total de muestras clasificadas como positivas. Mide la probabilidad de que una muestra clasificada como positiva sea verdaderamente positiva.
Valor de predicción negativa	$\frac{TN}{FN + TN}$	Da una medida del porciento de los verdaderos negativos del total de muestras clasificadas como negativas. Mide la probabilidad de que una muestra clasificada como negativa sea verdaderamente negativa.

2.3.2 Espacio ROC

Para evaluar el rendimiento del clasificador se utilizan las curvas ROC (*Receiver Operator Curve*, Curva de operación del receptor). En esta curva se representa el valor de la razón de *VP* contra la razón de *FP*, mediante la variación del umbral de decisión. El umbral de decisión es aquel que decide si una instancia x , a partir del vector de salida del clasificador, pertenece o no a cada una de las clases. Se usa el área denominada AUC, (*Area Under the Curve*, Área bajo la curva ROC) como un indicador de la calidad del clasificador. En tanto dicha área esté más cercana a la unidad, el comportamiento del clasificador está más cercano al clasificador perfecto (100% de *VP* con un 0% de *FP*) (Díaz-Barrios et al., 2015).

En la Imagen 2.4 se muestran los elementos del espacio ROC. En un problema de clasificación binaria el espacio ROC es una representación en dos dimensiones con la proporción de ejemplos positivos detectados (*true positive rate*) en el eje Y y la proporción de falsos positivos (*false positive rate*) en el eje X . El punto (0,0) representa a un clasificador que predice todas las instancias como negativas, mientras que el punto (1,1) corresponde a un clasificador que predice todas las instancias como positivas. Por otro lado el punto (0,1) representa una clasificación perfecta mientras que el punto (1,0) corresponde al clasificador que falla en todas sus predicciones. Por último la línea diagonal $y = x$ representa la estrategia de clasificar aleatoriamente (Pe & Verdejo, 2009).

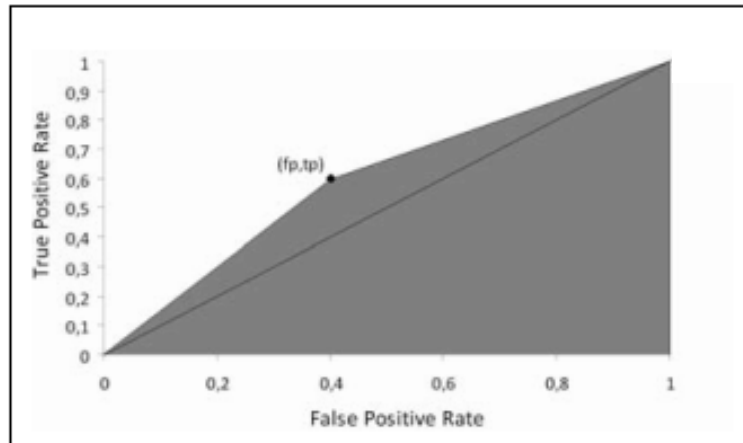


Imagen 2.4 Área bajo la curva (AUC) del punto (0.4, 0.6) (Rodrigo, Peñas, & Verdejo, 2009)

Dada una curva ROC se puede utilizar el área bajo la curva (*Area Under the ROC Curve, AUC*) como indicador del rendimiento del clasificador (Bradley, 1997), sin embargo, el cálculo tradicional de AUC no permite variar la importancia que se le da a los indicadores *tp rate* o *fp rate*. En este caso se debe adoptar una modificación de AUC que permita dicho ajuste. (Pe & Verdejo, 2009).

2.4 Técnicas de estimación

En el diseño de clasificadores a partir de un conjunto de muestras, se busca construir un clasificador que tenga un error pequeño, por lo que es necesario estimar el error que tendrá el clasificador en nuevas muestras no incluidas en el conjunto de aprendizaje, pero que se distribuya de manera similar. En general si se dispone de un conjunto de datos D del que se conoce un determinado parámetro y para todos los elementos de dicho conjunto, y se quiere estimar el valor de esa función para el dominio global, se debe buscar un estimador, que es una variable aleatoria, que estime el parámetro de la población a partir de la muestra contenida en el conjunto de datos D . A continuación se mencionan algunas técnicas de estimación. (Javier & Navarro, 2001b)

2.4.1 Error aparente

La estimación del error con este tipo de estimador se obtiene como la proporción de muestras del conjunto de aprendizaje que han sido erróneamente clasificadas. El clasificador se entrena y se comprueba su tasa de error utilizando el mismo conjunto de aprendizaje, por lo que la estimación es optimista ya que las muestras a clasificar ya han sido vistas previamente por el clasificador. Este efecto en ciertos clasificadores con una alta capacidad de memorización como puede ser un clasificador basado en el vecino más cercano, lleva a considerar una tasa de error bastante baja, resultando la misma, mayor en muestras no vistas anteriormente por el clasificador en el proceso de aprendizaje. El uso de este estimador tiende a generar clasificadores que poseen un mal comportamiento cuando se presentan nuevos casos, es decir, tendrán poca capacidad de generalización, por supuesto, siempre que no exista un gran número de muestras en cuyo caso la tasa de error estimada coincidirá con la tasa de error real. A este efecto de poca capacidad de predicción se le denomina sobreespecialización o sobreajuste.

2.4.2 Entrenamiento y prueba o holdout

El problema de la sobreespecialización que aparece con la utilización del error aparente se debe a que el mismo conjunto de muestras se utiliza en el proceso de inducción del clasificador y de validación del mismo, que no corresponde con las condiciones de funcionamiento reales ya que el clasificador inducido deberá clasificar muestras que no se le han presentado previamente en el proceso de inducción y el rendimiento en esas nuevas muestras puede diferir significativamente del que existía en el proceso de aprendizaje. Para evitar la sobreespecialización del clasificador se utiliza la técnica de remuestreo aleatorio (*random resampling*) del conjunto de aprendizaje que consiste en estimar la tasa de error como la tasa de error promedio de los clasificadores derivados de las particiones de prueba aleatoria e independiente generadas. Estas técnicas se denominan

de remuestreo ya que intentan reproducir el proceso de inducción y funcionamiento (proceso de clasificación) de un clasificador. Un tipo de remuestreo aleatorio es la estimación por entrenamiento y prueba o *hold-out* donde se utiliza un conjunto de muestras para obtener el clasificador y otro diferente para estimar el error, con el fin de eliminar el efecto de la sobreespecialización. Este método consiste en dividir el conjunto de aprendizaje en dos, el primer conjunto es utilizado para llevar a cabo el entrenamiento del clasificador y el resto de las muestras no utilizadas se emplea para estimar la tasa de error del clasificador. Una cuestión importante en el uso de este estimador es cuántas muestras se deben utilizar en el conjunto de entrenamiento de prueba para que la tasa de error estimada sea lo más próxima a la tasa de error real. En principio, el número de muestras necesarias para el conjunto de prueba no debe ser elevado, además, es posible conocer cuál es la diferencia de la tasa de error estimada con la real en función de este número. Una división utilizada con frecuencia es tomar $2/3$ de las muestras para el proceso de aprendizaje y el $1/3$ restante para comprobar el error del clasificador. Un inconveniente que se detecta en este estimador es que solo se utiliza una parte de las muestras disponibles para llevar a cabo el aprendizaje, con lo que se pierde información útil en el proceso de inducción del clasificador.

2.4.3 Validación cruzada

Esta técnica evita la ocultación de parte de las muestras al algoritmo de inducción y la consiguiente pérdida de información. En la validación cruzada (*cross-validation*) el conjunto de aprendizaje se divide en k particiones mutuamente exclusivas $D_1, D_2, \dots, D_k$, conteniendo todas aproximadamente el mismo número de muestras. En muchos casos se suele indicar este caso como validación cruzada con k particiones (*k-fold cross-validation*). A partir de k particiones se obtienen k clasificadores, utilizando como conjunto de aprendizaje para el clasificador i – ésimo todas las particiones menos la partición i – ésima y el error se estima sobre las muestras de la partición

no utilizada en el aprendizaje, por último el error se obtiene como la media de los errores de los k clasificadores.

2.4.4 Bootstrap

Esta técnica se propone para reducir la alta variabilidad que exhibe la validación cruzada en muestras pequeñas (menos de 30 muestras), consiguiendo un aumento de eficiencia comparable a un aumento en el tamaño de la muestra de un 60% (Efron & Tibshirani, n.d.). Dentro de esta tipología existen diversos tipos de estimadores, entre los más utilizados se encuentra el e_0 y el .632. En ambos casos se toman r muestras del conjunto de aprendizaje con reemplazamiento, que conformarán el conjunto de aprendizaje. Todas aquellas muestras que no formen parte del conjunto de aprendizaje se tomarán como conjunto de prueba.

2.5 Descripción de la empresa

Limones Mónica S.A. de C.V. es una empresa ubicada en Martínez de la Torre, Ver con dirección Gregorio Torres Quintero, Col. El Cañizo, dicha empresa está dedicada a la cadena productiva de lima persa principalmente en calidad de exportación, dentro de su filosofía empresarial existe el compromiso sobre la máxima satisfacción de los clientes a través de la alta calidad del producto que ésta ofrece. La planta cuenta con diferentes áreas de producción y administrativas en las que opera un equipo de cien colaboradores aproximadamente.

Según lo que se establece en (Fernández, 2018) las principales áreas y procesos dentro de la cadena productiva de lima persa en la empresa son: área de báscula, área de recepción, área de selección, lavado, sanitizado, secado, encerado, secado de cera, clasificación de fruta la cual se apoya en maquinaria clasificadora por calibres y calidad, área de embalaje y área de almacenamiento. Dentro del mercado a nivel internacional se encuentran los países: Europa, Japón y Estados Unidos, sin embargo, esto no es una limitante para atender también al mercado nacional. Para el proceso de compra y selección de proveedores Limones Mónica S.A de C.V. utiliza como criterio de selección de fruta, la calidad descrita por la empresa (color, textura y tamaño de fruta) con base en

los requerimientos del cliente, así como el precio de compra, que varía dependiendo la temporada de abastecimiento. El principal esfuerzo es seleccionar proveedores de lima persa confiables capaces de cumplir con todas las necesidades del cliente en tiempo y forma, para ello la empresa ha determinado seleccionar a los proveedores bajo los siguientes criterios de estudio: calidad, tiempo de entrega, volumen de producción y confiabilidad. La información de los proveedores de lima persa está integrada en una base de datos que se conforma por 3 agrupamientos, de acuerdo a la ubicación de los 23 proveedores la cual se establece de la siguiente manera: el primer agrupamiento (ubicación norte), comprende los municipios, El Pital con los proveedores P12 y P16, San Rafael con los proveedores P1, P7 y P11 y Papantla con los proveedores P2, P8 y P19. El segundo agrupamiento (ubicación centro), está conformada por Martínez de la Torre con los proveedores P3, P6, P14, P17, P18, P20, Libertad con el proveedor P5, Coapeche con el proveedor P21 y Misantla con el proveedor P23. El tercer agrupamiento (ubicación sur), la conforman los proveedores foráneos, Tabasco con el proveedor P22, Oaxaca con el proveedor P10 y Tlaltelela, Xalapa con el proveedor P4 y P9.

2.6 Estado del arte

En este apartado se exponen aportaciones de diferentes autores tomando como base la revisión de la literatura que sirve de contexto en el desarrollo de la propuesta de investigación.

La logística es aquella parte del proceso de la cadena de suministro que planea, implementa y controla el flujo y almacenamiento de productos y servicios, y su información relacionada, desde el punto donde se originan hasta el punto donde se consumen, en forma eficiente y al menor costo posible, para satisfacer los requerimientos de los clientes. (Segura Sánchez, 2005)

Partiendo de este enfoque podemos establecer que la empacadora de lima persa Mónica S.A. de C.V forma parte de una CS¹⁴ y el cumplimiento de los objetivos organizacionales como la satisfacción de la demanda de los clientes están dados en

¹⁴ Acrónimo de Cadena de Suministro

función del buen desempeño y control de cada uno de los elementos de la cadena de suministro.

En la industria alimentaria, las CS son sistemas complejos en continuo cambio que involucran a muchos participantes, entre ellos, proveedores, productores, distribuidores, comercializadores mayoristas y minoristas, entidades de regulación, consumidores, etc. y que tiene como fin el cumplimiento de las expectativas de los clientes. La coordinación entre los elementos de la CS es efectiva si se conoce y gestiona la información, es decir, los flujos de materiales y servicios que pueden considerarse como un sistema de elementos que comparten información para cumplir un objetivo. En las CS agroalimentarias, como en otras áreas, la innovación juega un papel importante por lo que se requiere conocer el entorno y las mejores formas de diseñar y gestionar la información, que se alimenta con las organizaciones responsables de proveer, producir, distribuir, procesar y comercializar productos agrícolas a los consumidores finales. (Vianchá Sánchez, 2014).

Todas las organizaciones cuentan con alguna clase de sistemas de información, (registros, documentos, etc.), sin embargo, para contar con un sistema funcional de información que satisfaga diversas necesidades, todos los datos medibles deben ser organizados de tal manera que sea fácil de registrarlos, almacenarlos, procesarlos, recuperarlos y comunicarlos según lo requieran los usuarios. En este sentido, la propuesta de solución del presente trabajo conduce a exponer un marco de referencia en el tema de selección de proveedores y métodos de algoritmos de clasificación que fundamenten el desarrollo de investigaciones similares.

2.6.1 Resultados de la revisión literaria

Las plataformas digitales de búsqueda especializada en literatura científica o académica utilizadas en el proceso de revisión se mencionan a continuación: Dialnet, Scielo, Springer, Science Research, Google Scholar y Books Google principalmente. Los resultados de la investigación documental corresponden a trabajos realizados que abordan el método de algoritmos de clasificación e hiperparametrización y trabajos

previos sobre el proceso de selección de proveedores basado en el análisis de datos históricos.

2.6.1.1 Investigaciones previas –Algoritmos de clasificación e hiperparametrización:

(Corso & Lorena, 2007), el estudio tiene como finalidad descubrir el nivel de confianza de los **algoritmos de clasificación supervisada basados en reglas que implementa el software WEKA**, la interrogante que este trabajo pretende develar es si este nivel de confianza aumenta o disminuye al disponer de mayor cantidad de datos de entrenamiento.

Los algoritmos se implementaron en un conjunto de datos nominales, los repositorios sobre los cuales se efectuaron las pruebas se cargaron con 300, 1420 y 2000 instancias, el test elegido para realizar todas las pruebas fue *Use training set*.

Con base en los resultados obtenidos se estableció un ranking teniendo en cuenta cuál de los algoritmos de clasificación basados en reglas, fue más confiable, los resultados obtenidos para el archivo cargado con 350 instancias representan el algoritmo *rules.NNge* con 100% de instancias correctamente clasificadas, con estadístico Kappa=1 y error absoluto=0, el modelo obtenido se visualiza por medio de un conjunto de siete reglas generadas al aplicar el algoritmo *NNge*, los resultados obtenidos para el archivo con 1420 instancias el algoritmo *NNge* en sus instancias correctamente clasificadas, estadístico kappa y error absoluto mantienen los valores anteriores, los resultados obtenidos para el archivo cargado con 2000 instancias el 100% de instancias correctamente clasificadas, estadístico Kappa=1 y error absoluto=0, están dados por los algoritmos *rules.NNge* y *rules.PART* por lo que se concluye que según los resultados obtenidos de las pruebas, la cantidad de instancias no influye de manera significativa en la bondad de los modelos generados por la aplicación de los algoritmos de clasificación basada en reglas.

(Lindauer & Hutter, 2014), se presenta el primer acercamiento al uso de la **configuración de algoritmos para optimizar la selección de algoritmos mediante un enfoque denominado *AutoFolio* para atender la problemática de selección de algoritmos** y responder a las siguientes interrogantes ¿Cuál enfoque de selección

utilizar? ¿Cómo seleccionar algoritmos y cómo establecer sus parámetros de manera efectiva? ¿Cómo establecer técnicas para mejorar la selección de algoritmos?, esta problemática se aborda de manera simultánea utilizando la configuración automática de algoritmos, los autores pretenden demostrar que pueden utilizar configuradores de algoritmos para configurar automáticamente *claspfolio2*, el cual implementa una gran variedad de diferentes enfoques de técnicas de selección de algoritmos y sus respectivos parámetros en un único *framework* altamente parametrizado.

En lugar del enfoque común de prueba y error manual, los autores proponen responder automáticamente a las preguntas anteriores mediante la configuración de algoritmos para configurar *frameworks* flexibles de selección de algoritmos, como *claspfolio 2*, mientras que la configuración manual es propensa a errores, sesgada por humanos y requiere mucho tiempo humano y conocimiento experto, el enfoque que se presenta es totalmente automático e imparcial y aprovecha el poder completo de una amplia gama de AS¹⁵, de este modo, facilita el uso más fácil y eficaz de la selección de algoritmos y hace que las técnicas de AS sean accesibles a una comunidad más amplia, *AutoFolio*, es un enfoque general para aplicar la configuración de algoritmos a la selección de algoritmos y proporcionar una implementación de código abierto basada en los configuradores de algoritmos *SMAC*, *ParamILS* y el selector de algoritmo *claspfolio 2*, se analiza empíricamente el rendimiento del enfoque *Autofolio* (experimentos basados en *claspfolio 2*) usando *sklearn 0.15.0* (paquete de aprendizaje automático de *Python*) y *SMAC 2.06.01*, *Autofolio* es ejecutado en los doce escenarios de selección de algoritmos que conforman la biblioteca de selección de algoritmos, estos escenarios comprenden una amplia variedad de problemas de combinación dura; cada uno de ellos incluye los datos de rendimiento de un rango de solucionadores (entre 2 y 31) para un conjunto de instancias, y características de instancia organizadas en grupos de características con costos asociados, el uso de repositorios en internet permite obtener detalles sobre todos los escenarios lo que facilita comparar *AutoFolio* de una manera justa y uniforme con otros métodos de selección de algoritmos, los autores mencionan un enfoque relacionado para el aprendizaje automático supervisado general, *Auto-WEKA*, como un sistema que aborda el problema combinado de seleccionar un algoritmo de aprendizaje automático

¹⁵ Acrónimo de Aprendizaje Supervisado

del *framework* de *WEKA* y optimizar sus hiperparámetros, *Auto-WEKA* y *AutoFolio* utilizan el mismo mecanismo de validación cruzada para definir el proceso de configuración.

Esta investigación demostró que el enfoque, denominado *AutoFolio*, puede mejorar significativamente el rendimiento de *claspfolio 2* en 11 de los 12 escenarios de la Biblioteca de selección de algoritmos, los autores afirman que los selectores de algoritmos han demostrado empíricamente probar el estado del arte para instancias heterogéneas, se comparó el rendimiento de la configuración por defecto de *claspfolio 2* (similar a SATzilla'11) con el de la configuración optimizada por *AutoFolio*, para todos los escenarios de selección, *AutoFolio* logró un mejor rendimiento en los datos de entrenamiento, y el rendimiento mejorado en los datos de prueba se obtuvo en todos los escenarios menos uno, las mejoras en el rendimiento de los datos de las pruebas fueron estadísticamente significativas, bajo criterio de los autores se presenta *AutoFolio*, como el primer enfoque para configurar automáticamente los selectores de algoritmos, *Autofolio* emplea 2 configuradores de algoritmos complementarios del estado del arte : *SMAC* y *ParamILS*, usando este enfoque basado en la selección de algoritmos altamente parametrizados *Claspfolio 2*, se mostró que los configuradores de algoritmos del estado del arte pueden encontrar automáticamente configuraciones optimizadas de selectores de algoritmos que funcionan significativamente mejor que los selectores configurados manualmente, el sistema *claspfolio 2* con configuración automática mostró mejoras en el rendimiento de un factor entre 1.3 y 12.4 en términos de puntajes PAR10 en comparación con la mejor solución única para el conjunto de algoritmos dado.

(Mohr, Wever, & Hüllermeier, 2018), en este trabajo presenta un **algoritmo para el servicio automatizado aplicado al área de aprendizaje máquina**, se considera el problema de encontrar una combinación con la mejor calidad (numérica) con la limitación de que esta medida de calidad puede no estar agregada a los servicios contenidos en la composición, asumiendo una calidad numérica que se optimiza, pero que no hay una forma (conocida) de asignar estáticamente cualidades a los servicios existentes y calcular la calidad de una combinación a partir de estas cualidades, el problema más comúnmente abordado es la combinación o configuración de un solo proceso, ya sea mediante la creación de instancias o la combinación de un flujo de trabajo abstracto o mediante la

creación de dicho proceso desde cero dada una descripción de comportamiento en términos de condiciones previas, se argumenta que, recientemente este enfoque ya no se ha abordado debido a que los servicios en el mundo real no parecían tan bien descritos como se esperaba, lo que descarta muchos enfoques que se basan en tales descripciones e incluso para los enfoques que no se basan en tales suposiciones, la combinación del servicio automatizado se ha resuelto principalmente en ejemplos que no han demostrado relevancia en escenarios del mundo real.

La principal contribución de este documento es demostrar que la composición de servicios automatizada puede superar la combinación de software manual o no orientada a servicios en aplicaciones del mundo real, el dominio del problema de combinación de servicio considerado es aprendizaje máquina automatizado, es decir, dados algunos datos de muestra, la tarea consiste en componer un canal de aprendizaje automático (que consiste en servicios de aprendizaje automático) que maximice la precisión de la clasificación en datos de la misma fuente, los autores mencionan que hasta ahora, este problema solo se ha abordado con herramientas específicas como *Auto-weka* y *auto-sklearn*, se presenta una descripción técnica detallada y una evaluación preliminar que muestra los beneficios de utilizar los servicios con el propósito de aprendizaje automático automatizado, similar a los enfoques de combinación existentes, este enfoque se basa en la planificación jerárquica, la diferencia principal con respecto a los enfoques existentes es que el proceso de búsqueda se basa en las medidas de rendimiento que se adquieren de la ejecución de los candidatos de la combinación, asimismo se hace una evaluación empírica en la que se compara el enfoque contra las herramientas tradicionales, para este caso de estudio se considera el dominio del aprendizaje automático automatizado, *Auto-ML*, el cual tiene como objetivo seleccionar y configurar automáticamente el algoritmo en la línea de aprendizaje automático, por lo general, consiste en uno o más algoritmos de preprocesador (componentes principales) y un algoritmo de clasificación (árboles de decisión, logística, regresión, etc.), los enfoques del estado del arte *auto-sklearn* y *Auto-WEKA* reducen el problema de sección combinada de algoritmos y el problema de optimización de hiperparámetros a un mero problema de optimización de hiperparámetros, considerando la selección de un algoritmo para función de preprocesamiento y un modelo de clasificador como parámetros adicionales para una

herramienta de optimización de hiperparámetros, estas herramientas están comprometidas con una determinada biblioteca (*scikit-learn* o *WEKA*) y a un lenguaje de programación específico (*Python* o *Java*), sin embargo, estas bibliotecas no son iguales ni una subestima a la otra, las implementaciones de ciertos algoritmos de aprendizaje automático generalmente difieren significativamente incluso para un algoritmo particular.

Como resultado del trabajo se presenta una evaluación experimental, que muestra que el enfoque basado en el servicio que combina los servicios *WEKA* y *scikit-learn* es a menudo mejor que usar la misma técnica de búsqueda con algoritmos de una sola biblioteca, y es incluso más competitivo con enfoques expertos como *Auto-WEKA* y *auto-sklearn*, para evaluar cuál es el beneficio de usar los servicios, se comparó *MLS-Plan* con otras herramientas de *Auto-ML* como *Auto-WEKA* (*WEKA*, *Java*), *auto-sklearn* y *TPOT* (*scikit-learn*, *Python*) para colocar los resultados en el contexto del estado del arte, los resultados se obtuvieron mediante 20 ejecuciones en 21 conjuntos de datos (repositorio *OpenML*) con un tiempo de espera de 1h, la importancia de una mejora o degradación se determinó utilizando la Prueba t con un umbral para la puntuación de 2.086, el tiempo de espera para la evaluación interna de una sola solución se estableció en 5 minutos (si lo permite la herramienta correspondiente), las corridas que no se apegaron a lo establecido fueron canceladas sin considerar sus resultados, en cada ejecución, se utilizó una división aleatoria y estratificada de los datos, el 70% para el aprendizaje (búsqueda) y el 30% para las pruebas, utilizando las mismas divisiones para todos los candidatos, los resultados muestran que el rendimiento de los enfoques varía mucho a través de los conjuntos de datos, demostrando que no existe un enfoque que sea el mejor entre la mayoría de los conjuntos de datos, *TPOT* parece dominar en conjuntos de datos pequeños, el enfoque de la evaluación es la comparación del *MLS-Plan* con los servicios de los *frameworks* *WEKA* y *scikit-learn*, por un lado, y por otra parte la misma estrategia de combinación utilizando algoritmos de una sola de estas bibliotecas, de esta manera, se adquiere conocimiento sobre el beneficio que se obtiene de la implementación orientada al servicio, el cual permite combinar algoritmos de ambas bibliotecas, los demás enfoques están destinados a proporcionar valores de referencia de herramientas reconocidas en el área respectiva, pero su desempeño no es relevante, asimismo se destaca la importancia que, a priori, no está del todo claro cuál de las combinaciones

sería mejor, aplicar *MLS-Plan* con los servicios *WEKA* y *scikit-learn*, no significa considerar el espacio de solución conjunta, ya que se considera que los algoritmos aparecen en ambas bibliotecas solo una vez, el espacio de búsqueda de *MLS-Plan* con algoritmos de ambas bibliotecas no es un superconjunto del espacio de búsqueda de *MLS-Plan* utilizando solo algoritmos *WEKA* o *scikit-learn* respectivamente, el espacio de búsqueda es mucho más grande en comparación con el uso de una sola de las bibliotecas, lo que hace más probable que las buenas soluciones necesiten más tiempo para ser encontradas, los resultados muestran que la combinación de las bibliotecas a menudo es significativa, mucho mejor que una o incluso las dos versiones de una sola biblioteca, la impresión general es que la variante basada en el servicio produce mejoras significativas, no solo sobre las otras variantes del *MLS-Plan*, sino a nivel global, con estos resultados se concluye que la combinación del servicio automatizado es un enfoque relevante para resolver el problema de *Auto-ML*, por lo que debe estar en la lista de técnicas de solución, los resultados se interpretan como evidencia de la utilidad de la combinación de servicios automatizada para el problema de *Auto-ML* del mundo real.

(Kotthoff, Thornton, Hoos, Hutter, & Leyton-Brown, 2016), los autores proponen **Auto-WEKA 2.0 para la selección automática de modelos y optimización de hiperparámetros en WEKA**, en búsqueda de resolver la problemática que presentan los usuarios cuando resulta difícil identificar el mejor enfoque para el conjunto de datos particular entre los muchos disponibles.

Se describe la nueva versión de *Auto-WEKA* como un sistema diseñado para ayudar a los usuarios a maximizar el rendimiento mediante la búsqueda automática a través del conjunto de los algoritmos de aprendizaje de *WEKA* y sus posibles configuraciones de hiperparámetros, utilizando un método de optimización Bayesiano, el paquete está estrechamente integrado con *WEKA*, lo que lo hace accesible para los usuarios finales como cualquier otro algoritmo de aprendizaje.

Los resultados obtenidos de la versión propuesta radica en que a diferencia de la versión anterior donde se debía descargar manualmente *Auto-Weka* y ejecutar de forma independiente de *WEKA*, *Auto-WEKA 2.0* ya está disponible a través del gestor de paquetes de *WEKA*, sin instalar software por separado, *Auto-WEKA* fue el primer método

para utilizar la optimización bayesiana para crear automáticamente una estructura de aprendizaje máquina altamente parametrizado con solo presionar un botón, este enfoque de aprendizaje automático automatizado (*AutoML*) también se ha aplicado recientemente a *Python* y *scikit-learn*, en *Auto-sklearn* un paquete similar a *Auto-WEKA*, *Auto-sklearn* usa el mismo optimizador bayesiano que *Auto-WEKA*, pero comprende un espacio más pequeño de modelos e hiperparámetros, ya que *scikit-learn* no implementa tantas diferentes técnicas de aprendizaje automático como *WEKA*; sin embargo, *Auto-sklearn* incluye técnicas adicionales de meta-aprendizaje, también es posible optimizar los hiperparámetros utilizando los propios paquetes de multi-búsqueda de *WEKA*, sin embargo, estos paquetes solo permiten ajustar un método de aprendizaje y un método de filtrado a la vez, la búsqueda maneja solo un hiperparámetro y los nombres de los hiperparámetros y los posibles valores deben ser especificados por el usuario, las cuatro características principales de *Autoweka 2.0* son: *Expande su enfoque anterior de clasificación permitiendo ahora algoritmos de regresión.*Admite la optimización de todas las métricas de rendimiento compatibles con *WEKA*.*Apoya las ejecuciones paralelas (en una sola máquina) para encontrar buenas configuraciones más rápido y guardar las N mejores configuraciones de cada ejecución en lugar de solo la única mejor.*Está totalmente integrado con *WEKA*, proporcionando simplicidad en una interfaz que permite pulsar un botón y no requiere conocimiento sobre los algoritmos de aprendizaje disponibles o sus hiperparámetros, solicitando al usuario el conjunto de datos, el límite de memoria a utilizar y el tiempo disponible para el proceso de aprendizaje.

(Thornton et al., 2012a), se propone el enfoque ***Auto-WEKA* para la selección automatizada y optimización de hiperparámetros de los algoritmos de clasificación**, en esta investigación los autores mencionan que ha habido una cantidad considerable de trabajos anteriores sobre la elección de los algoritmos de aprendizaje y, por separado, en la optimización de hiperparámetros, sin embargo, la propuesta de su investigación se basa en atender ambos problemas de manera conjunta ya que según determinan no existe algún trabajo que se ocupe de una problemática similar, es decir, se enfrentan al desafío considerable de elegir el algoritmo de aprendizaje correcto para un problema dado, además, la mayoría de los algoritmos tienen los llamados hiperparámetros que deben ajustarse a valores adecuados para obtener un buen rendimiento en un contexto

de uso dado, la dificultad de este desafío y la cantidad de esfuerzo requerido para evaluar alternativas alientan a los usuarios simplemente a elegir un algoritmo basado en su reputación o atractivo intuitivo y a dejar los hiperparámetros establecidos en sus valores predeterminados, lo que conduce a un rendimiento por debajo del mejor método.

Su propuesta demuestra la viabilidad de utilizar un enfoque totalmente automatizado para elegir tanto un algoritmo de aprendizaje como sus hiperparámetros, aprovechando las innovaciones recientes en la optimización Bayesiana, el enfoque es aplicado específicamente a la gama completa de clasificadores implementados en *WEKA*, que abarcan 3 métodos ensamblados, 14 meta-métodos, 30 clasificadores base y una amplia gama de ajustes de hiperparámetros para cada uno de éstos, se utilizan 10 conjuntos de datos de un repositorio en línea y se muestra mejor rendimiento de clasificación que el de la validación cruzada completa sobre la configuración predeterminada de hiperparámetros de los 47 algoritmos de clasificación, cada conjunto de datos se particionó en una división aleatoria 70% (datos de entrenamiento) y 30% (datos de prueba) para evaluar las configuraciones encontradas por los diversos métodos de optimización, se evaluaron las configuraciones de hiperparámetros basadas en una validación cruzada estándar de 10 dobleces para *Ex-Def*, *Hoeffding races*, *TPE*, and *SMAC*.

Los autores concluyen que el enfoque *Auto-WEKA*, permitirá a los usuarios de los algoritmos de aprendizaje automático tomar mejores decisiones y, por lo tanto, obtener un mejor rendimiento de manera totalmente automatizada, los resultados que se presentan tienen en cuenta los problemas de clasificación, se estudia el rendimiento que se puede lograr aplicando métodos automatizados y técnicas de selección de hiperparámetros en *Auto-WEKA*, en los 10 conjuntos de datos, *TPE* y *SMAC* (los métodos de optimización Bayesianos) dieron un rendimiento mejor o igual que *Hoeffding races* y *Ex-Def* (que solo realizan la selección del modelo), entre los métodos de optimización bayesiana, *SMAC* se desempeñó mejor en 8 de 10 conjuntos de datos, los clasificadores seleccionados por cada uno de los métodos para cada conjunto de datos, *TPE* y *SMAC* poseen una tendencia para elegir entre el mismo conjunto de clasificadores para cada conjunto de datos, cualitativamente, los métodos más comunes seleccionados son los métodos ensamblados y árboles aleatorios.

(Thornton, Hutter, Hoos, & Leyton-Brown, 2012b), la tesis **Auto-WEKA: Selección combinada y optimización de hiperparámetros de algoritmos de aprendizaje máquina supervisado**, presenta un enfoque ante la problemática común que la mayoría de los usuarios presenta debido a la falta de comprensión profunda sobre la terminología y los mecanismos asociados con cada algoritmo de aprendizaje.

El enfoque sugiere que dado un conjunto de datos, se seleccione automática y simultáneamente un algoritmo de aprendizaje y configure sus hiperparámetros para optimizar su rendimiento, los autores llaman a este problema: el problema combinado de selección de algoritmo y optimización de hiperparámetros (CASH), se proporciona el acercamiento a una herramienta, *Auto-WEKA*, que requiere una participación mínima del usuario y proporciona una solución para el problema *CASH*, buscando en los algoritmos de aprendizaje proporcionados en la distribución estándar de *WEKA*, el problema *CASH* consiste en dos subproblemas principales: la selección de algoritmo y la optimización de hiperparámetros, el trabajo se centra en los problemas de aprendizaje automático supervisado.

Para dar solución al problema *CASH*, se creó la herramienta *Auto-WEKA*, que resuelve este problema para todos los algoritmos de clasificación y regresión en combinación con todos los selectores / evaluadores de características implementados en el paquete estándar *WEKA*, *Auto-weka* se evaluó en 21 conjuntos de datos de clasificación, optimizando la tasa de error de clasificación, con el 100% correspondiente a los clasificadores que no hicieron predicciones correctas, después de filtrar los algoritmos de aprendizaje no válidos, *Auto-WEKA* buscó en un espacio de hiperparámetros de 786 dimensiones, la búsqueda sobre los hiperparámetros de todos los clasificadores base arrojó mejores resultados que *Ex-Def* en 17/21 casos, subrayando la importancia de no solo elegir el algoritmo correcto sino también de configurar bien sus hiperparámetros, *Auto-WEKA* fue efectiva para optimizar su función objetivo dada para la clasificación; sin embargo, esto no es suficiente para concluir que produce modelos que se generalizan bien, como resultado se muestra qué métodos base fueron elegidos, *AdaBoostM1* seleccionó frecuentemente *SVM* en conjuntos de datos pequeños, pero nunca en grandes, mientras que los árboles aleatorios y de *REP* fueron elegidos a menudo para grandes conjuntos de datos, en el clasificador *MultiClass*, los dos métodos seleccionados

con mayor frecuencia fueron la regresión logística y los *SVM*, la regresión logística, así como el árbol aleatorio seleccionado con frecuencia por *AdaBoostM1*, a menudo no se seleccionaron como clasificadores base por sí solos, se utilizaron 13 conjuntos de datos para evaluar el desempeño de *Auto-WEKA* en tareas de regresión entre las variantes de *Auto-WEKA*, *SMAC* a menudo encontró la mejor configuración de hiperparámetro en comparación con *TPE* (en 10/13 casos) y *I / F-Race* (en 8/13 casos).

2.6.1.2 Investigaciones previas – Evaluación y selección de proveedores basada en técnicas de IA.

(Lambert, Lasserre, Olivia, Rivera, & Montalvo, n.d.), presenta una **propuesta basada en técnicas de IA** atendiendo a la problemática que está relacionada particularmente con la dificultad que tiene una exportadora de cítricos para identificar y seleccionar productores-proveedores confiables que aseguren un abasto y calidad de fruto en atención a la demanda de sus mercados.

Se establece dicha propuesta a fin de sincronizar la cadena de suministros mediante una arquitectura metodológica para la selección del proveedor-productor para una exportadora de limón persa basada en tres técnicas de la IA para determinar el impacto de la selección de proveedores en la cadena de suministro a partir de criterios de selección no tradicionales del contexto de la producción de limón persa y relacionado con las buenas prácticas agrícolas. Se plantea el uso de una Red Neuronal Artificial (RNA) para la predicción del rendimiento de huerto y calidad de fruto a partir de datos históricos sobre la trazabilidad de cultivo en huerto. Un modelo de Lógica Difusa (LD) que considera la evaluación del rendimiento del huerto y la calidad del fruto a partir de la experiencia y conocimiento del experto. Un Algoritmo Genético (AG) que optimiza el resultado de un modelo matemático multicriterio que maximice la producción en huerto, y maximice la calidad del fruto.

Como resultado se llega a una comparación de las tres técnicas: el modelo de lógica difusa basado en la experticia de los citricultores y personas dedicadas a este dominio, la RNA sustentada en casos históricos, y el modelo matemático para optimizar parámetros inciertos, donde el autor expone que se ha encontrado una abundancia de trabajos que atienden este problema en empresas del sector manufacturero y de

servicios, haciendo uso de diferentes herramientas de la ingeniería, técnicas de la inteligencia artificial, y modelos matemáticos, donde este caso no es así para el sector agrícola, que si bien se presentan investigaciones que han hecho uso de diversas técnicas que incluyen a las de la IA, entre ellas no se reportan trabajos que atiendan el problema de selección de proveedores en dicho sector.

(Fernández, 2018), resuelve el problema de **evaluar los proveedores para una empacadora de la industria citrícola basado en técnica Deep Learning**, en el trabajo se expone que en la actualidad existe una gran preocupación de la empacadora ya que cotidianamente los pedidos del producto final salen fuera del tiempo estimado y no llegan en la fecha programada al cliente, lo que impide reunir el volumen y calidad de fruta solicitada por el tiempo.

El trabajo propone seleccionar proveedores de lima persa apoyado de la técnica *Deep Learning*, para asegurar el abastecimiento de fruta con las características de exportación por la empresa Limones Mónica, S.A. de C.V. para apoyar el área de proceso en cuanto a la calidad, exigencias y prioridades que el cliente requiere se propone un modelo de selección de proveedores mediante la técnica *Deep Learning* y la metodología *AHP*, el cual permitirá con claridad analizar cada proveedor, para identificar a los proveedores con calidad de exportación que se requieran en el momento para atender la necesidad de fruta requerida del programa de ventas.

Los proveedores de Lima persa que abastecen a la empresa Limones Mónica, S.A. de C.V. deben de cumplir criterios de calidad normativos descritos por el cliente, como lo son: color, textura y tamaño de la fruta, la empresa ha determinado seleccionar a los proveedores de Lima persa bajo los siguientes criterios de estudio: calidad, tiempo de entrega, volumen de producción y confiabilidad, la base de datos se construyó con información histórica de los 23 proveedores que abastecen a la empresa en temporada baja y temporada alta para los años comprendidos del 2014 al 2017. Los criterios a evaluar mediante AHP se describen con base a su importancia, donde cada proveedor es evaluado en una escala en particular para cada criterio los cuales fueron definidos por el experto en producción de Lima persa. La estructura jerárquica mediante AHP presenta las alternativas de decisión (23 proveedores), y los criterios a evaluar (calibres) que son descritos y argumentados por la empresa en base a su calidad, para ponderar cada una

de las alternativas de acuerdo a los criterios, se utilizó la escala de Saaty, se implementó la técnica Deep Learning, que es basada en las Redes Neuronales Artificiales. El desarrollo de las técnicas están sujetas a los programas de software de Excel y Weka, para una selección de proveedores de Lima persa, para el desarrollo de la técnica Deep Learning, se utilizaron los clasificadores Meta (Ensamblados)+ DI4jMlpClassifier (Perceptrón Multicapa), el cual permitió realizar las combinaciones de entrenamiento, de cada una de las bases de datos de la temporada baja y la temporada alta. En el experimento se seleccionó el método de clasificación Meta (ensamblado), el cual contiene una serie de clasificadores que fueron combinados con el clasificador DI4jMlpClassifier (Perceptrón multicapa), en los resultados de estas combinaciones el clasificador que obtuvo mayor porcentaje de menor error fue el clasificador FilteredClassifier, siendo utilizado para el segundo experimento con la combinación de FilteredClassifier + DI4jMlpClassifier (Perceptrón multicapa). Se eligió la validación cruzada como validación del modelo, en los experimentos se utilizó a partir de 10 folds en adelante, con un porcentaje de 66.6% con el que aprende el modelo por default. Los primeros resultados que se obtuvieron con la combinación de los clasificadores Meta+DI4jMlpClassifier, para la temporada alta y para la temporada baja demuestran que el clasificador FilteredClassifier obtuvo el mayor porcentaje de menor error, esto quiere decir, que es el más confiable para profundizar los experimentos y obtener el menor error posible. Como resultado final se describe el porcentaje de rendimiento de los 23 proveedores (3 zonas) de lima persa para las temporadas (alta y baja) mostrando quién tiene el mejor resultado (porcentaje) para cubrir un demanda específica.

(Machado & Garrido, 2001), abordan la **selección de proveedores internacionales** como un proceso marcado por su complejidad, ante la necesidad de evaluar a los diferentes suministradores en función de criterios tanto de carácter cuantitativo como cualitativo provocando dificultad para encontrar un suministrador que destaque en todos los criterios.

Se propone una técnica de apoyo a la decisión AHP¹⁶, se trata de una técnica multicriterio que dispone de un software específico, denominado *Expert Choice*, el cual

¹⁶ AHP por sus siglas en Inglés. *Analytic Hierarchy Process*

posibilita la automatización de su metodología, haciéndola más asequible para que pueda ser empleada sin necesidad de disponer de grandes conocimientos matemáticos o informáticos, el trabajo de investigación muestra la metodología a través de una aplicación práctica como es la selección de proveedores internacionales en el comercio exterior.

Como resultado se observa la construcción de la jerarquía representativa del problema, se realiza una comparación por pares de los elementos del segundo nivel con respecto al objetivo global empleando una escala de medida generando una matriz de comparaciones para encontrar los pesos que midan la importancia relativa de cada criterio en relación a su meta global, lo que permitirá al decisor comparar las distintas alternativas de elección con respecto a cada criterio considerado, para posteriormente cotejar por pares los proveedores con relación al primer criterio lo cual genera los pesos que representan la importancia relativa de cada uno de los suministradores en relación al primer criterio. Una vez finalizada la etapa de comparaciones, la aplicación del método AHP termina con un proceso que, a partir de los resultados obtenidos y mediante una agregación multiplicativa de los mismos entre niveles jerárquicos, permite seleccionar al mejor suministrador. Esta metodología posibilita también la medición de la consistencia del sujeto decisor en la etapa de realización de comparaciones, para medir las propiedades de transitividad y secuencia de comparaciones se utiliza un indicador denominado ratio de consistencia, el cual su cálculo es más complejo, pero este inconveniente se solventa empleando el Software *Expert Choice* el cual permite testar la sensibilidad de la solución final ante cambios en la información disponible. Las técnicas multicriterio como AHP está diseñada como herramienta de modelización flexible y adaptable a las necesidades específicas de cualquier organización y permite, no sólo medir el grado de consistencia del decisor al realizar comparaciones sino, también, conocer la estructura del problema por medio de una jerarquía representativa y automatizar su metodología mediante *Expert Choice*.

(Aguilar, n.d.), ante la problemática de modelizar el proceso de **toma de decisiones con criterios múltiples**, se presenta una alternativa de solución frente a este proceso en los que entran en juego: una decisión a ser tomada, los eventos desconocidos que pueden afectar el o los resultados, los posibles cursos de acción, y el o los resultados

mismos, mediante los modelos multicriterio el decisor puede estimar las posibles implicaciones que puede tomar cada curso de acción, para obtener una mejor comprensión de las vinculaciones entre sus acciones y sus objetivos. Los métodos y modelos para la toma de decisiones con criterios múltiples, proveen herramientas de utilidad a la hora de analizar problemas complejos.

El trabajo presenta un resumen derivado del análisis de la literatura relacionada con modelos de preferencia en el cual se han identificado cinco importantes modelos específicos, cuyos marcos conceptuales se comentan brevemente en el documento: El proceso analítico jerárquico, el método Electra III, el método Promethee, el método MacBeth y el método NAIADE.

El autor concluye que los métodos para la toma de decisiones con múltiples criterios son útiles para ayudar a resolver problemas complejos, en los que es necesaria la comparación de alternativas, considerando varios puntos de vista simultáneamente, debido al grado de conflicto que presentan los temas de las organizaciones y a la subjetividad de muchos de sus aspectos, los métodos para la toma de decisiones con múltiples criterios, son potencialmente útiles para el análisis de decisiones en dicho campo.

Capítulo 3

Metodología de la Investigación

El marco metodológico del presente trabajo se divide en dos aspectos, el primero se refiere al desarrollo metodológico el cual se sustenta en la implementación de la etapas de la metodología de investigación (Alcántara, n.d.) y el segundo aspecto atiende a la metodología de selección de proveedores. (Valdivieso, 2010), implementando el enfoque de selección combinada de algoritmos e hiperparametrización (Thornton, Hutter, Hoos, & Leyton-Brown, 2013).

3.1 Desarrollo metodológico

A continuación se establecen los pasos del procedimiento metodológico que sustenta la propuesta de investigación con el fin de formalizar los objetivos descritos en esta tesis.

3.1.1 Diseño de investigación

El presente trabajo se define como un proyecto de investigación y evaluación a nivel exploratorio dado que analiza información secundaria disponible, el proyecto se aplica a datos históricos de proveeduría en la empacadora Mónica de Martínez de la Torre, Veracruz, donde el conjunto de datos incorpora la recolección de información registrada durante el periodo 2014 a 2018, bajo el entendido que servirá como preámbulo para conducir a una investigación conclusiva ya que la naturaleza de los datos es de tipo

cuantitativa y los resultados generados de la investigación permiten ser utilizados como entrada para la toma de decisiones, asimismo se atiende a la definición del término: investigación conclusiva, ya que este tipo de investigación está diseñada como apoyo a los decisores para evaluar y seleccionar la mejor de sus alternativas, implementando por una parte un diseño de investigación experimental pura donde para este caso se aplican técnicas de minería de datos que permiten controlar dos grupos de datos (entrenamiento y prueba) y por otra parte se establece la investigación aplicada la cual se orienta a demostrar la validez de la técnica de selección combinada de clasificadores mediante el método de hiperparametrización en el proceso de discriminación de proveedores para la empacadora en mención.

3.1.2 Definición de las variables

En el proceso de proveeduría de lima persa se identifican algunas variables que son tomadas en cuenta para evaluar el desempeño de los proveedores con base en criterios específicos de cada empresa, algunas variables a considerar pueden ser: variables del sistema (fertilizante, características del suelo y huerto...), variables inciertas o de medio ambiente (huracán, exceso de lluvias, sequías, plagas...), variables del producto (precio, calidad, volumen de producción...), criterios del proveedor (capacidad de respuesta, cumplimiento de entrega a tiempo...) entre otras.

3.1.3 Población y muestra

La población de estudio se compone de los datos históricos sobre los proveedores que atiende las necesidades de requerimiento de limón persa en la empacadora Mónica S.A de C.V. donde la muestra se construye a partir de la técnica implementada al conjunto de datos (*cross-validation*, *percentage split* y muestra representativa) y el muestreo se realiza atendiendo la definición propia de cada técnica.

3.1.4 Instrumentos

El principal instrumento del presente trabajo se centra en una base de datos sobre producción citrícola derivada de estudios anteriores la cual se analiza y se separa en conjuntos de datos clasificados por temporalidad de producción (temporada alta y baja),

a su vez los registros son agrupados por zona de producción (norte, centro y sur), dichos registros se conforman por atributos recabados durante los años 2014-2018, donde se contempla a los estados de Veracruz y Tabasco como los principales productores de limón persa (Fernández, 2018).

3.1.5 Procedimientos

El procedimiento para recolectar la información se realizó directamente de fuente documental secundaria (base de datos histórica de proveeduría), el formato de la base de datos es un documento digital en hoja de cálculo del cual se generaron seis conjuntos de datos que corresponden a las tres zonas de producción en las dos temporalidades. Los archivos generados son convertidos a formato de valores separado por comas (csv)¹⁷ para ser tratados mediante técnicas de minería de datos en la plataforma de software para el aprendizaje automático *WEKA* donde se define el origen de cada conjunto de datos para implementar el clasificador (*classifier*) en este caso la selección combinada de clasificadores e hiperparametrización, seleccionando el atributo (*proveedor*) como clase objetivo y posteriormente se define el modo de entrenamiento (*test options*). Los tres métodos de prueba para el clasificador seleccionado que son objeto de estudio del presente trabajo se describen a continuación:

Validación cruzada, donde el conjunto de n datos se divide en $n \text{ folds} = 10$, por cada parte se construye el clasificador con las $n \text{ folds} - 1$ partes restantes y se prueba con esa, así por cada una de las n particiones.

Percentage Split(66%), en este método de prueba se define el *percentage Split* = 66% con el que se construirá el clasificador (datos de entrenamiento) y con la parte restante se probará (datos de prueba).

Muestra representativa, cuando se hace este tipo de muestra, se debe preguntar: dado que una población es de N tamaño, ¿Cuál es el menor número de unidades muestrales (personas, organizaciones, etc.) necesarias para conformar la muestra (n) que asegure

¹⁷ csv por sus siglas en Inglés. *Comma-separated values*

un determinado nivel de error estándar?, la respuesta a esta pregunta busca encontrar una muestra que sea representativa del universo o población con cierta posibilidad de error (se pretende minimizar) y nivel de confianza (maximizar), así como la probabilidad. El porcentaje estimado de la muestra es la probabilidad de ocurrencia del fenómeno (representatividad de la muestra *versus* no representatividad, la cual se estima sobre marcos de muestreo previos o se define, la certeza total siempre es igual a uno, las posibilidades a partir de esto son p de que sí ocurra y q de que no ocurra ($p + q = 1$). Cuando no se tiene marcos de muestreo previo, se usa un porcentaje estimado de 50%, es decir, se asume que $p = q = 50\%$, (resulta lo más común, particularmente cuando se selecciona por vez primera una muestra en una población). Finalmente, el nivel deseado de confianza es el complemento del error máximo aceptable (porcentaje de “acertar en la representatividad de la muestra”), si el error elegido es de 5%, el nivel deseado de confianza será de 95%, (los niveles más comunes son de 95 y 99%).(Hernández, Fernández, & Baptista, 2010). Para obtener el valor de Z_{α} se asigna el valor de la tabla de distribución normal para Z según el nivel de confianza deseado, es decir, para sustituir en la Ecuación 3.1 Cálculo de la muestra de población finita) se coloca en el lugar de Z_{α} el valor constante del nivel (que se obtiene de la tabla de la distribución normal estándar $N(0,1)$ y no el porcentaje deseado de confianza. En la Tabla 3.1 se muestran los valores para los niveles de confianza Z más utilizados y su correspondiente nivel de significación α .

*Tabla 3.1 Valores frecuentemente utilizados de la distribución normal para Z .
Elaboración propia. Fuente (Dra & Gladys, 2011)*

Z_{α}	2.576	2.24	1.96	1.645	1.44	1.28
Nivel de confianza	99%	97.50%	95%	90%	85%	80%

En el presente trabajo se utiliza este tipo de muestreo como modo de entrenamiento donde se requiere determinar la muestra final n del conjunto de datos N , en el cual se define un porcentaje con el que se construirá el clasificador atendiendo a la Ecuación 3.1 Cálculo de la muestra de población finita) asumiendo un nivel deseado de confianza Z_{α} del 95% con una probabilidad p de 50% y un nivel de error del 5% para cada conjunto de datos.

$$n = \frac{NZ\alpha^2 p*q}{d^2*(N-1) + Z\alpha^2*p*q}$$

Ecuación 3.1 Cálculo de la muestra de población finita

Donde:

- N Representa el tamaño de la población (cantidad total de tuplas o registros del conjunto de datos)
- Z Nivel deseado de confianza, $1 - \alpha$ (95%)
- Z_α Valor obtenido de la tabla de distribución normal estándar cuando el nivel de significación $\alpha = .05$ (1.96)
- p Probabilidad de éxito o porcentaje esperado de la muestra (.5)
- q Probabilidad de fracaso, es decir, $1 - p$ (.5)
- d Precisión, error máximo admisible en términos de precisión (.05)

En la Tabla 3.2 se muestran los valores que sirven de apoyo para el cálculo de la muestra para cada conjunto de datos del caso de estudio.

Tabla 3.2 Tabla de apoyo para el cálculo del tamaño de la muestra con nivel de confianza = 95%.
Fuente: Elaboración propia.

Z	α	$Z\alpha$	$Z\alpha^2$	d	d^2
95%	0.05	1.96	3.8416	0.05	0.0025

La información del origen de datos, el tamaño de la muestra calculado a través de la sustitución en la Ecuación 3.1 y el porcentaje para asignar como valor en el muestreo bajo el criterio *percentage split* se establecen en la Tabla 3.3.

Tabla 3.3 Valores de la población y muestra obtenida para cada conjunto de datos.
Fuente: Elaboración propia.

Origen de datos	N	n	<i>percentage split</i>
alta_norte	336	180	46.30%
alta_centro	476	213	55.20%
alta_sur	448	208	53.50%
baja_norte	336	180	46.30%
baja_centro	1792	317	82.30%
baja_sur	448	208	53.50%

3.1.6 Análisis de datos

Los conjuntos de datos tratados en este caso de estudio se refieren a datos secundarios los cuales se analizan de manera cuantitativa y son extraídos de la base de datos histórica de proveeduría como parte del trabajo de otro autor. Después de realizar la clasificación para cada conjunto de datos según el modo de entrenamiento elegido, se verifican las técnicas de clasificación mediante los indicadores de rendimiento, en el cual se analizan los datos del clasificador y la matriz de confusión obtenida.

3.2 Metodología para la selección de proveedores

Existen múltiples enfoques que se han propuesto para resolver el problema de la priorización y captura del proceso de decisión dentro de una organización de compras (Valdivieso, 2010) . Derivado de la revisión de la literatura sobre tópicos de selección de proveeduría se concluye que el proceso se lleva a cabo mediante la selección de diversos atributos considerados por el decisor, éstos dirigidos ya sea por las políticas de la empresa o por juicio de los expertos, donde la categorización de dichos atributos está orientada por la importancia de las variables, es decir, por el peso o prioridad que éstas reciban en función del alcance o cumplimiento sobre los estándares establecidos. Algunas de las variables consideradas en un proceso de decisión multicriterio para el caso de la selección de proveedores en industrias citrícolas son las siguientes: Capacidad de respuesta, precio del producto, calidad del producto, otras características del producto, volumen de producción, temporalidad, etc. cabe señalar que existen variables cualitativas y cuantitativas, las cuales pueden ser evaluadas en diferentes contextos según el objetivo que se persiga, asimismo, existen diferentes métodos para la selección de proveedores (Castro et al., 2009): métodos de precalificación o aproximación, métodos categóricos, técnicas difusas, análisis envolvente de datos, razonamiento basado en casos, método AHP, método ASP, etc.

3.2.1 Definición del problema

En el presente trabajo, la descripción formal del problema para atender el caso de selección de proveedores en la empacadora de limones Mónica, se lleva a cabo tomando como base la inicialización de una propuesta de modelo para la planificación en la cadena de suministro (Roma,2019), la cual se especifica a continuación:

En el apartado 1.2 (*Descripción del problema*) se detalla la problemática detectada que concierne a procesos de proveeduría de la empacadora que se menciona en el párrafo anterior, para formalizar esta parte de la realidad se toma como referencia la propuesta citada donde asume que S proveedores $S = \{1...s\}$, están disponibles para suministrar los w tipos de productos requeridos, $W = \{1...w\}$, para el caso de la empacadora se toma a W como la lima persa en sus diferentes calibres. El horizonte de planificación está definido por t períodos, $T = \{1...t\}$, los valores del conjunto T corresponden entonces a las temporadas de producción de lima persa durante los meses del año (temporada alta y temporada baja).

Existe la situación en la que algunos proveedores solo puedan suministrar un subconjunto de W . Sk es el conjunto de proveedores que pueden producir y proporcionar el artículo k . Para que el producto sea considerado como parte de la propuesta es necesario $Sk > 1$ o no habrá decisión que tomar (única fuente de opción), es decir, un único proveedor, tomando en cuenta la misma situación del lado del proveedor, Wg es el conjunto de productos que el proveedor g puede producir/entregar, pero en este caso $Wg \geq 1$, ya que si incluso el proveedor solo puede ofrecer uno de los productos, este debería ser considerado en el proceso de toma de decisión.

3.2.2 Descripción del método

Se aborda el caso de estudio de la empacadora de limones Mónica basado en datos históricos de proveeduría usando la técnica de clasificación en minería de datos implementando en el software para el aprendizaje automático el método de *selección combinada de algoritmos y optimización de hiperparámetros* para la selección de proveedores. Se utiliza la herramienta para combinar los resultados de los clasificadores individuales con el objetivo de maximizar la exactitud, asegurando la ventaja del

paradigma de combinación de clasificadores en contra de los modelos individuales donde se aborda separadamente el modelo de selección y la optimización de hiperparámetros, dicho paradigma permite obtener resultados de alta calidad en un tiempo razonable y con el mínimo esfuerzo humano. Este método abarca el rango completo de clasificadores y selector de características en la herramienta utilizada, el cual se describió a detalle en el apartado 2.1.4 (*Selección de clasificadores e hiperparametrización*).

3.2.3 Identificación de atributos

La base histórica se conforma por registros de proveeduría de la empresa de cítricos en la que se identifican los atributos relevantes tomando en consideración el juicio de personas expertas apoyando su decisión en los datos históricos de proveeduría, para el caso específico de la empacadora los atributos a considerar para evaluar a los proveedores son los siguientes (Fernández, 2018).

Calidad: Categorización de la fruta en función de criterios descritos por las especificaciones del cliente los cuales se refieren a características de la fruta como: color, textura y tamaño.

Tiempo de entrega: Velocidad de respuesta con la cual el proveedor entrega la fruta.

Volumen de producción: Capacidad de atender la necesidad requerida de fruta

Confiabilidad: Grado de confianza con el cual el proveedor ha demostrado abastecer de fruta a la empresa de acuerdo a los requerimientos del cliente con base en su registro histórico.

La categorización de las variables se muestra en la Tabla 3.4, donde se observan los criterios de inclusión tomados en cuenta en el caso de estudio y su descripción.

Tabla 3.4 Operacionalización de las variables del caso de estudio.
Fuente: Elaboración propia.

Variable	Definición conceptual	Definición operacional	Indicadores	Fuente / Escala de medición	Dimensión
year	Año en el que se obtuvo la producción de lima persa.	Está dada por la variable cuantitativa que corresponde al año de producción con datos históricos de cuatro años.	2014 2015 2016 2017	Datos históricos de proveeduría	Producción
calidad	Categorización de la fruta en función de criterios descritos por las especificaciones del cliente los cuales se refieren a características de la fruta como: color, textura y tamaño.	Se mide a través de la característica <i>calibre o tamaño</i> del atributo <i>calidad de la fruta</i> , mediante seis posibles indicadores.	110 150 175 200 230 250	Datos históricos de proveeduría	Producto
especificación	Categorización de la fruta en función de características físicas de preselección.	Se mide a través del atributo <i>calidad de la fruta</i> , mediante diez posibles indicadores.	torreón R-grande R-chica T-encerado TO-encerado coleadado verde normal genérico SJ	Datos históricos de proveeduría	Producto
marzo abril mayo junio julio agosto septiembre octubre noviembre diciembre enero febrero	Mes del año en el que se obtuvo la producción de lima persa. (temporalidad alta)	Está dada por la variable cuantitativa que corresponde a la cantidad de toneladas de producción mensual durante la temporada alta con datos históricos de cuatro años.	Atributo cuantitativo que varía según la cantidad de toneladas producidas en el mes.	Datos históricos de proveeduría	Producción
	Mes del año en el que se obtuvo la producción de lima persa. (temporalidad baja)	Está dada por la variable cuantitativa que corresponde a la cantidad de toneladas de producción mensual durante la temporada baja con datos históricos de cuatro años.	Atributo cuantitativo que varía según la cantidad de toneladas producidas en el mes.	Datos históricos de proveeduría	Producción
cliente	Se refiere a las entidades que se abastecen con el producto.	Está dada por la etiqueta que corresponde al tipo de mercado al cual se dirige la producción, mediante cuatro posibles indicadores.	nacional japón EUA EUA-Europa	Datos históricos de proveeduría	Mercado nacional e internacional
proveedor	Se refiere a las entidades que abastecen a la empresa en las cuales se identifican a 23 proveedores de Lima persa en temporada baja y temporada alta.	Está dada por la etiqueta que corresponde al número de proveedor del que se obtiene la producción, mediante 23 posibles indicadores.	prov-1 prov-2 prov-3 prov-4 prov-5 prov-6 prov-7 prov-8 prov-9 prov-10 prov-11 prov-12 prov-13 prov-14 prov-15 prov-16 prov-17 prov-18 prov-19 prov-20 prov-21 prov-22 prov-23	Datos históricos de proveeduría	Huertos propios Huertos locales Huertos foráneos Centros de acopio Intermediarios
expertix	Ponderación obtenida como resultado de evaluar cada proveedor con base en una escala particular para cada criterio definido por el experto en producción de Lima persa.	Está dada por la variable cuantitativa que corresponde al indicador obtenido después de aplicar la metodología AHP determinada por una escala que permite identificar la importancia de los criterios según el criterio del experto.	Atributo cuantitativo de proveeduría/Escala SAATY, Me que varía según el resultado de la evaluación del proveedor.		Proveedor

3.2.4 Evaluación del proveedor mediante una técnica

La evaluación del proveedor según la importancia de los criterios establecida por los expertos se realiza mediante el análisis de la matriz de confusión generada y los indicadores de exactitud que son resultado de aplicar la técnica de minería de datos basada en clasificación.

Capítulo 4

Experimentos y Resultados

La fase experimental se llevó a cabo en una computadora sobre la plataforma Windows 8.1 PRO con procesador Intel ® Core ™ i3 CPU M380 a @2.53 GHz, con una memoria instalada de 4 GB y sistema operativo de 64 bits, procesador x64. Los conjuntos de datos se cargaron en la plataforma de software de aprendizaje automático *WEKA* versión estable 3.8.3 para implementar los algoritmos de clasificación como se detalló en el apartado 3.1.5.

4.1 Análisis e interpretación de los resultados

En este apartado se presentan los principales resultados después de procesar la información contenida en los *dataset* mediante la implementación de las técnicas de clasificación. Los indicadores se observan respectivamente en la Tabla 4.1, Tabla 4.2 y Tabla 4.3 aplicando los criterios de muestreo para el clasificador: *cross-validation (10 folds)*, *percentage split=66%* y muestra representativa. Para cada conjunto de datos se observa el mejor clasificador encontrado, el número de instancias correctamente e incorrectamente clasificadas y el porcentaje de exactitud del clasificador según la matriz de confusión resultante.

4.2 Análisis de los indicadores de rendimiento del clasificador

En esta sección, se muestran los resultados que se obtuvieron con la implementación del clasificador *autoweka* bajo los criterios de muestreo establecidos en el estudio, donde se destaca el mejor algoritmo clasificador y se determina el criterio de muestreo con mayor porcentaje de exactitud encontrado para cada conjunto de datos.

4.2.1 Resultados de clasificación bajo el criterio de muestreo: Cross-validation (10 folds)

En la Tabla 4.1 se pueden visualizar los resultados obtenidos en la clasificación para la temporada alta y para la temporada baja al momento de implementar el criterio de muestreo: *cross-validation*, donde se observa el mejor algoritmo clasificador para cada conjunto de datos

*Tabla 4.1 Mejor algoritmo clasificador y porcentaje de exactitud para cada conjunto de datos.
Criterio de muestreo: Cross-validation (10 folds).
Fuente: Elaboración propia*

Cross-validation: 10 folds						
DATA SET SOURCE	RESULT	BEST CLASSIFIER		CLASSIFIES INSTANCES		TOTAL INSTANCES
				CORRECTLY	INCORRECTLY	
alta_norte	autoweka	classifiers.meta	Vote	300	36	336
				89.29%	10.71%	
alta_centro	autoweka	classifiers.trees	RandomForest	337	139	476
				70.80%	29.20%	
alta_sur	autoweka	classifiers.meta	AttributeSelectedClassifier	392	56	448
				87.50%	12.50%	
baja_norte	autoweka	classifiers.trees	RandomForest	294	42	336
				87.50%	12.50%	
baja_centro	autoweka	classifiers.rules	JRip	1373	419	1792
				76.62%	23.38%	
baja_sur	autoweka	classifiers.bayes	NaiveBayes	399	49	448
				89.06%	10.94%	

4.2.2 Resultados de clasificación bajo el criterio de muestreo: Percentage split(66%).

La Tabla 4.2 muestra los resultados obtenidos en la clasificación para cada conjunto de datos bajo el criterio de muestreo: *percentage split (66%)*, donde se observa el mejor algoritmo clasificador para cada conjunto de datos.

Tabla 4.2 Mejor algoritmo clasificador y porcentaje de exactitud para cada conjunto de datos.
Criterio de muestreo: Percentage split (66%).
Fuente: Elaboración propia.

Percentage split: 66%						
DATASET SOURCE	RESULT	BEST CLASSIFIER		CLASSIFIES INSTANCES		TOTAL INSTANCES
				CORRECTLY	INCORRECTLY	
alta_norte	autoweka	classifiers.meta	RandomSubSpace	98	16	114
				85.9649%	14.0351%	
alta_centro	autoweka	classifiers.meta	RandomCommittee	105	57	162
				64.8148%	35.1852%	
alta_sur	autoweka	classifiers.trees	J48	128	24	152
				84.2105%	15.7895%	
baja_norte	autoweka	classifiers.trees	Bayes.Net	101	13	114
				88.5965%	11.4035%	
baja_centro	autoweka	classifiers.rules	JRip	457	152	609
				75.0411%	24.9589%	
baja_sur	autoweka	classifiers.bayes	NaiveBayes	137	15	152
				90.1316%	9.8684%	

4.2.3 Resultados de clasificación bajo el criterio de muestreo: Muestra representativa.

En la Tabla 4.3 se pueden visualizar los resultados obtenidos en la clasificación para la temporada alta y para la temporada baja al momento de implementar el criterio de muestreo: muestra representativa, donde se observa el mejor algoritmo clasificador para cada conjunto de datos

Tabla 4.3 Mejor algoritmo clasificador y porcentaje de exactitud para cada conjunto de datos.
Criterio de muestreo: Percentage split (Muestra representativa).
Fuente: Elaboración propia.

MUESTRA REPRESENTATIVA						
SOURCE	RESULT	BEST CLASSIFIER		CLASSIFIES INSTANCES		TOTAL INSTANCES
				CORRECTLY	INCORRECTLY	
alta_norte	autoweka	classifiers.meta	Vote	161	19	180
				89.4444%	10.5556%	
alta_centro	autoweka	classifiers.meta	RandomCommittee	143	70	213
				67.1320%	32.8638%	
alta_sur	autoweka	classifiers.meta	AttributeSelectedClassifier	175	33	208
				84.1346%	15.8540%	
baja_norte	autoweka	classifiers.trees	RandomForest	159	21	180
				88.3333%	11.6667%	
baja_centro	autoweka	classifiers.rules	JRip	235	82	317
				74.1325%	25.8675%	
baja_sur	autoweka	classifiers.bayes	NaiveBayes	183	25	208
				87.9808%	12.0192%	

La Tabla 4.4 concentra los valores de los elementos que conforman la Ecuación 3.1 (Cálculo de la muestra de población finita) y la variable calculada (n) que equivale al tamaño de la muestra representativa implementada en el clasificador, la cual sirvió de apoyo para aplicar el criterio de muestreo: muestra representativa para cada conjunto de datos.

Tabla 4.4 Cálculo del tamaño de la muestra para cada conjunto de datos.
Fuente: Elaboración propia.

		$n = \frac{NZ\alpha^2 p \cdot q}{d^2 \cdot (N-1) + Z\alpha^2 \cdot p \cdot q}$					
N Tamaño de población		Z Nivel deseado de confianza	$Z\alpha$ Valor obtenido de la tabla de distribución normal estándar	p Probabilidad de éxito o porcentaje esperado de la muestra	q Probabilidad de fracaso	d Precisión, error máximo admisible en términos de precisión	n
origen de datos	(cantidad total de tuplas o registros)	(1- α)	cuando el nivel de significación $\alpha=0.05$ (1.96)	0.5	(1-p)	0.05	
alta_norte	336	95	1.96	0.5	0.5	0.05	179
alta_centro	476	95	1.96	0.5	0.5	0.05	213
alta_sur	448	95	1.96	0.5	0.5	0.05	207
baja_norte	1792	95	1.96	0.5	0.5	0.05	316
baja_centro	336	95	1.96	0.5	0.5	0.05	179
baja_sur	448	95	1.96	0.5	0.5	0.05	207

Una vez que se realiza la tarea de minería a dichos datos (clasificación) mediante la elección del *classifier* para cada conjunto de datos bajo los parámetros establecidos según el modo de entrenamiento o criterio de muestreo (*cross-validation*, *percentage split=66%* o *percentage split: muestra representativa*), se verifican las técnicas de clasificación mediante las métricas de validación más comunes (indicadores de rendimiento del clasificador) donde se analizan principalmente los datos obtenidos de la implementación del clasificador y la matriz de confusión resultante. Los valores generados de dicha implementación se concentran en la Tabla 11 donde se muestra el mejor algoritmo clasificador (*best classifier*) correspondiente al criterio de muestreo de clasificación implementado y el nivel de exactitud o instancias correctamente clasificadas (*classifies instances correctly*) para cada conjunto de datos.

4.2.4 Comparación de resultados de clasificación para cada conjunto de datos bajo los criterios de muestreo establecidos en el estudio.

Las Tablas 4.1, 4.2 y 4.3 sirvieron para generar el siguiente concentrado (Tabla 4.5) el cual permite visualizar, para cada conjunto de datos, los mejores algoritmos encontrados en la clasificación y el rendimiento del clasificador bajo los criterios de muestreo: *cross-validation*, *percentage split* (66%) y muestra representativa.

*Tabla 4.5 Comparación de resultados mediante diferentes criterios de muestreo.
Resultados: Mejor algoritmo clasificador y porcentaje de exactitud.
Fuente: Elaboración propia.*

Mejor algoritmo clasificador y porcentaje de exactitud para cada criterio de muestreo										
DATASET SOURCE	CRITERIO DE MUESTREO						SPLIT:	CLASSIFIES INSTANCES CORRECTLY		
	CROSS-VALIDATION 10 FOLDS		PERCENTAGE SPLIT 66.66 %		MUESTRA_REPRESENTATIVA			CROSS- VALIDATION 10 FOLDS	PERCENTAGE SPLIT 66.65 %	MUESTRA REPRESENTATIVA
	ALGORITMO (BEST CLASSIFIER)									
alta_norte	classifiers.meta	Vote	classifiers.meta	RandomSubSpace	classifiers.meta	Vote	46.3%	300 89.29%	98 85.96%	161 89.44%
alta_centro	classifiers.trees	RandomForest	classifiers.meta	RandomCommittee	classifiers.meta	RandomCommittee	55.2%	337 70.80%	105 64.81%	143 67.13%
alta_sur	classifiers.meta	AttributeSelectedClassifier	classifiers.trees	J48	classifiers.meta	AttributeSelectedClassifier	53.5%	392 87.50%	128 84.21%	175 84.13%
baja_norte	classifiers.trees	RandomForest	classifiers.trees	Bayes.Net	classifiers.trees	RandomForest	46.3%	294 87.50%	101 88.60%	159 88.33%
baja_centro	classifiers.rules	JRip	classifiers.rules	JRip	classifiers.rules	JRip	82.3%	1373 76.62%	457 75.04%	235 74.13%
baja_sur	classifiers.bayes	NaiveBayes	classifiers.bayes	NaiveBayes	classifiers.bayes	NaiveBayes	53.5%	399 89.06%	137 90.13%	183 87.98%

La información de la Tabla 4.5 permite comparar los criterios de muestreo implementados (*cross-validation*, *percentage split* (66%), muestra representativa) y su respectivo porcentaje de exactitud. De acuerdo al total de instancias correctamente clasificadas en cada origen de datos se puede determinar que para el conjunto alta_norte el porcentaje de exactitud más alto corresponde al criterio: muestra representativa; para los conjuntos: alta_centro, alta_sur y baja_centro los porcentajes más altos de exactitud corresponden al criterio: *cross-validation* mientras que para los conjuntos: baja_norte y baja_sur el criterio de muestreo: *percentage split* (66%) contiene los porcentajes más altos de exactitud.

Los resultados de la comparación anterior se resumen en la Tabla 4.6 la cual muestra para cada conjunto de datos el mejor algoritmo clasificador y el porcentaje de exactitud para el criterio de muestreo donde se obtuvo mayor rendimiento.

Tabla 4.6 Mejor algoritmo clasificador y porcentaje de exactitud para el criterio de muestreo con mayor rendimiento sobre cada conjunto de datos
Fuente: Elaboración propia.

Mejor algoritmo clasificador y porcentaje de exactitud para el criterio de muestreo con mayor rendimiento				
DATASET SOURCE	ALGORITMO (BEST CLASSIFIER)		CRITERIO DE MUESTREO CON MAYOR RENDIMIENTO	EXACTITUD (%)
alta_norte	classifiers.meta	Vote	Muestra Representativa Split: 46.3%	89.44%
alta_centro	classifiers.trees	RandomForest	Cross-Validation	70.80%
alta_sur	classifiers.meta	AttributeSelectedClassifier	Cross-Validation	87.50%
baja_norte	classifiers.trees	Bayes.Net	Percentage Split 66.66 %	88.60%
baja_centro	classifiers.rules	JRip	Cross-Validation	76.62%
baja_sur	classifiers.bayes	NaiveBayes	Percentage Split 66.66 %	90.13%

4.3 Análisis de la matriz de confusión y del resumen detallado por clase

Los resultados de la clasificación se presentan como información base en la matriz de confusión, en esta se muestra en forma de matriz, en qué han consistido, concretamente, las equivocaciones del clasificador ya que en $m(i, j)$ se tiene el número de ejemplares que son de la clase i y se han clasificado como de la clase j . Por tanto, la diagonal principal contiene enteros altos y el resto de celdas números próximos a cero o cero. (Bot, 2007)

En este apartado se muestran los valores resultantes de la matriz de confusión para cada conjunto de datos tomando como base el criterio de muestreo donde se obtuvo mayor rendimiento. A continuación se muestran las matrices obtenidas durante las pruebas realizadas implementando la técnica de clasificación.

La Tabla 4.7 muestra la matriz de confusión generada bajo el criterio de muestreo con mayor rendimiento: *muestra representativa* para el origen de datos: *alta_norte*.

Tabla 4.7 Matriz de confusión para el criterio de muestreo con mayor rendimiento: muestra representativa.
 Origen de datos: alta_norte.
 Fuente: Elaboración propia.

=== Confusion Matrix ===			
a	b	c	<-- classified as
52	4	1	a = prov-19
5	56	0	b = prov-2
1	8	53	c = prov-8

Tomando de referencia a un clasificador como una frontera que separa los datos (instancias) pertenecientes a distintas clases, la matriz de confusión permite visualizar a qué clase pertenecen dichas instancias, cabe mencionar que sólo se hace la interpretación de la Tabla 4.7 partiendo de la idea que dicha interpretación es similar al resto de las matrices de confusión.

En este caso se pueden corroborar los datos de la Tabla 4.5 (*classifies instances correctly=161, 89.44%*) calculando el número de instancias correctamente clasificadas que se obtiene de la suma de la diagonal principal de la matriz (52,56,53), el total de las instancias restantes (5,1,4,8,1,0) corresponden a las que están clasificadas incorrectamente (*classifies instances incorrectly=19,10.55%*), en la matriz se observa que de la clase “a” cuatro instancias son confundidas como clase “b” y una instancia es confundida como clase “c”. De la clase “b”, cinco las confunde como “a”. De la clase “c” ocho las confunde como “b” y una como “a”.

Los valores que resultan de la matriz de confusión sirven para calcular las métricas que permiten evaluar el modelo de clasificación mismas que se obtienen aplicando las fórmulas descritas en el apartado 2.3 (Métricas de validación. Indicadores del rendimiento del clasificador) o bien consultando los resultados de la implementación de los algoritmos para cada conjunto de datos como se muestra en la Tabla 4.8.

Tabla 4.8 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:
muestra representativa. Origen de datos: alta_norte.
Fuente: Elaboración propia.

=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.912	0.049	0.897	0.912	0.904	0.860	0.976	0.951	prov-19
	0.918	0.101	0.824	0.918	0.868	0.798	0.980	0.945	prov-2
	0.855	0.008	0.981	0.855	0.914	0.878	0.991	0.975	prov-8
Weighted Avg.	0.894	0.053	0.901	0.894	0.895	0.845	0.982	0.957	

La tabla anterior permite visualizar la información generada en el software de aprendizaje después de aplicar el proceso de clasificación, en el resumen detallado por clase se encuentran medidas que se refieren a cada una de las clases las cuales permiten conocer la efectividad del clasificador.

TP hace referencia a *true positive* (observaciones que el modelo clasificó correctamente para cada clase), FP a *false positives* (observaciones que el modelo clasificó como de esa clase cuando en realidad no lo eran). La precisión se refiere fracción de ejemplares que se han clasificado como de la clase correspondiente y que, en realidad, son de esa clase, seguidamente se encuentra el *recall* (sensibilidad) que se refiere a la fracción de ejemplos de la clase de todo el conjunto que se clasifican correctamente y por último el f-measure que está dada por una combinación de la precisión y el *recall*. (Bot, 2007)

A partir de estas estadísticas que aparecen en la Tabla 4.8 se toma como referencia el valor máximo de la métrica de validación *recall* o sensibilidad, la cual corresponde a la tasa de verdaderos positivos de la clase considerada, en este caso para el conjunto de datos: alta_norte, el valor máximo del atributo *recall* permite establecer al proveedor 2 como la mejor clase, esta información coincide con el análisis de la matriz de confusión genera para el mismo conjunto de datos (Tabla 4.7) donde se infiere también al proveedor 2 como mejor clase obteniendo el máximo valor de instancias correctamente clasificadas en su diagonal principal (56 instancias).

En la Tabla 4.9 se muestra la matriz de confusión generada bajo el criterio de muestreo con mayor rendimiento: *cross-validation* para el origen de datos: *alta_centro*.

Tabla 4.9 Matriz de confusión para el criterio de muestreo con mayor rendimiento: *cross-validation*.
Origen de datos: *alta_centro*
. Fuente: Elaboración propia.

=== Confusion Matrix ===																	
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	<-- classified as
22	3	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	a = prov-1
2	21	1	0	2	0	0	0	0	0	1	0	0	0	0	1	0	b = prov-11
1	1	21	0	1	0	0	0	1	0	1	0	1	0	0	0	1	c = prov-12
2	2	0	19	3	0	0	0	0	0	0	0	1	0	0	0	1	d = prov-13
1	1	0	1	15	4	1	0	5	0	0	0	0	0	0	0	0	e = prov-14
6	0	0	0	5	16	0	0	1	0	0	0	0	0	0	0	0	f = prov-15
1	2	0	1	1	0	22	0	0	0	0	0	0	0	0	0	1	g = prov-16
1	3	0	0	1	0	1	21	0	0	0	0	0	1	0	0	0	h = prov-17
1	2	0	0	1	2	0	0	21	0	0	0	0	0	1	0	0	i = prov-18
3	3	0	0	0	1	0	0	0	21	0	0	0	0	0	0	0	j = prov-20
2	2	1	0	0	0	0	0	0	0	17	4	0	0	1	0	1	k = prov-21
4	1	0	0	0	0	1	1	0	0	2	19	0	0	0	0	0	l = prov-22
3	1	1	0	0	0	1	0	0	0	0	0	21	0	0	0	1	m = prov-23
2	3	0	0	0	1	1	0	0	0	0	0	0	16	4	0	1	n = prov-3
2	2	1	0	0	0	0	0	1	0	0	0	1	1	20	0	0	o = prov-5
1	4	0	0	1	0	0	0	1	0	0	0	0	0	0	21	0	p = prov-6
2	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	24	q = prov-7

En este caso para el conjunto de datos: *alta_centro* se observa el valor máximo en la diagonal (24 instancias), este valor indica que la clasificación considera al proveedor 7 como la mejor clase obtenida.

La tabla 4.10 muestra el resumen detallado por clase para el conjunto de datos: *alta_centro* aplicando el criterio de muestreo con mayor rendimiento: *cross-validation*.

Tabla 4.10 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:
cross-validation. Origen de datos: alta_centro.
Fuente: Elaboración propia.

=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.786	0.076	0.393	0.786	0.524	0.518	0.918	0.468	prov-1
	0.750	0.067	0.412	0.750	0.532	0.520	0.886	0.466	prov-11
	0.750	0.009	0.840	0.750	0.792	0.782	0.873	0.767	prov-12
	0.679	0.004	0.905	0.679	0.776	0.772	0.908	0.762	prov-13
	0.536	0.036	0.484	0.536	0.508	0.477	0.876	0.428	prov-14
	0.571	0.025	0.593	0.571	0.582	0.556	0.909	0.587	prov-15
	0.786	0.011	0.815	0.786	0.800	0.788	0.981	0.867	prov-16
	0.750	0.002	0.955	0.750	0.840	0.838	0.886	0.795	prov-17
	0.750	0.020	0.700	0.750	0.724	0.707	0.893	0.695	prov-18
	0.750	0.002	0.955	0.750	0.840	0.838	0.928	0.798	prov-20
	0.607	0.009	0.810	0.607	0.694	0.685	0.852	0.655	prov-21
	0.679	0.009	0.826	0.679	0.745	0.735	0.830	0.620	prov-22
	0.750	0.007	0.875	0.750	0.808	0.799	0.945	0.801	prov-23
	0.571	0.004	0.889	0.571	0.696	0.699	0.859	0.602	prov-3
	0.714	0.013	0.769	0.714	0.741	0.726	0.912	0.709	prov-5
	0.750	0.002	0.955	0.750	0.840	0.838	0.883	0.797	prov-6
	0.857	0.013	0.800	0.857	0.828	0.817	0.944	0.888	prov-7
	Weighted Avg.	0.708	0.018	0.763	0.708	0.722	0.711	0.899	0.689

En los valores generados en el resumen anterior se visualiza el valor máximo para el atributo *recall* el cual corresponde a la clase: proveedor 7, esto permite confirmar la mejor clase obtenida en la matriz de confusión para el mismo conjunto de datos (Tabla 4.9) donde el mayor número de instancias correctamente clasificadas (24) corresponde también al proveedor 7.

La Tabla 4.11 muestra la matriz de confusión generada bajo el criterio de muestreo con mayor rendimiento: *cross-validation* para el origen de datos: *alta_sur*.

Tabla 4.11 Matriz de confusión para el criterio de muestreo con mayor rendimiento:
cross-validation. Origen de datos: alta_sur.
Fuente: Elaboración propia.

=== Confusion Matrix ===				
a	b	c	d	<-- classified as
103	2	1	6	a = prov-10
3	100	8	1	b = prov-22
8	1	103	0	c = prov-4
16	6	4	86	d = prov-9

La diagonal principal de la tabla anterior permite visualizar las instancias correctamente clasificadas para el conjunto de datos: alta_sur, en la cual se observa que los valores máximos de dicha diagonal corresponden a dos clases: proveedor 10 y proveedor 4 donde se obtuvo un valor de 103 instancias correctamente clasificadas.

A continuación se muestra en la Tabla 4.12 el resumen detallado por clase para el conjunto de datos: alta_sur bajo el criterio de muestreo con mayor rendimiento: cross-validation.

Tabla 4.12 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:
cross-validation. Origen de datos: alta_sur.
Fuente: Elaboración propia.

=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.920	0.080	0.792	0.920	0.851	0.801	0.974	0.919	prov-10
	0.893	0.027	0.917	0.893	0.905	0.874	0.973	0.955	prov-22
	0.920	0.039	0.888	0.920	0.904	0.871	0.988	0.966	prov-4
	0.768	0.021	0.925	0.768	0.839	0.798	0.936	0.898	prov-9
Weighted Avg.	0.875	0.042	0.881	0.875	0.875	0.836	0.968	0.934	

Tomando como referencia los valores máximos en el indicador *recall* del resumen anterior, se obtiene como mejores clases al proveedor 10 y al proveedor 4, esta determinación coincide con los datos generados en la matriz de confusión para el conjunto de datos: alta_sur (Tabla 4.11) donde también clasifica al proveedor 10 y al proveedor 4 como sus mejores clases.

En la Tabla 4.13 se muestra la matriz de confusión generada bajo el criterio de muestreo con mayor rendimiento (*percentage-split=66%*) para el origen de datos: *baja_norte*.

Tabla 4.13 Matriz de confusión para el criterio de muestreo con mayor rendimiento: *percentage split=66%*. Origen de datos: *baja_norte*.
Fuente: Elaboración propia.

=== Confusion Matrix ===			
a	b	c	<-- classified as
29	3	1	a = prov-19
2	39	0	b = prov-2
7	0	33	c = prov-8

Los valores en la diagonal de la matriz anterior permiten determinar al proveedor 2 como la mejor clase con un valor máximo de 39 instancias correctamente clasificadas

Tabla 4.14 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento: *percentage split=66%*. Origen de datos: *baja_norte*.
Fuente: Elaboración propia.

=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.879	0.111	0.763	0.879	0.817	0.739	0.932	0.889	prov-19
	0.951	0.041	0.929	0.951	0.940	0.905	0.974	0.919	prov-2
	0.825	0.014	0.971	0.825	0.892	0.847	0.988	0.966	prov-8
Weighted Avg.	0.886	0.052	0.895	0.886	0.887	0.836	0.967	0.927	

De igual forma se observa en la Tabla 4.14 (resumen detallado por clase) que para el conjunto de datos: *baja_norte* la mejor clase es el proveedor 2 con el valor máximo en la métrica de *recall*.

La Tabla 4.15 muestra la matriz de confusión generada bajo el criterio de muestreo con mayor rendimiento (*cross-validation*) para el origen de datos: *baja_centro*.

Tabla 4.15 Matriz de confusión para el criterio de muestreo con mayor rendimiento:
cross-validation. Origen de datos: baja_centro.
Fuente: Elaboración propia

=== Confusion Matrix ===																
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<-- classified as
84	6	1	0	2	0	0	3	0	7	4	1	0	1	1	2	a = prov-1
0	91	3	0	0	2	0	0	0	5	0	0	1	1	2	7	b = prov-11
1	2	88	2	0	1	2	3	2	2	0	1	0	0	0	8	c = prov-12
0	0	1	77	8	5	0	3	1	13	0	1	1	1	1	0	d = prov-13
0	0	2	6	73	6	2	3	1	9	0	0	1	1	2	6	e = prov-14
0	1	0	3	3	80	7	1	2	6	0	2	0	0	0	7	f = prov-15
0	1	1	1	0	6	84	1	1	3	0	1	0	0	1	12	g = prov-16
2	1	2	3	0	3	1	86	1	7	2	0	0	0	1	3	h = prov-17
1	0	3	0	1	1	4	3	89	2	0	1	0	1	0	6	i = prov-18
2	0	0	2	0	0	0	4	0	96	0	1	1	0	0	6	j = prov-20
1	3	1	0	0	0	1	2	0	14	79	0	0	1	0	10	k = prov-21
1	1	3	3	1	0	1	3	2	5	0	86	1	1	0	4	l = prov-23
1	1	0	1	0	6	2	2	0	7	0	1	86	0	0	5	m = prov-3
1	6	2	0	0	0	1	0	1	3	3	0	0	85	2	8	n = prov-5
0	0	1	1	0	3	0	5	0	7	2	0	2	0	85	6	o = prov-6
0	0	1	1	1	0	0	0	0	5	0	0	0	0	0	104	p = prov-7

En la tabla anterior se puede observar que el valor máximo en la diagonal clasifica al proveedor 7 como mejor clase con un valor de 104 instancias correctamente clasificadas.

La determinación anterior se fundamenta tomando como referencia el valor máximo del indicador *recall* para el conjunto de datos: baja_centro como se muestra en la Tabla 4.16 donde clasifica al proveedor 7 como la mejor clase.

Tabla 4.16 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento:
cross-validation. Origen de datos: baja_centro.
Fuente: Elaboración propia.

=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.750	0.006	0.894	0.750	0.816	0.808	0.966	0.862	prov-1
	0.813	0.013	0.805	0.813	0.809	0.796	0.958	0.861	prov-11
	0.786	0.013	0.807	0.786	0.796	0.783	0.961	0.838	prov-12
	0.688	0.014	0.770	0.688	0.726	0.711	0.959	0.771	prov-13
	0.652	0.010	0.820	0.652	0.726	0.716	0.948	0.785	prov-14
	0.714	0.020	0.708	0.714	0.711	0.692	0.961	0.801	prov-15
	0.750	0.013	0.800	0.750	0.774	0.760	0.961	0.836	prov-16
	0.768	0.020	0.723	0.768	0.745	0.727	0.919	0.793	prov-17
	0.795	0.007	0.890	0.795	0.840	0.831	0.981	0.867	prov-18
	0.857	0.057	0.503	0.857	0.634	0.628	0.965	0.786	prov-20
	0.705	0.007	0.878	0.705	0.782	0.775	0.959	0.807	prov-21
	0.768	0.005	0.905	0.768	0.831	0.824	0.929	0.844	prov-23
	0.768	0.004	0.925	0.768	0.839	0.833	0.973	0.883	prov-3
	0.759	0.004	0.924	0.759	0.833	0.828	0.970	0.866	prov-5
	0.759	0.006	0.895	0.759	0.821	0.813	0.954	0.856	prov-6
	0.929	0.054	0.536	0.929	0.680	0.682	0.987	0.897	prov-7
	Weighted Avg.	0.766	0.016	0.799	0.766	0.773	0.763	0.959	0.835

En la Tabla 4.17 se muestra la matriz de confusión generada bajo el criterio de muestreo con mayor rendimiento (*percentage-split=66%*) para el origen de datos: *baja_sur*.

Tabla 4.17 Matriz de confusión para el criterio de muestreo con mayor rendimiento:
percentage split=66%. Origen de datos: *baja_sur*.
Fuente: Elaboración propia.

=== Confusion Matrix ===				
a	b	c	d	<-- classified as
34	0	2	3	a = prov-10
0	26	0	3	b = prov-22
0	0	34	3	c = prov-4
3	0	1	43	d = prov-9

Los valores en la diagonal principal de la matriz anterior permiten visualizar las instancias correctamente clasificadas para el conjunto de datos: *baja_sur*, donde se observa que la mejor clase corresponde al proveedor 9 con un valor de 43 instancias correctamente clasificadas.

La tabla 4.18 muestra el resumen detallado de la clasificación para el conjunto de datos: *baja_sur*, donde se visualiza que la mejor clase está determinada para el proveedor 4, sin embargo, en este caso, la discriminación la apoyaremos en los datos generados en la matriz de confusión para el mismo conjunto de datos ya que la matriz contiene la información base de donde surgen el resto de las métricas.

*Tabla 4.18 Resumen detallado por clase para el criterio de muestreo con mayor rendimiento: percentage-split=66%. Origen de datos: baja_sur.
Fuente: Elaboración propia.*

=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.872	0.027	0.919	0.872	0.895	0.860	0.990	0.963	prov-10
	0.897	0.000	1.000	0.897	0.945	0.936	0.996	0.975	prov-22
	0.919	0.026	0.919	0.919	0.919	0.893	0.992	0.974	prov-4
	0.915	0.086	0.827	0.915	0.869	0.808	0.986	0.967	prov-9
Weighted Avg.	0.901	0.040	0.906	0.901	0.902	0.866	0.991	0.970	

4.4 Análisis de las métricas de validación aplicadas

En este apartado se analizan los resultados obtenidos en la clasificación para determinar los indicadores de exactitud los cuales permiten validar la efectividad del clasificador mediante la implementación de las principales métricas de validación apoyadas en el desarrollo de las fórmulas del apartado 2.4 (Métricas de validación. Indicadores del rendimiento del clasificador) y en el resumen detallado por clase generado en el software *Weka*.

4.4.1 Resultados obtenidos mediante el desarrollo de fórmulas para el cálculo de: *overall accuracy*, precisión, sensibilidad, especificidad y *FP Rate*.

Para evaluar la clasificación y determinar el mejor criterio de muestreo implementado, se toman en cuenta los indicadores globales de rendimiento para los conjuntos de datos bajo los criterios de muestreo que resultaron con mayor desempeño mismos que se

especificaron en el apartado 4.3 (Análisis de la matriz de confusión y el resumen detallado por clase), dichos indicadores corresponden a la precisión global, la precisión por clase, la sensibilidad, la especificidad entre otros, los cuales fueron calculados ya sea mediante las fórmulas o generados en *Weka* de forma global y por clase.

A manera de referencia en la Tabla 4.19 se muestran los resultados globales de los indicadores de rendimiento que se obtuvieron bajo el desarrollo de las fórmulas correspondientes.

*Tabla 4.19 Valor de los indicadores de rendimiento mediante el desarrollo de fórmulas para cada conjunto de datos.
Fuente: Elaboración propia.*

Métricas de desempeño calculadas mediante el desarrollo de fórmulas					
DATASET SOURCE	OVERALL ACCURACY	PRECISIÓN	SENSIBILIDAD (RECALL) TP RATE	ESPECIFICIDAD TN RATE	1-ESPECIFICIDAD FP RATE
	=TP/N	=$\bar{X}$ (TP/TP+FP)	=$\bar{X}$ (TP/ TP+FN)	=$\bar{X}$ (TN/TN+FP)	=1-TN RATE
	INDICADORES				
alta_norte	0.894	0.901	0.895	0.947	0.053
alta_centro	0.708	0.763	0.708	0.982	0.018
alta_sur	0.875	0.881	0.875	0.958	0.042
baja_norte	0.886	0.887	0.885	0.945	0.055
baja_centro	0.766	0.799	0.766	0.984	0.016
baja_sur	0.901	0.916	0.901	0.965	0.035

4.4.2 Resultados generados en weka para las principales métricas de desempeño en clasificadores: overall accuracy, precisión, sensibilidad, FP Rate y curva roc.

En la Tabla 4.20 se pueden visualizar los resultados globales de los indicadores de rendimiento de la clasificación generados en *Weka* para cada conjunto de datos.

Tabla 4.20 Valor de los indicadores de rendimiento generados en weka para cada conjunto de datos.
Fuente: Elaboración propia.

Métricas de desempeño generadas en WEKA					
DATASET SOURCE	OVERALL ACCURACY	PRECISIÓN	SENSIBILIDAD (RECALL) TP RATE	1-ESPECIFICIDAD FP RATE	CURVA ROC
	INDICADORES				
alta_norte	0.894	0.901	0.894	0.053	0.982
alta_centro	0.708	0.763	0.708	0.018	0.899
alta_sur	0.875	0.881	0.875	0.042	0.968
baja_norte	0.886	0.895	0.886	0.052	0.967
baja_centro	0.766	0.799	0.766	0.016	0.959
baja_sur	0.901	0.906	0.901	0.04	0.991

De manera global se puede observar que la clasificación es buena tomando en cuenta el valor de la sensibilidad (recall) mediante el desarrollo de fórmulas (Tabla 4.19) y los valores generados en el software *Weka* (Tabla 4.20), están por encima del 80%.

4.5 Análisis del espacio ROC

Es importante recordar que un gráfico ROC se construye mediante la representación de los pares (1-especificidad, sensibilidad) obtenidos al considerar todos los posibles valores de corte en una prueba, correspondiendo el eje *y* a la sensibilidad y el eje *x* a (1-especificidad) de cada uno de ellos. Ambos ejes incluyen valores entre 0 y 1 (0% a 100%). A modo de referencia, en todo gráfico de curva ROC se traza una línea desde el punto 0,0 al punto 1,1, llamada diagonal de referencia o línea de no-discriminación (Cerdeira & Cifuentes, 2012)

La curva ROC nos proporciona una representación global de la exactitud en la clasificación, dicha curva es necesariamente creciente, propiedad que refleja el compromiso existente entre sensibilidad y especificidad. La exactitud de la prueba aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo. Si la discriminación fuera perfecta (100% de sensibilidad y 100% de especificidad) pasaría por dicho punto. (López & Pita, 1998)

La Tabla 4.21 muestra los datos de apoyo que permiten generar los gráficos ROC, en el cual se establecen los puntos de corte para cada conjunto de datos tomando en cuenta la clase mejor clasificada, así mismo solo a manera de referencia se muestra el valor del área ROC para la exactitud de la clasificación.

Tabla 4.21 Valores de los ejes (x,y) para construir los gráficos ROC.
Fuente: Elaboración propia.

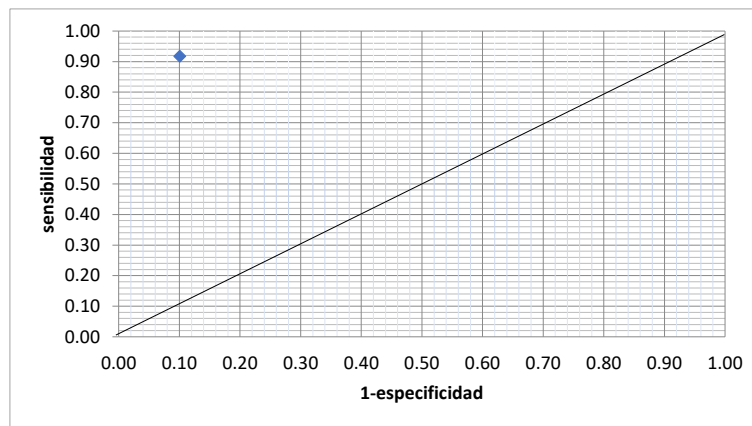
Concentrado de los pares (x,y) para la mejor clase				
DATASET SOURCE	1-ESPECIFICIDAD	SENSIBILIDAD (RECALL)	ROC AREA	CLASS
	FP RATE	TP RATE		
	INDICADORES			
alta_norte	0.101	0.918	0.980	prov-2
alta_centro	0.013	0.857	0.944	prov-7
alta_sur	0.080	0.920	0.974	prov-10
	0.039	0.920	0.988	prov-4
baja_norte	0.041	0.951	0.974	prov-2
baja_centro	0.054	0.929	0.987	prov-7
baja_sur	0.026	0.919	0.992	prov-4

Los valores del área ROC permiten visualizar un buen desempeño de manera general en el proceso de clasificación ya que en todos los casos se logra maximizar el área bajo la curva ROC, con un valor cercano a 1, esto indica la calidad que existe en el clasificador.

A continuación se presentan los gráficos ROC donde el método que se implementó para construir el espacio ROC es el método empírico que consiste simplemente en representar el par (*FP RATE*, *TP RATE*) – es decir el par (1-especificidad, sensibilidad) (López & Pita, 1998) del valor que define la exactitud de la clasificación en la mejor clase para cada conjunto de datos en este estudio.

4.5.1 Gráfico ROC para el conjunto de datos: alta_norte

Con base en la información concentrada en la Tabla 4.21 se toma como referencia las métricas obtenidas en la clasificación del conjunto de datos alta norte, para generar el gráfico que se representa en la Imagen 4.1, donde se utilizó el par (.101, .918) para la clase proveedor 2.

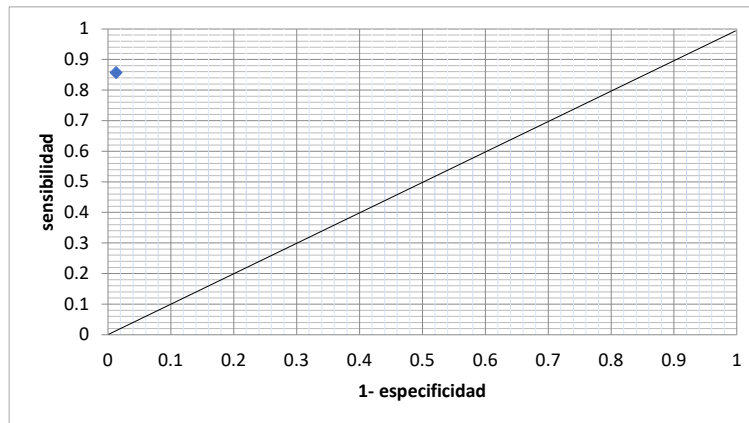


*Imagen 4.1 Gráfico ROC para la clase: proveedor 2.
Origen de datos: alta_norte.
Fuente: Elaboración propia.*

En el gráfico anterior se puede visualizar que la exactitud en la clasificación es buena para la clase proveedor 2 ya que la intersección de los indicadores se posiciona arriba de la diagonal de referencia (45°).

4.5.2 Gráfico ROC para el conjunto de datos: alta_centro

En este caso se utilizaron como referencia las métricas obtenidas en la clasificación del conjunto de datos alta centro, que consiste en el par (.013, .857) para la clase proveedor 7.

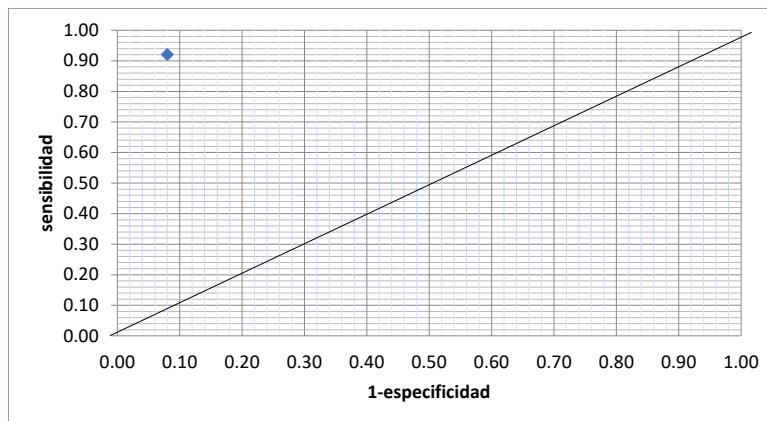


*Imagen 4.2 Gráfico ROC para la clase: proveedor 7.
Origen de datos: alta_centro.
Fuente: Elaboración propia.*

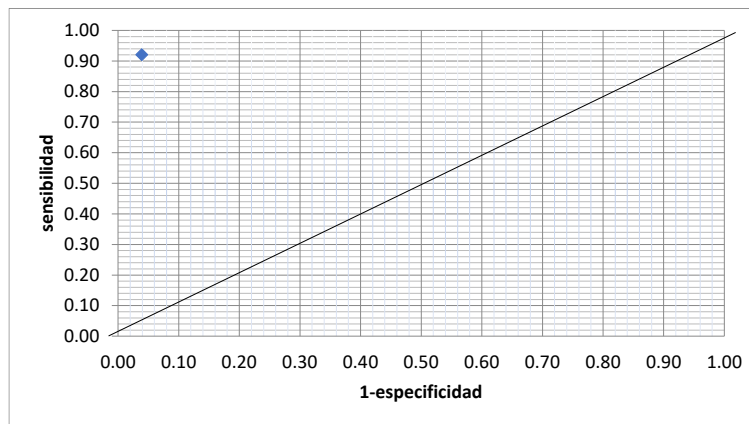
En la Imagen 4.2 se puede apreciar que la exactitud en los resultados es buena para la clase proveedor 7 ya que el punto de corte se posiciona arriba de la diagonal de no discriminación.

4.5.3 Gráfico ROC para el conjunto de datos: alta_sur

En el conjunto de datos alta_sur se utilizaron los pares (.080), (.920) y (.039), (.920) para las mejores clases proveedor 10 y proveedor 4 respectivamente, mismos que corresponden a las métricas obtenidas en la clasificación y permiten construir los gráficos que se representan en la Imagen 4.3 y la Imagen 4.4.



*Imagen 4.3 Gráfico ROC para la clase: proveedor 10.
Origen de datos: alta_sur.
Fuente: Elaboración propia.*

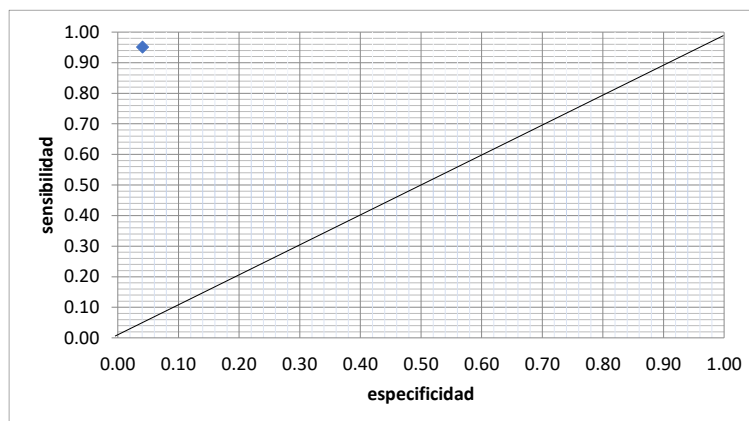


*Imagen 4.4 Gráfico ROC para la clase: proveedor 4.
Origen de datos: alta_sur.
Fuente: Elaboración propia*

Las representaciones anteriores permiten apreciar que la exactitud en la clasificación es buena ya que los puntos en las Imágenes 4.3 y 4.4. se encuentran por encima de la diagonal de referencia o diagonal de no-discriminación.

4.5.4 Gráfico ROC para el conjunto de datos: baja_norte

La imagen 4.5 representa el gráfico para los puntos (.041, .951), tomando como referencia las métricas obtenidas en la clasificación del conjunto de datos baja norte, para la clase proveedor 2.

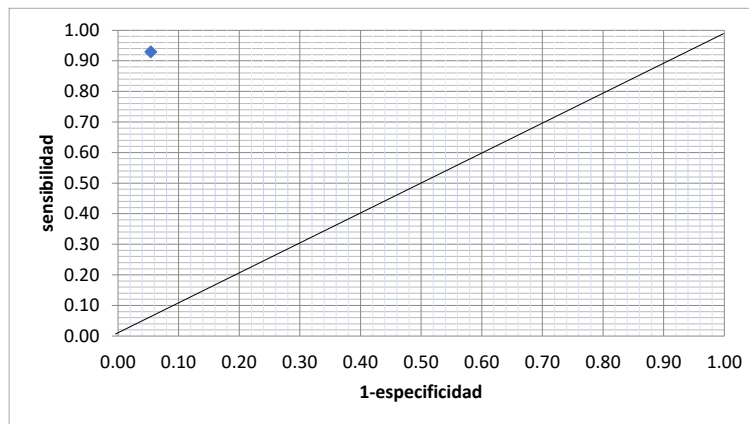


*Imagen 4.5 Gráfico ROC para la clase: proveedor 2.
Origen de datos: baja_norte.
Fuente: Elaboración propia.*

De acuerdo al gráfico anterior se puede apreciar que la exactitud en la clasificación es buena, ya que la intersección de los puntos queda arriba de la diagonal imaginaria.

4.5.5 Gráfico ROC para el conjunto de datos: baja_centro

Derivado de las métricas obtenidas en la clasificación del conjunto de datos para la clase proveedor 7 se utilizaron los indicadores (.054, .929), para generar el gráfico representado en la Imagen 4.6.

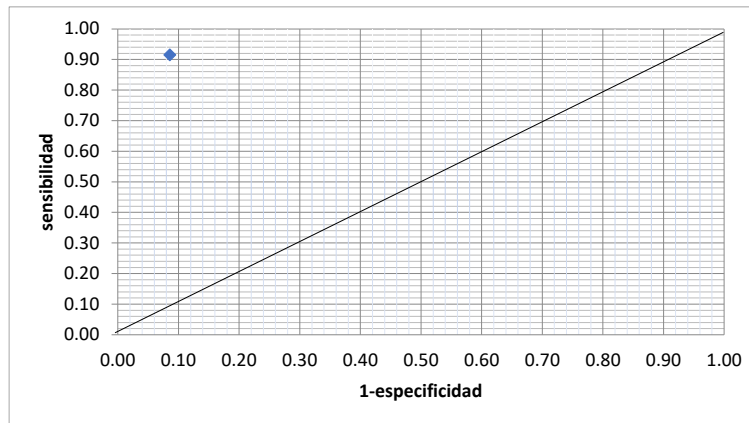


*Imagen 4.6 Gráfico ROC para la clase: proveedor 7.
Origen de datos: baja_centro.
Fuente: Elaboración propia.*

El gráfico anterior permite determinar que la exactitud en los resultados es buena para la clase proveedor 7 ya que el punto de corte se posiciona arriba de la diagonal de no discriminación.

4.5.6 Gráfico ROC para el conjunto de datos: baja_sur

Tomando como referencia las métricas obtenidas en la clasificación del conjunto de datos baja_sur, se utilizó el par (.026,.919) para la clase proveedor 4, lo cual permite construir el gráfico que se representa en la Imagen 4.7.



*Imagen 4.7 Gráfico ROC para la clase: proveedor 4.
Origen de datos: baja_sur.
Fuente: Elaboración propia.*

En el gráfico anterior se puede visualizar que la exactitud en la clasificación es buena para la clase proveedor 4 ya que la intersección de los indicadores se posiciona arriba de la diagonal de referencia (45°).

4.6 Indicadores utilizados para la discriminación de clases (proveedores)

Después de analizar los valores de la matriz de confusión generada para cada conjunto de datos bajo el criterio de muestreo con mayor rendimiento, se determinó cuál fue la clase con mayor número de instancias clasificadas correctamente, así mismo, dichos valores de la matriz sirvieron para calcular las métricas que permiten evaluar el modelo de clasificación, generando así los indicadores de precisión detallada por clase lo cual permitió realizar la discriminación de clases mediante el valor máximo para la métrica de sensibilidad (*recall*).

En la Tabla 4.22 se muestra para cada conjunto de datos, el valor máximo encontrado en la diagonal principal de la matriz de confusión y el valor máximo de sensibilidad encontrado en el resumen detallado por clase, de igual forma se muestra para ambos casos el nombre de la clase con mayor rendimiento en la clasificación.

Tabla 4.22 Valor de los indicadores con mayor rendimiento para la discriminación de clases sobre cada conjunto de datos.

Fuente: Elaboración propia.

INDICADORES PARA LA DISCRIMINACIÓN DE PROVEEDORES				
DATASET SOURCE	CONFUSION MATRIZ		DETAILED ACCURACY BY CLASS	
	CLASSIFIES INSTANCES CORRECTLY	CLASSIFIED AS	SENSIBILIDAD (RECALL) TP RATE	CLASSIFIED AS
	=TP/N		= (TP/ TP+FN)	
alta_norte	56	prov-2	0.918	prov-2
alta_centro	24	prov-7	0.857	prov-7
alta_sur	103	prov-10, prov-4	0.920	prov-10, prov-4
baja_norte	39	prov-2	0.951	prov-2
baja_centro	104	prov-7	0.929	prov-7
baja_sur	43	prov-9	0.919	prov-4

La información de la tabla anterior permite realizar la discriminación de clases determinando que para la matriz de confusión del conjunto de datos: alta_norte el valor máximo en su diagonal principal es: 56 dando como resultado al proveedor 2 como la clase mejor clasificada la cual coincide con la clase de mayor rendimiento en el indicador de sensibilidad, clasificando a proveedor 2 como la mejor clase. Para el conjunto de datos: alta_centro el valor máximo en su diagonal principal es: 24 clasificando al proveedor 7 como su mejor clase, así mismo el valor máximo del indicador de sensibilidad clasifica al proveedor 7. En la diagonal principal de la matriz de confusión para el conjunto de datos alta_sur se clasifica al proveedor 10 y al proveedor 4 como mejores clases con un valor de 103 instancias correctamente clasificadas, de igual forma, las clases con mayor rendimiento en el indicador de sensibilidad son: proveedor 10 y proveedor 4. En el conjunto de datos: baja_norte el valor máximo en la diagonal principal es: 39 clasificando al proveedor 2 como su mejor clase y el valor máximo del indicador de sensibilidad también clasifica al proveedor 2. Para el conjunto de datos: baja_centro tanto el valor máximo de la diagonal principal como el valor máximo del indicador de sensibilidad clasifican al proveedor 7 como la mejor clase. En el conjunto de datos: baja_sur la matriz de confusión clasifica en el valor más alto de su diagonal al proveedor 9 y el indicador de sensibilidad clasifica al proveedor 4 como mejor clase; en este caso se puede discriminar las clases eligiendo al proveedor 9 en el cual se toma como referencia los datos resultantes de la matriz de confusión por ser la información base en el proceso de clasificación.

4.7 Indicadores de confiabilidad para el total de las clases

En este apartado se muestran los indicadores del grado de confianza que existe para cada una de las 23 clases (proveedores de lima persa) en las dos temporalidades de estudio (alta y baja). La Tabla 4.23 muestra el concentrado del porcentaje de confiabilidad para los proveedores en temporada alta, los datos en la tabla se encuentran ordenados de manera descendente según el valor en la columna de confiabilidad, en la tabla se observa que las tuplas resaltadas corresponden al proveedor con mayor rendimiento en una determinada zona de abastecimiento (sur, norte o centro).

*Tabla 4.23 Porcentaje de confiabilidad de los proveedores.
Origen de datos: Temporada alta.
Fuente: Elaboración propia.*

Porcentaje de confiabilidad de los proveedores en temporada alta		
Proveedor	Confiabilidad (%)	Zona
prov-10	91.96%	sur
prov-4	91.96%	sur
prov-2	91.80%	norte
prov-19	91.23%	norte
prov-22	89.29%	sur
prov-7	85.71%	centro
prov-8	85.48%	norte
prov-1	78.57%	centro
prov-16	78.57%	centro
prov-9	76.79%	sur
prov-11	75.00%	centro
prov-12	75.00%	centro
prov-17	75.00%	centro
prov-18	75.00%	centro
prov-20	75.00%	centro
prov-23	75.00%	centro
prov-6	75.00%	centro
prov-5	71.43%	centro
prov-13	67.86%	centro
prov-21	60.71%	centro
prov-15	57.14%	centro
prov-3	57.14%	centro
prov-14	53.57%	centro

La información que se concentra en la tabla anterior está basada en los indicadores que arrojan las matrices de confusión para cada conjunto de datos en la temporada alta, ésta se constituye por la clase (proveedor), el porcentaje de rendimiento (confiabilidad) y la zona (ubicación); para lo cual se describen los siguientes hallazgos:

- Los proveedores foráneos 10 y 4 que corresponden a la zona sur, (con ubicación en Oaxaca y Tlaltelela, Xalapa respectivamente), presentan el mejor resultado con un porcentaje de confiabilidad de 91.96%.
- Los proveedores 2 y 19, ambos con ubicación zona norte (Papantla), tienen un rendimiento de 91.80% y 91.23% respectivamente.
- Los proveedores 22, 7 y 8 que corresponden a la zona sur, centro y norte respectivamente tienen un rendimiento superior al 80%.
- El rendimiento de los proveedores para los escenarios anteriores estima que son confiables para cubrir las necesidades de la demanda, sin embargo, para el resto de los proveedores (1, 16, 9, 11, 12, 17, 18, 20, 23, 6, 5, 13, 21, 15, 3, 14) su porcentaje de confiabilidad es inferior a 80%.

La Tabla 4.24 muestra el concentrado del porcentaje de confiabilidad para los proveedores en la temporada baja, al igual que la tabla anterior, los datos en esta tabla están ordenados de forma descendente según el valor en la columna de confiabilidad donde se observa que los renglones sombreados corresponden al proveedor con mayor rendimiento en una determinada zona (norte, centro o sur).

Tabla 4.24 Porcentaje de confiabilidad de los proveedores.
 Origen de datos: Temporada baja.
 Fuente: Elaboración propia.

Porcentaje de confiabilidad de los proveedores en temporada baja		
Proveedor	Confiabilidad (%)	Zona
prov-2	95.12%	norte
prov-7	92.86%	centro
prov-4	91.89%	sur
prov-9	91.49%	sur
prov-22	89.66%	sur
prov-19	87.88%	norte
prov-10	87.18%	sur
prov-20	85.71%	centro
prov-8	82.50%	norte
prov-11	81.25%	centro
prov-18	79.46%	centro
prov-12	78.57%	centro
prov-17	76.79%	centro
prov-23	76.79%	centro
prov-3	76.79%	centro
prov-5	75.89%	centro
prov-6	75.89%	centro
prov-1	75.00%	centro
prov-16	75.00%	centro
prov-15	71.43%	centro
prov-21	70.54%	centro
prov-13	68.75%	centro
prov-14	65.18%	centro

El concentrado que se muestra en la tabla anterior se deriva de la información base de las matrices de confusión para cada conjunto de datos en la temporada baja lo cual permite describir las siguientes observaciones:

- El proveedor 2 que corresponde a la zona norte (Papantla), presenta el mejor resultado con un porcentaje de confiabilidad de 95.12%.
- El proveedor 7 con ubicación zona centro tienen un rendimiento de 92.86%.
- Los proveedores 4 y 9 con ubicación zona sur tienen un rendimiento de 91.89% y 91.49% respectivamente.

-
- Los proveedores 19 y 8; proveedores 20 y 11; proveedores 22 y 10 que corresponden a la zona norte, centro y sur respectivamente tienen un rendimiento superior al 80%.
 - El rendimiento de los proveedores anteriores demuestra que son confiables para cubrir las necesidades de la demanda, sin embargo, para el resto de los proveedores (18, 12, 17, 23, 3, 5, 6, 1, 16, 15, 21, 13, 14) su porcentaje de confiabilidad es inferior a 80%.

De manera general las Tablas 4.23 y 4.24 permiten visualizar el rendimiento para cada uno de los proveedores de lima persa en la temporada alta y baja, de esta forma es posible determinar que los mejores indicadores de rendimiento se concentran en la temporada baja y corresponden a los proveedores 2 y 7 con valor de confiabilidad de 95.12% y 92.86% respectivamente, esto representa que para dichos proveedores en la temporada baja la capacidad de proveeduría es superior respecto a la temporada alta.

4.8 Comparación de resultados con el estado del arte

De acuerdo a los hallazgos encontrados, se acepta la hipótesis general que establece que es posible discriminar proveedores de lima persa con base en su historial de abastecimiento mediante la *selección combinada de clasificadores e hiperparametrización* con un alto grado de confianza.

Para realizar los experimentos, el clasificador se implementó en el software de aprendizaje automático WEka para combinar los resultados de los clasificadores individuales abarcando el rango completo de clasificadores. Los criterios de muestreo utilizados en el estudio fueron: cross-validation, percentage split (66%) y muestra representativa.

Estos resultados guardan relación con el estudio realizado en (Fernández, 2018) donde se aplican algoritmos para seleccionar proveedores con base en su historial de abastecimiento en el cual se abordan modelos de clasificación, pero el trabajo al cual se hace referencia difiere con el presente estudio en el sentido que en éste se propone el paradigma de combinación de clasificadores en contra de los modelos individuales de clasificación.

Las siguientes tablas muestran la comparación entre los resultados obtenidos en el estudio referido y los resultados derivados de la presente investigación, los cuales corresponden al mejor algoritmo clasificador (Tabla 4.25), mejor criterio de muestreo (Tabla 4.26), mejor clase encontrada (Tabla 4.27) y mayores porcentajes de confiabilidad de las clases (Tabla 4.28).

Tabla 4.25 Comparación de selección combinada de algoritmos con los resultados reportados en (Fernández, 2018)

temporada	dataset	meta.AutoWEKAClassifier (Estudio realizado)		Meta+DI4jMIPClassifier Estudio reportado en (Fernández,2018)
		ALGORITMO (BEST CLASSIFIER)		
ALTA	alta_norte	classifiers.meta	Vote	FilteredClassifier
	alta_centro	classifiers.trees	RandomForest	FilteredClassifier
	alta_sur	classifiers.meta	AttributeSelectedClassifier	FilteredClassifier
BAJA	baja_norte	classifiers.trees	Bayes.Net	FilteredClassifier
	baja_centro	classifiers.rules	JRip	FilteredClassifier
	baja_sur	classifiers.bayes	NaiveBayes	FilteredClassifier

Tabla 4.26 Comparación del mejor criterio de muestreo con los resultados reportados en (Fernández, 2018)

temporada	dataset	CRITERIO DE MUESTREO CON MAYOR RENDIMIENTO			
		Exactitud (%)			
		(Estudio realizado)		Estudio reportado en (Fernández,2018)	
ALTA	alta_norte	Muestra Representativa (Split: 46.3%)	89.44%	Cross-Validation (300 folds)	86.90%
	alta_centro	Cross-Validation (10 folds)	70.80%	Cross-Validation (100 folds)	73.08%
	alta_sur	Cross-Validation (10 folds)	87.50%	Cross-Validation (400 folds)	83.25%
BAJA	baja_norte	Percentage Split (66.66 %)	88.60%	Cross-Validation (200 folds)	89.58%
	baja_centro	Cross-Validation (10 folds)	76.62%	Cross-Validation (200 folds)	76.97%
	baja_sur	Percentage Split (66.66 %)	90.13%	Cross-Validation (200 folds)	90.62%

Tabla 4.27 Comparación de la mejor clase (proveedor) con los resultados reportados en (Fernández, 2018)

temporada	dataset	MEJOR CLASE (PROVEEDOR)			
		Confiabilidad (%)			
		(Estudio realizado)		Estudio reportado en (Fernández,2018)	
ALTA	alta_norte	PROV-2	91.80%	PROV-8	91.10%
	alta_centro	PROV-7	85.71%	PROV-7	79.30%
	alta_sur	PROV-10 PROV-4	91.96%	PROV-22	84.80%
BAJA	baja_norte	PROV-2	95.12%	PROV-8	92.00%
	baja_centro	PROV-7	92.86%	PROV-7	82.70%
	baja_sur	PROV-4	91.89%	PROV-22	93.80%

Tabla 4.28 Comparación de los mayores porcentajes de confiabilidad de los proveedores con los resultados reportados en (Fernández, 2018)

temporada	CONFIABILIDAD DE LOS PROVEEDORES			
	Confiabilidad (%)			
	(Estudio realizado)		Estudio reportado en (Fernández,2018)	
ALTA	PROV-10	91.96%	PROV-8	91.10%
	PROV-4	91.96%	PROV-2	87.50%
	PROV-2	91.80%	PROV-19	85.70%
	PROV-19	91.23%	PROV-22	84.80%
	PROV-22	89.29%	PROV-4	83.90%
BAJA	PROV-2	95.12%	PROV-22	93.80%
	PROV-7	92.86%	PROV-8	92.00%
	PROV-4	91.89%	PROV-4	91.10%
	PROV-9	91.49%	PROV-2	89.30%
	PROV-22	89.66%	PROV-9	89.30%

Capítulo 5

Conclusión y Trabajo Futuro

5.1 Conclusión

En el estudio realizado sobre la información generada en el ámbito de proveeduría de la exportadora de Lima persa “Mónica, S.A. de C.V”, de Martínez de La Torre, Veracruz; se llegó a la conclusión general que es posible discriminar proveedores de lima persa con base en su historial de abastecimiento mediante la *selección combinada de clasificadores e hiperparametrización*.

El presente trabajo permitió obtener y analizar la base de datos histórica de proveeduría y a su vez agrupar los datos con base en la temporada y zona de abastecimiento para establecer un formato estandarizado los cuales fueron procesados en la plataforma de software para el aprendizaje automático (*Weka*), donde se implementó el método de clasificación *selección combinada de algoritmos e hiperparametrización*, dicha clasificación permitió obtener resultados los cuales fueron analizados para validar su efectividad mediante la implementación de las principales métricas de evaluación de clasificadores, los indicadores fueron analizados directamente de los resultados generados en *Weka* y a su vez mediante el desarrollo de las fórmulas de validación de clasificadores las cuales se utilizaron para generar los indicadores de especificidad y sensibilidad o precisión global.

Los criterios de muestreo implementados en el clasificador fueron: *cross-validation*, *percentage split* (66%) y *percentage split* (muestra representativa). La evaluación del modelo de clasificación permitió determinar el criterio de muestreo con mayor rendimiento dando como resultado que para el conjunto de datos: alta_norte el porcentaje más alto

de exactitud está dado para el criterio: muestra representativa (*split*:46.3%) con un valor de 89.44%. En los conjuntos de datos: *alta_centro*, *alta_sur* y *baja_centro* el mejor criterio de muestreo está determinado como *cross-validation*, con un porcentaje de 70.80%, 87.50% y 76.62% de manera respectiva; mientras que para los conjuntos de datos: *baja_norte* y *baja_sur* el criterio de muestreo con mayor rendimiento es *percentage split* (66%), con un porcentaje de 88.60% y 90.13% respectivamente.

Los valores resultantes de la matriz de confusión en el proceso de clasificación y las métricas que se implementaron para evaluar el modelo en cada conjunto de datos permitieron realizar la discriminación de clases y así mismo obtener los indicadores del grado de confianza que existe para cada una de las 23 clases (proveedores de lima persa) en las dos temporalidades de estudio (alta y baja), lo que a su vez permitió identificar al proveedor con mayor rendimiento en una determinada zona de abastecimiento (norte, centro o sur). Con el fundamento anterior se puede inferir que para el conjunto de datos: *alta_norte* la mejor clase es el proveedor 2; para el conjunto de datos: *alta_centro* la clase mejor clasificada es el proveedor 7; las clases con mayor rendimiento para el conjunto de datos: *alta_sur* son el proveedor 10 y el proveedor 4; en el conjunto de datos: *baja_norte* el proveedor 2 es la mejor clase; para el conjunto de datos: *baja_centro* el proveedor 7 es el mejor clasificado, por último, para el conjunto de datos: *baja_sur* la mejor clase es el proveedor 9.

Tomando como referencia el concentrado del porcentaje de confiabilidad para los proveedores en *temporada alta* se determinó que los proveedores 10 y 4 que corresponden a la zona sur, presentan el mejor rendimiento con un porcentaje de confiabilidad de 91.96% y en *temporada baja* el proveedor 2 que corresponde a la zona norte presenta el mejor resultado con un porcentaje de confiabilidad de 95.12%. Con base en los resultados sobre el rendimiento de los proveedores en las dos temporadas, es importante reconocer que el proveedor 14 es la clase que obtuvo menor rendimiento, por otra parte, el proveedor 4 y el proveedor 2 se encuentran bien clasificados como proveedores con alto nivel de confiabilidad tanto en la temporada alta como en la temporada baja, con un porcentaje superior a 90%; cabe mencionar que el proveedor por excelencia con mayor índice de rendimiento es el *proveedor 2* con una

confiabilidad de 95.12% lo cual corresponde a su capacidad de proveeduría en temporada baja.

A manera de conclusiones se puede mencionar que este estudio permitió obtener los siguientes resultados:

- Se generaron los conjuntos de datos agrupados con base en la temporada y zona de abastecimiento.
- Se encontró el mejor algoritmo clasificador para cada conjunto de datos.
- Se determinó el criterio de muestreo con mayor rendimiento en el proceso de clasificación.
- Se realizó la discriminación de clases tomando en cuenta los valores resultantes de la matriz de confusión y las métricas de evaluación de desempeño implementadas.
- Se obtuvo un concentrado con los indicadores del grado de confianza que existe para cada proveedor en las dos temporalidades de estudio (alta y baja).
- Se identificó al proveedor con mayor rendimiento en una determinada zona de abastecimiento (norte, centro o sur).

El desarrollo de la presente tesis y los resultados obtenidos pueden ser vistos como una técnica la cual permite obtener cierto grado de confianza al momento de elegir un proveedor para un pedido determinado de la empacadora en mención.

5.2 Trabajo futuro

Los indicadores de confiabilidad que se obtuvieron con el desarrollo de este trabajo mediante la *selección combinada de algoritmos e hiperparametrización* aplicados a la base de datos histórica de proveeduría se consideran porcentajes aceptables, sin embargo, éstos pueden ser objeto de estudio para investigaciones futuras, lo cual representa un área de oportunidad que permite ser estudiada y atendida para fortalecer un nuevo trabajo de investigación integrando alguno de los siguientes aspectos:

-
- Incrementar el histórico de datos del *dataset*, ampliando la investigación para obtener y usar la base de datos actualizada.
 - Realizar la clasificación mediante la selección de atributos explorando algoritmos diferentes al que se propone en el trabajo.
 - Implementar nuevos clasificadores según se actualiza el estado del arte.
 - Identificar otros casos de estudio afines a la implementación de base de datos históricas, para estudiar el comportamiento del proceso realizado en otro escenario.
 - Modelar formalmente el proceso implementado, para generalizarlo.
 - Automatizar el proceso llevado a cabo, mediante el desarrollo de una herramienta de software que permita la discriminación de proveedores con alto grado de confianza.

Referencias

- Aguilar, M. A. S. (n.d.). *Toma de Decisiones con Criterios Múltiples: un resumen conceptual*.
- Alcántara, R. (n.d.). El marco metodológico de la tesis ¿Cómo elaborarlo. Retrieved from https://www.academia.edu/31681715/El_marco_metodológico_de_la_tesis_Cómo_elaborarlo
- Bot, J. A. (2007). Herramientas de Minería de datos : WEKA (Waikato Environment for Knowledge Analysis), 1–9.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*. [https://doi.org/https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/https://doi.org/10.1016/S0031-3203(96)00142-2)
- Castro, S., Ariel, W., Gómez, C., Danilo, Ó., Franco, O., Fernanda, L., ... De, S. (2009). Disponible en: <http://www.redalyc.org/articulo.oa?id=20511730008>.
- Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. *Revista Chilena de Infectología*, 29(2), 138–141. <https://doi.org/10.4067/S0716-10182012000200003>
- Corso, I., & Lorena, C. (2007). Aplicación de algoritmos de clasificación supervisada usando, 11. Retrieved from http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf
- Díaz-Barrios, H., Alemán-Rivas, Y., Cabrera-Hernández, L., Morales-Hernández, A., Chávez-Cárdenas, M. del C., & Casas-Cardoso, G. M. (2015). *Revista Cubana de ciencias informáticas RCCi = Cuban journal of computer science. Revista Cubana de Ciencias Informáticas* (Vol. 9). Univ. de las Ciencias Informaticas (UCI). Retrieved from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992015000400012
- Dra, M., & Gladys, B. C. (2011). Volumen 10 2011 Email : Rev . Med . Act . Clin @ gmail . com
Página 476 Revista de Actualización Clínica, 476–480.
- Efron, B., & Tibshirani, R. (n.d.). Cross-Validation and the Bootstrap : Estimating the Error Rate of a Prediction Rule 1 Introduction, 1–28.
- Fernández, M. Y. G. (2018). *Evaluación de proveedores para la empacadora limones Mónica S.A. de C.V. basado en la técnica Deep Learning*. Instituto Tecnológico Superior de Misantla.
- Hernández, Fernández, & Baptista. (2010). *Metodología de la Investigación*.
- Javier, J., & Navarro, L. (2001a). Organización jerárquica de las técnicas de Aprendizaje Automático.
- Javier, J., & Navarro, L. (2001b). Selección de atributos en aprendizaje automático basada en teoría de la información.
- Kohavi, R., & Provost, F. (1998). Glossary of Terms Journal of Machine Learning. Retrieved June 1, 2018, from <http://robotics.stanford.edu/~ronnyk/glossary.html>
- Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2016). Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, 17, 1–5.
- Lambert, G. F., Lasserre, A. A. A., Olivia, B., Rivera, I., & Montalvo, N. (n.d.). Una propuesta basada en técnicas de IA para seleccionar proveedores en una exportadora de cítricos, 1–6.

-
- Lindauer, M., & Hutter, F. (2014). AutoFolio : Algorithm Configuration for Algorithm Selection. *Proceedings of the International Symposium on Wireless Communication Systems*, 1(5), 26–80.
- López, I., & Pita. (1998). Curvas ROC. *Cad Aten Primaria*, 5(4), 229–235.
- Machado, J. J. G., & Garrido, N. P. (2001). La selección de proveedores internacionales.
- Michie, E. D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine Learning , Neural and Statistical Classification.
- Mitchell, T. M. (2017). CHAPTER 14 Key Ideas in Machine Learning, 1–11.
- Mohr, F., Wever, M., & Hüllermeier, E. (2018). Automated Machine Learning Service Composition. Retrieved from <http://arxiv.org/abs/1809.00486>
- Moreno Ribas, A. (1994). *Aprendizaje automático*. Retrieved from <https://upcommons.upc.edu/handle/2099.3/36157>
- Pavón Pérez, F., & Vega Sánchez Sebastián Dormido Canto, J. A. (n.d.). Resumen de la tesis: Generación de Conocimiento basado en Aprendizaje Automático y Aplicación en Diferentes Sectores AIPAKA (Artificial Intelligence in a Process for Automated Knowledge Acquisition and Applications).
- Pe, A., & Verdejo, F. (2009). Comparacion de Enfoques para Evaluar la Validacion de Repuestas, 43, 277–285.
- Segura Sánchez, J. (2005). Una larga polémica: el tâtonnement walrasiano. *Investigaciones de Historia Economica*, 1(2), 75–104. [https://doi.org/10.1016/S1698-6989\(05\)70003-5](https://doi.org/10.1016/S1698-6989(05)70003-5)
- Silva, J. L. (2015). Aprendizaje supervisado eficiente para el análisis de datos geoespaciales a gran escala, (c).
- Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2012a). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. <https://doi.org/10.1145/2487575.2487629>
- Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2012b). *Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms*. <https://doi.org/10.1145/2487575.2487629>
- Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA, (August), 847. <https://doi.org/10.1145/2487575.2487629>
- Valdivieso, I. C. (2010). Culcyt // Modelado SELECCIÓN DE PROVEEDORES USANDO EL MÉTODO MOORA, (40).
- Vianchá Sánchez, Z. H. (2014). Models and configurations of supply chains in perishable goods. *Ingeniería y Desarrollo*, 32(1), 138–154. <https://doi.org/10.14482/inde.32.1.4577>

Anexos

Experimentos realizados en weka implementando meta.AutoWEKAClassifier

Anexo 1 Conjunto de datos: alta_norte bajo el criterio de muestreo: cross-validation (10 folds).

```
Time taken to build model: 791.6 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      300           89.2857 %
Incorrectly Classified Instances    36           10.7143 %
Kappa statistic                    0.8393
Mean absolute error                 0.0861
Root mean squared error             0.2295
Relative absolute error             19.3613 %
Root relative squared error         48.6863 %
Total Number of Instances          336

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.866   0.045   0.907     0.866   0.886     0.831   0.965    0.905    prov-19
                0.938   0.080   0.854     0.938   0.894     0.839   0.969    0.959    prov-2
                0.875   0.036   0.925     0.875   0.899     0.851   0.968    0.960    prov-8
Weighted Avg.   0.893   0.054   0.895     0.893   0.893     0.840   0.968    0.941

=== Confusion Matrix ===

  a   b   c  <-- classified as
97   9   6 |  a = prov-19
5 105   2 |  b = prov-2
5   9  98 |  c = prov-8
```

Anexo 2 Conjunto de datos: alta_centro bajo el criterio de muestreo: cross-validation (10 folds).

```
Time taken to build model: 850.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      337           70.7983 %
Incorrectly Classified Instances    139           29.2017 %
Kappa statistic                    0.6897
Mean absolute error                 0.0542
Root mean squared error             0.1674
Relative absolute error             48.932 %
Root relative squared error         71.1445 %
Total Number of Instances          476
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Class
	0.786	0.076	0.393	0.786	0.524	0.518	0.918	0.468	prov-1
	0.750	0.067	0.412	0.750	0.532	0.520	0.886	0.466	prov-11
	0.750	0.009	0.840	0.750	0.792	0.782	0.873	0.767	prov-12
	0.679	0.004	0.905	0.679	0.776	0.772	0.908	0.762	prov-13
	0.536	0.036	0.484	0.536	0.508	0.477	0.876	0.428	prov-14
	0.571	0.025	0.593	0.571	0.582	0.556	0.909	0.587	prov-15
	0.786	0.011	0.815	0.786	0.800	0.788	0.981	0.867	prov-16
	0.750	0.002	0.955	0.750	0.840	0.838	0.886	0.795	prov-17
	0.750	0.020	0.700	0.750	0.724	0.707	0.893	0.695	prov-18
	0.750	0.002	0.955	0.750	0.840	0.838	0.928	0.798	prov-20
	0.607	0.009	0.810	0.607	0.694	0.685	0.852	0.655	prov-21
	0.679	0.009	0.826	0.679	0.745	0.735	0.830	0.620	prov-22
	0.750	0.007	0.875	0.750	0.808	0.799	0.945	0.801	prov-23
	0.571	0.004	0.889	0.571	0.696	0.699	0.859	0.602	prov-3
	0.714	0.013	0.769	0.714	0.741	0.726	0.912	0.709	prov-5
	0.750	0.002	0.955	0.750	0.840	0.838	0.883	0.797	prov-6
	0.857	0.013	0.800	0.857	0.828	0.817	0.944	0.888	prov-7
Weighted Avg.	0.708	0.018	0.763	0.708	0.722	0.711	0.899	0.689	

=== Confusion Matrix ===

```

a b c d e f g h i j k l m n o p q  <-- classified as
22 3 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 | a = prov-1
2 21 1 0 2 0 0 0 0 0 0 1 0 0 0 0 1 | b = prov-11
1 1 21 0 1 0 0 0 0 1 0 1 0 1 0 0 0 | c = prov-12
2 2 0 19 3 0 0 0 0 0 0 0 0 1 0 0 0 | d = prov-13
1 1 0 1 15 4 1 0 5 0 0 0 0 0 0 0 0 | e = prov-14
6 0 0 0 5 16 0 0 1 0 0 0 0 0 0 0 0 | f = prov-15
1 2 0 1 1 0 22 0 0 0 0 0 0 0 0 0 1 | g = prov-16
1 3 0 0 1 0 1 21 0 0 0 0 0 1 0 0 0 | h = prov-17
1 2 0 0 1 2 0 0 21 0 0 0 0 0 1 0 0 | i = prov-18
3 3 0 0 0 1 0 0 0 21 0 0 0 0 0 0 0 | j = prov-20
2 2 1 0 0 0 0 0 0 0 17 4 0 0 1 0 1 | k = prov-21
4 1 0 0 0 0 1 1 0 0 2 19 0 0 0 0 0 | l = prov-22
3 1 1 0 0 0 1 0 0 0 0 0 21 0 0 0 1 | m = prov-23
2 3 0 0 0 1 1 0 0 0 0 0 0 16 4 0 1 | n = prov-3
2 2 1 0 0 0 0 0 1 0 0 0 1 1 20 0 0 | o = prov-5
1 4 0 0 1 0 0 0 1 0 0 0 0 0 0 21 0 | p = prov-6
2 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 24 | q = prov-7

```

Anexo 3 Conjunto de datos: alta_sur bajo el criterio de muestreo: *cross-validation* (10 folds).

Time taken to build model: 850.56 seconds

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      392          87.5 %
Incorrectly Classified Instances    56           12.5 %
Kappa statistic                    0.8333
Mean absolute error                 0.0978
Root mean squared error             0.2229
Relative absolute error             26.0722 %
Root relative squared error         51.4815 %
Total Number of Instances          448

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.920	0.080	0.792	0.920	0.851	0.801	0.974	0.919	prov-10
	0.893	0.027	0.917	0.893	0.905	0.874	0.973	0.955	prov-22
	0.920	0.039	0.888	0.920	0.904	0.871	0.988	0.966	prov-4
	0.768	0.021	0.925	0.768	0.839	0.798	0.936	0.898	prov-9
Weighted Avg.	0.875	0.042	0.881	0.875	0.875	0.836	0.968	0.934	

```
=== Confusion Matrix ===

  a   b   c   d  <-- classified as
103   2   1   6 |  a = prov-10
  3 100   8   1 |  b = prov-22
  8   1 103   0 |  c = prov-4
 16   6   4  86 |  d = prov-9
```

Anexo 4 Conjunto de datos: baja_norte bajo el criterio de muestreo: *cross-validation* (10 folds).

Time taken to build model: 792.73 seconds

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      294          87.5 %
Incorrectly Classified Instances    42           12.5 %
Kappa statistic                    0.8125
Mean absolute error                 0.0948
Root mean squared error             0.2163
Relative absolute error             21.3207 %
Root relative squared error         45.878 %
Total Number of Instances          336

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.813	0.040	0.910	0.813	0.858	0.796	0.965	0.951	prov-19
	0.866	0.054	0.890	0.866	0.878	0.818	0.976	0.954	prov-2
	0.946	0.094	0.835	0.946	0.887	0.829	0.986	0.958	prov-8
Weighted Avg.	0.875	0.063	0.878	0.875	0.874	0.814	0.976	0.955	

```
=== Confusion Matrix ===
```

```

a  b  c  <-- classified as
91  8 13 |  a = prov-19
 7 97  8 |  b = prov-2
 2  4 106 |  c = prov-8

```

Anexo 5 Conjunto de datos: baja_centro bajo el criterio de muestreo: cross-validation (10 folds).

```
Time taken to build model: 841.36 seconds
```

```
=== Stratified cross-validation ===
```

```
=== Summary ===
```

```

Correctly Classified Instances      1373           76.6183 %
Incorrectly Classified Instances    419           23.3817 %
Kappa statistic                    0.7506
Mean absolute error                 0.0433
Root mean squared error             0.1392
Relative absolute error             36.9173 %
Root relative squared error         57.4923 %
Total Number of Instances          1792

```

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.750	0.006	0.894	0.750	0.816	0.808	0.966	0.862	prov-1	
0.813	0.013	0.805	0.813	0.809	0.796	0.958	0.861	prov-11	
0.786	0.013	0.807	0.786	0.796	0.783	0.961	0.838	prov-12	
0.688	0.014	0.770	0.688	0.726	0.711	0.959	0.771	prov-13	
0.652	0.010	0.820	0.652	0.726	0.716	0.948	0.785	prov-14	
0.714	0.020	0.708	0.714	0.711	0.692	0.961	0.801	prov-15	
0.750	0.013	0.800	0.750	0.774	0.760	0.961	0.836	prov-16	
0.768	0.020	0.723	0.768	0.745	0.727	0.919	0.793	prov-17	
0.795	0.007	0.890	0.795	0.840	0.831	0.981	0.867	prov-18	
0.857	0.057	0.503	0.857	0.634	0.628	0.965	0.786	prov-20	
0.705	0.007	0.878	0.705	0.782	0.775	0.959	0.807	prov-21	
0.768	0.005	0.905	0.768	0.831	0.824	0.929	0.844	prov-23	
0.768	0.004	0.925	0.768	0.839	0.833	0.973	0.883	prov-3	
0.759	0.004	0.924	0.759	0.833	0.828	0.970	0.866	prov-5	
0.759	0.006	0.895	0.759	0.821	0.813	0.954	0.856	prov-6	
0.929	0.054	0.536	0.929	0.680	0.682	0.987	0.897	prov-7	
Weighted Avg.	0.766	0.016	0.799	0.766	0.773	0.763	0.959	0.835	

```
=== Confusion Matrix ===
```

```

  a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   <-- classified as
84   6   1   0   2   0   0   3   0   7   4   1   0   1   1   2 | a = prov-1
0  91   3   0   0   2   0   0   0   5   0   0   1   1   2   7 | b = prov-11
1   2  88   2   0   1   2   3   2   2   0   1   0   0   0   8 | c = prov-12
0   0   1  77   8   5   0   3   1  13   0   1   1   1   1   0 | d = prov-13
0   0   2   6  73   6   2   3   1   9   0   0   1   1   2   6 | e = prov-14
0   1   0   3   3  80   7   1   2   6   0   2   0   0   0   7 | f = prov-15
0   1   1   1   0   6  84   1   1   3   0   1   0   0   1  12 | g = prov-16
2   1   2   3   0   3   1  86   1   7   2   0   0   0   1   3 | h = prov-17
1   0   3   0   1   1   4   3  89   2   0   1   0   1   0   6 | i = prov-18
2   0   0   2   0   0   0   4   0  96   0   1   1   0   0   6 | j = prov-20
1   3   1   0   0   0   1   2   0  14  79   0   0   1   0  10 | k = prov-21
1   1   3   3   1   0   1   3   2   5   0  86   1   1   0   4 | l = prov-23
1   1   0   1   0   6   2   2   0   7   0   1  86   0   0   5 | m = prov-3
1   6   2   0   0   0   1   0   1   3   3   0   0  85   2   8 | n = prov-5
0   0   1   1   0   3   0   5   0   7   2   0   2   0  85   6 | o = prov-6
0   0   1   1   1   0   0   0   0   5   0   0   0   0  104 | p = prov-7

```

Anexo 6 Conjunto de datos: baja_sur bajo el criterio de muestreo: cross-validation (10 folds).

Time taken to build model: 771.21 seconds

```
=== Stratified cross-validation ===
```

```
=== Summary ===
```

```

Correctly Classified Instances      399          89.0625 %
Incorrectly Classified Instances     49          10.9375 %
Kappa statistic                     0.8542
Mean absolute error                  0.0671
Root mean squared error              0.1829
Relative absolute error              17.9016 %
Root relative squared error          42.2293 %
Total Number of Instances           448

```

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.839	0.024	0.922	0.839	0.879	0.842	0.978	0.960	prov-10
	0.938	0.009	0.972	0.938	0.955	0.940	0.997	0.991	prov-22
	0.893	0.048	0.862	0.893	0.877	0.836	0.988	0.970	prov-4
	0.893	0.065	0.820	0.893	0.855	0.805	0.980	0.925	prov-9
Weighted Avg.	0.891	0.036	0.894	0.891	0.891	0.856	0.986	0.962	

```
=== Confusion Matrix ===
```

```

  a   b   c   d   <-- classified as
94   1   7  10 | a = prov-10
0  105   3   4 | b = prov-22
4   0  100   8 | c = prov-4
4   2   6  100 | d = prov-9

```

Anexo 7 Conjunto de datos: alta_norte bajo el criterio de muestreo: percentage_spit (66%).

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances	98	85.9649 %
Incorrectly Classified Instances	16	14.0351 %
Kappa statistic	0.7878	
Mean absolute error	0.1009	
Root mean squared error	0.2286	
Relative absolute error	22.661 %	
Root relative squared error	48.3507 %	
Total Number of Instances	114	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.818	0.037	0.900	0.818	0.857	0.805	0.979	0.957	prov-19
	0.951	0.137	0.796	0.951	0.867	0.789	0.982	0.969	prov-2
	0.800	0.041	0.914	0.800	0.853	0.786	0.980	0.964	prov-8
Weighted Avg.	0.860	0.074	0.868	0.860	0.859	0.792	0.981	0.964	

=== Confusion Matrix ===

```
a b c <-- classified as
27 3 3 | a = prov-19
 2 39 0 | b = prov-2
 1 7 32 | c = prov-8
```

Anexo 8 Conjunto de datos: alta_centro bajo el criterio de muestreo: percentage_spit (66%).

Time taken to build model: 780.75 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances	105	64.8148 %
Incorrectly Classified Instances	57	35.1852 %
Kappa statistic	0.6253	
Mean absolute error	0.0397	
Root mean squared error	0.1612	
Relative absolute error	35.8429 %	
Root relative squared error	68.4011 %	
Total Number of Instances	162	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.020	0.813	1.000	0.897	0.892	0.999	0.979	prov-1
	0.700	0.099	0.318	0.700	0.438	0.422	0.827	0.573	prov-11
	0.545	0.053	0.429	0.545	0.480	0.441	0.943	0.811	prov-12
	0.600	0.020	0.667	0.600	0.632	0.610	0.771	0.625	prov-13
	0.444	0.039	0.400	0.444	0.421	0.386	0.691	0.253	prov-14
	0.444	0.000	1.000	0.444	0.615	0.656	0.934	0.795	prov-15
	0.909	0.026	0.714	0.909	0.800	0.790	0.952	0.915	prov-16
	0.889	0.033	0.615	0.889	0.727	0.722	0.935	0.805	prov-17
	0.625	0.019	0.625	0.625	0.625	0.606	0.922	0.777	prov-18
	0.750	0.000	1.000	0.750	0.857	0.860	0.875	0.762	prov-20
	0.250	0.038	0.143	0.250	0.182	0.162	0.608	0.269	prov-21
	0.308	0.000	1.000	0.308	0.471	0.539	0.840	0.717	prov-22
	0.800	0.007	0.889	0.800	0.842	0.834	0.892	0.812	prov-23
	0.444	0.000	1.000	0.444	0.615	0.656	0.715	0.475	prov-3
	0.700	0.020	0.700	0.700	0.700	0.680	0.834	0.509	prov-5
	0.625	0.000	1.000	0.625	0.769	0.783	0.856	0.675	prov-6
	0.700	0.000	1.000	0.700	0.824	0.829	0.994	0.911	prov-7
Weighted Avg.	0.648	0.022	0.744	0.648	0.656	0.656	0.871	0.710	

=== Confusion Matrix ===

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	<-- classified as
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	a = prov-1
0	7	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	b = prov-11
0	4	6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	c = prov-12
0	0	0	6	3	0	0	1	0	0	0	0	0	0	0	0	0	0	d = prov-13
0	3	0	0	4	0	1	0	1	0	0	0	0	0	0	0	0	0	e = prov-14
1	1	0	0	3	4	0	0	0	0	0	0	0	0	0	0	0	0	f = prov-15
1	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	g = prov-16
0	0	0	0	0	0	0	0	8	0	0	0	0	1	0	0	0	0	h = prov-17
0	0	1	0	0	0	0	0	2	5	0	0	0	0	0	0	0	0	i = prov-18
0	2	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	j = prov-20
0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	k = prov-21
1	0	1	0	0	0	1	0	1	0	5	4	0	0	0	0	0	0	l = prov-22
0	1	1	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	m = prov-23
0	1	0	0	0	0	1	0	0	0	0	0	0	4	3	0	0	0	n = prov-3
0	0	1	0	0	0	0	2	0	0	0	0	0	0	7	0	0	0	o = prov-5
0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	p = prov-6
0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	7	0	q = prov-7

Anexo 9 Conjunto de datos: alta_sur bajo el criterio de muestreo: percentage_spit (66%).

Time taken to build model: 784.56 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.06 seconds

=== Summary ===

Correctly Classified Instances	128	84.2105 %
Incorrectly Classified Instances	24	15.7895 %
Kappa statistic	0.7891	
Mean absolute error	0.0912	
Root mean squared error	0.2519	
Relative absolute error	24.191 %	
Root relative squared error	57.8316 %	
Total Number of Instances	152	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.923	0.106	0.750	0.923	0.828	0.768	0.941	0.822	prov-10
	0.897	0.049	0.813	0.897	0.852	0.817	0.950	0.875	prov-22
	0.892	0.043	0.868	0.892	0.880	0.841	0.967	0.879	prov-4
	0.702	0.010	0.971	0.702	0.815	0.768	0.911	0.873	prov-9
Weighted Avg.	0.842	0.050	0.859	0.842	0.841	0.795	0.940	0.862	

=== Confusion Matrix ===

```
a  b  c  d  <-- classified as
36  2  0  1 | a = prov-10
 0 26  3  0 | b = prov-22
 3  1 33  0 | c = prov-4
 9  3  2 33 | d = prov-9
```

Anexo 10 Conjunto de datos: baja_norte bajo el criterio de muestreo: percentage_spit (66%).

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	101	88.5965 %
Incorrectly Classified Instances	13	11.4035 %
Kappa statistic	0.8289	
Mean absolute error	0.09	
Root mean squared error	0.2372	
Relative absolute error	20.2089 %	
Root relative squared error	50.1883 %	
Total Number of Instances	114	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.879	0.111	0.763	0.879	0.817	0.739	0.932	0.889	prov-19
	0.951	0.041	0.929	0.951	0.940	0.905	0.974	0.919	prov-2
	0.825	0.014	0.971	0.825	0.892	0.847	0.988	0.966	prov-8
Weighted Avg.	0.886	0.052	0.895	0.886	0.887	0.836	0.967	0.927	

=== Confusion Matrix ===

```
a b c <-- classified as
29 3 1 | a = prov-19
 2 39 0 | b = prov-2
 7 0 33 | c = prov-8
```

Anexo 11 Conjunto de datos: baja_centro bajo el criterio de muestreo: percentage_spit (66%).

Time taken to build model: 850.48 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.15 seconds

=== Summary ===

Correctly Classified Instances	457	75.0411 %
Incorrectly Classified Instances	152	24.9589 %
Kappa statistic	0.7341	
Mean absolute error	0.0334	
Root mean squared error	0.156	
Relative absolute error	28.4992 %	
Root relative squared error	64.3918 %	
Total Number of Instances	609	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.810	0.005	0.919	0.810	0.861	0.853	0.981	0.802	prov-1
	0.750	0.014	0.750	0.750	0.750	0.736	0.909	0.596	prov-11
	0.780	0.011	0.842	0.780	0.810	0.798	0.940	0.738	prov-12
	0.683	0.018	0.737	0.683	0.709	0.689	0.866	0.546	prov-13
	0.618	0.010	0.778	0.618	0.689	0.677	0.845	0.524	prov-14
	0.733	0.135	0.220	0.733	0.338	0.350	0.799	0.174	prov-15
	0.608	0.007	0.886	0.608	0.721	0.715	0.916	0.647	prov-16
	0.763	0.007	0.879	0.763	0.817	0.808	0.952	0.739	prov-17
	0.771	0.007	0.871	0.771	0.818	0.809	0.935	0.731	prov-18
	0.743	0.007	0.867	0.743	0.800	0.791	0.918	0.684	prov-20
	0.686	0.010	0.800	0.686	0.738	0.726	0.840	0.614	prov-21
	0.667	0.002	0.963	0.667	0.788	0.791	0.882	0.692	prov-23
	0.872	0.018	0.773	0.872	0.819	0.808	0.929	0.686	prov-3
	0.783	0.000	1.000	0.783	0.878	0.877	0.926	0.823	prov-5
	0.872	0.005	0.919	0.872	0.895	0.888	0.961	0.825	prov-6
	0.906	0.009	0.853	0.906	0.879	0.872	0.973	0.837	prov-7
Weighted Avg.	0.750	0.015	0.829	0.750	0.776	0.768	0.913	0.675	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<-- classified as
34	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	a = prov-1
0	24	1	1	0	4	0	0	1	0	1	0	0	0	0	0	b = prov-11
0	1	32	0	0	6	0	0	2	0	0	0	0	0	0	0	c = prov-12
1	1	1	28	4	5	0	0	0	1	0	0	0	0	0	0	d = prov-13
0	0	1	5	21	5	0	0	0	0	0	0	2	0	0	0	e = prov-14
0	0	0	1	0	22	3	0	0	0	0	0	4	0	0	0	f = prov-15
0	1	0	0	0	15	31	0	0	0	0	1	2	0	0	1	g = prov-16
0	0	1	0	0	6	0	29	0	1	1	0	0	0	0	0	h = prov-17
0	0	0	0	0	5	0	0	27	0	2	0	0	0	1	0	i = prov-18
0	0	0	0	1	5	0	0	0	26	1	0	0	0	1	1	j = prov-20
0	0	0	0	0	2	0	4	0	2	24	0	2	0	0	1	k = prov-21
1	0	2	2	1	6	0	0	0	0	0	26	0	0	1	0	l = prov-23
1	1	0	1	0	1	1	0	0	0	0	0	34	0	0	0	m = prov-3
0	3	0	0	0	5	0	0	1	0	0	0	0	36	0	1	n = prov-5
0	1	0	0	0	3	0	0	0	0	0	0	0	0	34	1	o = prov-6
0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	29	p = prov-7

Anexo 12 Conjunto de datos: baja_sur el criterio de muestreo: percentage_spit (66%).

Time taken to build model: 768.47 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances	137	90.1316 %
Incorrectly Classified Instances	15	9.8684 %
Kappa statistic	0.8666	
Mean absolute error	0.0669	
Root mean squared error	0.1722	
Relative absolute error	17.743 %	
Root relative squared error	39.5355 %	
Total Number of Instances	152	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.872	0.027	0.919	0.872	0.895	0.860	0.990	0.963	prov-10
	0.897	0.000	1.000	0.897	0.945	0.936	0.996	0.975	prov-22
	0.919	0.026	0.919	0.919	0.919	0.893	0.992	0.974	prov-4
	0.915	0.086	0.827	0.915	0.869	0.808	0.986	0.967	prov-9
Weighted Avg.	0.901	0.040	0.906	0.901	0.902	0.866	0.991	0.970	

=== Confusion Matrix ===

```
a b c d <-- classified as
34 0 2 3 | a = prov-10
0 26 0 3 | b = prov-22
0 0 34 3 | c = prov-4
3 0 1 43 | d = prov-9
```

Anexo 13 Conjunto de datos: alta_norte bajo el criterio de muestreo: muestra representativa.

Time taken to build model: 795.5 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances	161	89.4444 %
Incorrectly Classified Instances	19	10.5556 %
Kappa statistic	0.8417	
Mean absolute error	0.0749	
Root mean squared error	0.215	
Relative absolute error	16.837 %	
Root relative squared error	45.5498 %	
Total Number of Instances	180	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.912	0.049	0.897	0.912	0.904	0.860	0.972	0.930	prov-19
	0.918	0.101	0.824	0.918	0.868	0.798	0.967	0.933	prov-2
	0.855	0.008	0.981	0.855	0.914	0.878	0.991	0.975	prov-8
Weighted Avg.	0.894	0.053	0.901	0.894	0.895	0.845	0.977	0.947	

=== Confusion Matrix ===

```

a  b  c  <-- classified as
52  4  1 | a = prov-19
 5 56  0 | b = prov-2
 1  8 53 | c = prov-8

```

Anexo 14 Conjunto de datos: alta_centro bajo el criterio de muestreo: muestra representativa.

Time taken to build model: 817.35 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances	143	67.1362 %
Incorrectly Classified Instances	70	32.8638 %
Kappa statistic	0.6503	
Mean absolute error	0.0391	
Root mean squared error	0.1437	
Relative absolute error	35.2138 %	
Root relative squared error	60.9217 %	
Total Number of Instances	213	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.563	0.000	1.000	0.563	0.720	0.737	0.914	0.661	prov-1
	0.769	0.000	1.000	0.769	0.870	0.871	0.972	0.818	prov-11
	1.000	0.255	0.203	1.000	0.338	0.389	0.990	0.939	prov-12
	0.667	0.000	1.000	0.667	0.800	0.808	0.771	0.694	prov-13
	0.533	0.015	0.727	0.533	0.615	0.599	0.937	0.529	prov-14
	0.600	0.000	1.000	0.600	0.750	0.767	0.974	0.811	prov-15
	0.882	0.000	1.000	0.882	0.938	0.935	0.985	0.912	prov-16
	0.750	0.000	1.000	0.750	0.857	0.860	0.970	0.799	prov-17
	0.750	0.015	0.750	0.750	0.750	0.735	0.966	0.797	prov-18
	0.615	0.000	1.000	0.615	0.762	0.775	0.955	0.699	prov-20
	0.429	0.029	0.333	0.429	0.375	0.354	0.917	0.494	prov-21
	0.286	0.000	1.000	0.286	0.444	0.522	0.966	0.779	prov-22
	0.750	0.000	1.000	0.750	0.857	0.860	0.970	0.799	prov-23
	0.643	0.020	0.692	0.643	0.667	0.645	0.972	0.814	prov-3
	0.750	0.015	0.750	0.750	0.750	0.735	0.965	0.755	prov-5
	0.636	0.000	1.000	0.636	0.778	0.790	0.957	0.704	prov-6
	0.700	0.000	1.000	0.700	0.824	0.831	0.964	0.751	prov-7
Weighted Avg.	0.671	0.021	0.862	0.671	0.718	0.726	0.951	0.757	

=== Confusion Matrix ===

```

a b c d e f g h i j k l m n o p q <-- classified as
9 0 3 0 0 0 0 0 0 0 0 0 0 4 0 0 0 | a = prov-1
0 10 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | b = prov-11
0 0 13 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | c = prov-12
0 0 1 8 3 0 0 0 0 0 0 0 0 0 0 0 0 | d = prov-13
0 0 6 0 8 0 0 0 1 0 0 0 0 0 0 0 0 | e = prov-14
0 0 2 0 0 6 0 0 2 0 0 0 0 0 0 0 0 | f = prov-15
0 0 2 0 0 0 15 0 0 0 0 0 0 0 0 0 0 | g = prov-16
0 0 3 0 0 0 0 9 0 0 0 0 0 0 0 0 0 | h = prov-17
0 0 3 0 0 0 0 0 9 0 0 0 0 0 0 0 0 | i = prov-18
0 0 5 0 0 0 0 0 0 8 0 0 0 0 0 0 0 | j = prov-20
0 0 4 0 0 0 0 0 0 0 3 0 0 0 0 0 0 | k = prov-21
0 0 4 0 0 0 0 0 0 0 6 4 0 0 0 0 0 | l = prov-22
0 0 3 0 0 0 0 0 0 0 0 0 9 0 0 0 0 | m = prov-23
0 0 2 0 0 0 0 0 0 0 0 0 0 9 3 0 0 | n = prov-3
0 0 3 0 0 0 0 0 0 0 0 0 0 0 9 0 0 | o = prov-5
0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 7 0 | p = prov-6
0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 7 | q = prov-7

```

Anexo 15 Conjunto de datos: alta_sur bajo el criterio de muestreo: muestra representativa

Time taken to build model: 775.43 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances	175	84.1346 %
Incorrectly Classified Instances	33	15.8654 %
Kappa statistic	0.7878	
Mean absolute error	0.102	
Root mean squared error	0.2472	
Relative absolute error	27.105 %	
Root relative squared error	56.8257 %	
Total Number of Instances	208	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.849	0.110	0.726	0.849	0.783	0.704	0.942	0.859	prov-10
	0.860	0.030	0.881	0.860	0.871	0.837	0.981	0.941	prov-22
	0.889	0.032	0.906	0.889	0.897	0.862	0.988	0.962	prov-4
	0.776	0.040	0.882	0.776	0.826	0.767	0.924	0.881	prov-9
Weighted Avg.	0.841	0.054	0.848	0.841	0.843	0.790	0.957	0.909	

=== Confusion Matrix ===

```

a b c d <-- classified as
45 2 0 6 | a = prov-10
3 37 3 0 | b = prov-22
5 1 48 0 | c = prov-4
9 2 2 45 | d = prov-9

```

Anexo 16 Conjunto de datos: baja_norte bajo el criterio de muestreo: muestra representativa

Time taken to build model: 801.77 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances	159	88.3333 %
Incorrectly Classified Instances	21	11.6667 %
Kappa statistic	0.8251	
Mean absolute error	0.0765	
Root mean squared error	0.2103	
Relative absolute error	17.21 %	
Root relative squared error	44.5648 %	
Total Number of Instances	180	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.877	0.098	0.806	0.877	0.840	0.763	0.962	0.949	prov-19
	0.918	0.076	0.862	0.918	0.889	0.830	0.981	0.943	prov-2
	0.855	0.000	1.000	0.855	0.922	0.891	0.993	0.984	prov-8
Weighted Avg.	0.883	0.057	0.892	0.883	0.885	0.830	0.979	0.959	

=== Confusion Matrix ===

```
a b c <-- classified as
50 7 0 | a = prov-19
5 56 0 | b = prov-2
7 2 53 | c = prov-8
```

Anexo 17 Conjunto de datos: baja_centro bajo el criterio de muestreo: muestra representativa

Time taken to build model: 788.61 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances	235	74.1325 %
Incorrectly Classified Instances	82	25.8675 %
Kappa statistic	0.7237	
Mean absolute error	0.0343	
Root mean squared error	0.144	
Relative absolute error	29.2399 %	
Root relative squared error	59.4384 %	
Total Number of Instances	317	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.789	0.010	0.833	0.789	0.811	0.799	0.949	0.836	prov-1
	0.615	0.007	0.800	0.615	0.696	0.691	0.898	0.707	prov-11
	0.739	0.003	0.944	0.739	0.829	0.825	0.966	0.875	prov-12
	0.842	0.020	0.727	0.842	0.780	0.768	0.939	0.833	prov-13
	0.714	0.017	0.667	0.714	0.690	0.675	0.908	0.787	prov-14
	0.647	0.027	0.579	0.647	0.611	0.589	0.915	0.687	prov-15
	0.611	0.028	0.733	0.611	0.667	0.632	0.892	0.744	prov-16
	0.722	0.013	0.765	0.722	0.743	0.728	0.923	0.679	prov-17
	0.762	0.000	1.000	0.762	0.865	0.866	0.963	0.855	prov-18
	0.722	0.007	0.867	0.722	0.788	0.780	0.952	0.690	prov-20
	0.650	0.010	0.813	0.650	0.722	0.711	0.848	0.622	prov-21
	0.682	0.010	0.833	0.682	0.750	0.738	0.941	0.835	prov-23
	0.870	0.014	0.833	0.870	0.851	0.839	0.968	0.915	prov-3
	0.750	0.007	0.882	0.750	0.811	0.802	0.978	0.857	prov-5
	0.864	0.014	0.826	0.864	0.844	0.833	0.963	0.853	prov-6
	1.000	0.089	0.308	1.000	0.471	0.530	0.993	0.885	prov-7
Weighted Avg.	0.741	0.016	0.792	0.741	0.754	0.744	0.936	0.793	

=== Confusion Matrix ===

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<-- classified as
15	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	a = prov-1
1	8	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	b = prov-11
0	0	17	0	0	1	1	0	0	0	0	1	0	1	0	2	c = prov-12	
0	0	0	16	2	1	0	0	0	0	0	0	0	0	0	0	0	d = prov-13
0	0	0	1	10	2	0	0	0	0	0	0	0	0	1	0	0	e = prov-14
1	0	0	0	0	11	2	0	0	0	0	0	1	0	0	2	f = prov-15	
0	1	1	1	1	0	22	1	0	0	0	1	0	0	1	7	g = prov-16	
0	0	0	1	0	0	2	13	0	0	1	0	1	0	0	0	0	h = prov-17
0	0	0	1	0	0	1	0	16	0	0	0	0	0	0	3	i = prov-18	
0	0	0	1	0	0	0	0	0	13	0	0	1	0	1	2	j = prov-20	
0	0	0	0	0	0	0	3	0	2	13	0	0	0	0	2	k = prov-21	
0	0	0	1	0	2	2	0	0	0	0	15	1	0	1	0	0	l = prov-23
0	0	0	0	1	1	0	0	0	0	0	0	20	0	0	1	0	m = prov-3
1	1	0	0	0	0	0	0	0	0	0	1	0	15	0	2	0	n = prov-5
0	0	0	0	0	0	0	0	0	0	1	0	0	0	19	2	0	o = prov-6
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	p = prov-7

Anexo 18 Conjunto de datos: baja_sur bajo el criterio de muestreo: muestra representativa

Time taken to test model on test split: 0.02 seconds

=== Summary ===

```

Correctly Classified Instances      183          87.9808 %
Incorrectly Classified Instances    25           12.0192 %
Kappa statistic                    0.8387
Mean absolute error                 0.1217
Root mean squared error             0.2204
Relative absolute error             32.3576 %
Root relative squared error         50.6568 %
Total Number of Instances          208
  
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.792	0.039	0.875	0.792	0.832	0.780	0.946	0.871	prov-10
	0.930	0.000	1.000	0.930	0.964	0.956	0.995	0.976	prov-22
	0.889	0.000	1.000	0.889	0.941	0.925	0.991	0.968	prov-4
	0.914	0.127	0.736	0.914	0.815	0.742	0.952	0.827	prov-9
Weighted Avg.	0.880	0.045	0.895	0.880	0.883	0.843	0.970	0.906	

=== Confusion Matrix ===

```

a b c d <-- classified as
42 0 0 11 | a = prov-10
0 40 0 3 | b = prov-22
1 0 48 5 | c = prov-4
5 0 0 53 | d = prov-9
  
```

Cálculo de las métricas de desempeño en clasificadores mediante el desarrollo de fórmulas

Anexo 19 Conjunto de datos: alta_norte bajo el criterio de muestreo con mayor rendimiento: muestra representativa

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2	=== Confusion Matrix ===														Porcentaje de confiabilidad de los proveedores en temporada alta		
3																	
4	a	b	c	<-- classified as													
5	52	4	1	a = prov-19													
6	5	56	0	b = prov-2													
7	1	8	53	c = prov-8													
8																	
9																	
10	52	56	53			TP	SUMA_TP	161									
11	6	12	1			FP											
12																	
13																	
14																	
15																	
16																	
17																	
18																	
19	Precisión global(overall accuracy): 0.894 (recall) 89.44%																
20	=TP/N																
21	Precisión: =TP/TP+FP																
22	Sensibilidad (recall): =TP/ TP+FN																
23	Especificidad =TN/TN+FP																
24	FP Rate =1-TN Rate																
25	Precision Recall																
26	0.897	0.897	0.912												FP Rate		
27	0.824	0.824	0.918												0.049	0.049	
28	0.981	0.981	0.855												0.101	0.101	
29	0.901	0.901	0.894												0.008	0.008	
30																	

Anexo 20 Conjunto de datos: alta _centro bajo el criterio de muestreo con mayor rendimiento: *cross-validation*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1																											
2																											
3																											
4																											
5																											
6																											
7																											
8																											
9																											
10																											
11																											
12																											
13																											
14																											
15																											
16																											
17																											
18																											
19																											
20																											
21																											
22																											
23																											
24																											
25																											
26																											
27																											
28																											
29																											
30																											
31																											
32																											
33																											
34																											

	Precisión:		Sensibilidad (recall):		Especificidad:		FP Rate		
			TP Rate		TN Rate				fp/n
35									
36									
37									
38									
39									
40									
41									
42									
43									
44									
45									
46									
47									
48									
49									
50									
51									
52									
53									
54									
55									
56									
57									

Anexo 21 Conjunto de datos: alta_sur bajo el criterio de muestreo con mayor rendimiento: *cross-validation*

	A	B	C	D	E	F	G	H	I	J	K	L	M	P	Q	R	S	T
2		Confusion Matrix											Porcentaje de confiabilidad de los proveedores en temporada alta					
3		a	b	c	d	<- classified as					TP	FN						
4		103	2	1	6	a = prov-10					103	9						
5	max_diag	3	100	8	1	b = prov-22					100	12						
6	103	8	1	103	0	c = prov-4					103	9						
7		16	6	4	86	d = prov-9					86	26						
8																		
9		103	100	103	86	TP		SUMA_TP	392									
10		27	9	13	7	FP												
11																		
12	CORRECTAS																	
13	INCORRECTAS																	
14	TOTAL																	
15																		
16	Precisión global(overall accuracy):																	
17	=TP/N																	
18																		
19	Precisión:																	
20	=TP/TP+FP																	
21																		
22																		
23																		
24																		
25																		
26	Precision																	
27	0.792																	
28	0.917																	
29	0.888																	
30	0.925																	
31	0.881																	

Anexo 22 Conjunto de datos: baja_norte bajo el criterio de muestreo con mayor rendimiento: *percentage-split*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
2		Confusion Matrix											Porcentaje de confiabilidad de los proveedores en temporada baja				
3		a	b	c	<- classified as						TP	FN					
4	max_diag	29	3	1	a = prov-19						29	4					
5		2	39	0	b = prov-2						39	2					
6		7	0	33	c = prov-8						33	7					
7																	
8																	
9	TP	29	39	33			SUMA_TP	101									
10	FP	9	3	1													
11																	
12	CORRECTAS																
13	INCORRECTAS																
14	TOTAL																
15																	
16	Precisión global(overall accuracy):																
17	=TP/N																
18																	
19	Precisión:																
20	=TP/TP+FP																
21																	
22																	
23																	
24																	
25	Precision																
26	0.763																
27	0.929																
28	0.971																
29	0.895																

Anexo 23 Conjunto de datos: baja_centro bajo el criterio de muestreo con mayor rendimiento: cross-validation.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
3		== Confusion Matrix ==																							
4		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<- classified as							
5		84	6	1	0	2	0	0	3	0	7	4	1	0	1	1	2	a = prov-1				TP		FN	
6		0	91	3	0	0	2	0	0	0	5	0	0	1	1	2	7	b = prov-11				91		21	
7		1	2	88	2	0	1	2	3	2	2	0	1	0	0	0	8	c = prov-12				88		24	
8		0	0	1	77	8	5	0	3	1	13	0	1	1	1	1	0	d = prov-13				77		35	
9		0	0	2	6	73	6	2	3	1	9	0	0	1	1	2	6	e = prov-14				73		39	
10	max_diag	0	1	0	3	3	80	7	1	2	6	0	2	0	0	0	7	f = prov-15				80		32	
11	104	0	1	1	1	0	6	84	1	1	3	0	1	0	0	1	12	g = prov-16				84		28	
12		2	1	2	3	0	3	1	86	1	7	2	0	0	0	1	3	h = prov-17				86		26	
13		1	0	3	0	1	1	4	3	89	2	0	1	0	1	0	6	i = prov-18				89		23	
14		2	0	0	2	0	0	0	4	0	96	0	1	1	0	0	6	j = prov-20				96		16	
15		1	3	1	0	0	0	1	2	0	14	79	0	0	1	0	10	k = prov-21				79		33	
16		1	1	3	3	1	0	1	3	2	5	0	86	1	1	0	4	l = prov-23				86		26	
17		1	1	0	1	0	6	2	2	0	7	0	1	86	0	0	5	m = prov-3				86		26	
18		1	6	2	0	0	0	1	0	1	3	3	0	0	85	2	8	n = prov-5				85		27	
19		0	0	1	1	0	3	0	5	0	7	2	0	2	0	85	6	o = prov-6				85		27	
20		0	0	1	1	1	0	0	0	0	5	0	0	0	0	0	104	p = prov-7				104		8	
21		84	91	88	77	73	80	84	86	89	96	79	86	86	85	85	104	TP SUMA_TP				1373			
22		10	22	21	23	16	33	21	33	11	95	11	9	7	7	10	90	FP							
23																									
24																									
25																									
29	CORRECTAS																								
30	INCORRECTAS																								
31	TOTAL																								
32	Precisión global(overall accuracy):																								
33	=TP/N																								

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
34		Precisión:								Sensibilidad (recall):						Especificidad:							FP Rate	
35		=TP/TP+FP								=TP/ TP+FN						=TN/TN+FP							=1-especificidad ó	
36		Precision						Recall						TN	FP									FP Rate
37		0.894	0.894					0.750	0.750	a	1670	10	0.994										0.006	0.006
38		0.805	0.805					0.813	0.813	b	1658	22	0.987										0.013	0.013
39		0.807	0.807					0.786	0.786	c	1659	21	0.988										0.013	0.013
40		0.770	0.770					0.688	0.688	d	1657	23	0.986										0.014	0.014
41		0.820	0.820					0.652	0.652	e	1664	16	0.990										0.010	0.010
42		0.708	0.708					0.714	0.714	f	1647	33	0.980										0.020	0.020
43		0.800	0.800					0.750	0.750	g	1659	21	0.988										0.013	0.013
44		0.723	0.723					0.768	0.768	h	1647	33	0.980										0.020	0.020
45		0.890	0.890					0.795	0.795	i	1669	11	0.993										0.007	0.007
46		0.503	0.503					0.857	0.857	j	1590	95	0.944										0.057	0.056
47		0.878	0.878					0.705	0.705	k	1669	11	0.993										0.007	0.007
48		0.905	0.905					0.768	0.768	l	1671	9	0.995										0.005	0.005
49		0.925	0.925					0.768	0.768	m	1673	7	0.996										0.004	0.004
50		0.924	0.924					0.759	0.759	n	1673	7	0.996										0.004	0.004
51		0.895	0.895					0.759	0.759	o	1670	10	0.994										0.006	0.006
52		0.536	0.536					0.929	0.929	p	1590	90	0.946										0.054	0.054
53		0.799	0.799					0.766	0.766					0.984									0.016	0.016
54																								

Anexo 24 Conjunto de datos: baja_sur bajo el criterio de muestreo con mayor rendimiento: percentage-split.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
3		== Confusion Matrix ==										Porcentaje de confiabilidad de los proveedores en temporada baja					
4		a	b	c	d	<- classified as				TP	FN						
5		34	0	2	3	a = prov-10				34	5						
6	max_diag	0	26	0	3	b = prov-22				26	3						
7	43	0	0	34	3	c = prov-4				34	3						
8		3	0	1	43	d = prov-9				43	4						
9																	
10																	
11	TP	34	26	34	43	SUMA TP		137									
12	FP	3	0	3	9												
13																	
18	CORRECTAS							137									
19	INCORRECTAS							15									
20	TOTAL							152									
21	Precisión global(overall accuracy):					0.901	90.13%										
22	=TP/N																
24	Precisión:					Sensibilidad (recall):				Especificidad:				1-especificidad			
25	=TP/TP+FP					=TP/TP+FN				=TN/TN+FP							
26	Precision									TN	FP						
27	0.919	0.92				0.872				a	110	3	0.973				
28	1.000	1				0.897				b	123	0	1.000				
29	0.919	0.92				0.919				c	112	3	0.974				
30	0.827	0.83				0.915				d	96	9	0.914				
31	0.906	0.92				0.901							0.965				

Instalación de paquetes en weka de la categoría *classification*

Anexo 25 Vista del proceso

