



UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO.



FACULTAD DE INGENIERÍA CIVIL
MAESTRÍA EN INFRAESTRUCTURA DEL TRANSPORTE EN LA
RAMA DE LAS VÍAS TERRESTRES
DIVISIÓN DE ESTUDIOS DE POSGRADO

TESIS:
DESARROLLO DE UN MODELO DE ESTIMACIÓN DEL ÍNDICE
DE REGULARIDAD INTERNACIONAL (IRI)

PARA OBTENER EL GRADO DE:
MAESTRO EN INGENIERÍA

PRESENTA:
ING. JULIO CESAR TORAL EQUIHUA

ASESOR:
DR. RAFAEL SOTO ESPITIA
CO-ASESORES:
DR. JOSÉ RICARDO SOLORIO MURILLO
DR. LUIS ALBERTO MORALES ROSALES

MORELIA, MICHOACAN. FEBRERO DEL 2023



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



Agradecimientos.

A mi familia, padres y hermanos que siempre me han brindado su apoyo incondicional para poder cumplir todos mis objetivos personales y académicos. A mis padres en especial porque me han brindado el soporte material y económico para poder centrarme en los estudios.

A los doctores Rafael Soto Espitia, Luis Alberto Morales Rosales y José Ricardo Solorio Murillo, por brindarme la oportunidad de desarrollar este proyecto con sus asesorías y consejos a lo largo del desarrollo de esta investigación. Además del tiempo invertido para poder finalizarla.

A la Universidad Michoacana de San Nicolás de Hidalgo, por otorgarme la oportunidad de formarme como profesionista.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico que facilitaron la realización de esta investigación.



Resumen.

La infraestructura vial en México es el conjunto de calles y carreteras que mejoran la conectividad y transporte apoyando el desarrollo socioeconómico. Un elemento importante de la infraestructura vial son los pavimentos, ya que es lo primero que perciben los usuarios al transitar. Para mantener en buenas condiciones los pavimentos es necesario contar con una planeación temprana de los trabajos de conservación.

La presente investigación describe la metodología utilizada para el desarrollo de un modelo de estimación de la evolución del Índice de Regularidad Internacional (IRI) para carreteras con pavimentos flexibles basado en variables explicativas como la profundidad de rodera, macrotextura y porcentaje de agrietamiento que describen el comportamiento del IRI. La base de datos utilizada es de la carretera Hermosillo – Santa Ana, que contiene información histórica de los años 2019 y 2020 de las variables que describen el comportamiento del IRI.

El análisis para determinar el modelo de estimación se llevó a cabo por medio de dos enfoques. El primer enfoque se ejecutó realizando un agrupamiento de los datos formando grupos donde los datos que lo conforman sean similares entre sí y los grupos sean diferentes, esto por medio del método del codo para encontrar el número de grupos para posteriormente formar los grupos por medio del algoritmo k-means. El segundo enfoque se llevó a cabo utilizando la base de datos completa.

El aprendizaje del modelo se llevó a cabo con cinco algoritmos de inteligencia artificial, los cuales son, regresión lineal, redes neuronales artificiales, support vector machine, random forest, m5 rules. Los 5 algoritmos fueron utilizados para los grupos del enfoque con datos particionados y para el enfoque sin datos particionados.

De los dos enfoques analizados se observó que utilizando la base de datos completa se obtienen mejores resultados. El algoritmo con el mejor desempeño de los cinco analizados fue el de random forest, dando como coeficiente de correlación un valor de 0.7262 y un error cuadrático medio de 0.0733.

Palabras clave: Infraestructura vial, pavimentos, variables explicativas, inteligencia artificial, coeficiente de correlación.



Abstract.

The road infrastructure in Mexico is the aggrupation of streets and highways that improve connectivity and transportation, supporting socioeconomic development. A vital element of the road infrastructure is the pavements, the first thing that users perceive when transiting the road. To keep the pavements in good condition is necessary to have an early planning of the conservation works.

This research describes the methodology used to develop a model for estimating the evolution of the International Roughness Index (IRI) for roads with flexible pavement based on explanatory variables, such as depth of ruts, macrotexture, and cracking percentage, which describe the IRI behavior. The database used is from the Hermosillo – Santa Ana highway, which contains historical information from 2019 and 2020 on the variables which describe the behavior of the IRI.

The analysis to determine the estimation model was carried out using two approaches. The first approach was accomplished by grouping the data in two different groups, one where the data is similar to each other and another where the data is different, using the elbow method to find the number of groups to form the groups by the k-means algorithm later. The second approach was accomplished using the complete database.

The model learning was obtained by experimenting with five artificial intelligence algorithms, linear regression, neural networks, support vector machine, random forest, and m5 rules. The five algorithms were applied to the groups of the partitioned and no-partitioned data approaches.

From the two analyzed approaches, we observed better results using the complete database. The algorithm with the best performance of the five analyzed was the random forest algorithm, giving a value of 0.7262 as a correlation coefficient and a mean square error of 0.0733.

Keywords: road infrastructure, pavements, explanate variables, artificial intelligence, correlation coefficient.



Contenido

1. Introducción.....	10
1.1. Planteamiento del Problema.....	12
1.2. Justificación.....	13
1.3. Objetivos	15
2. Estado del Arte.....	16
2.1. Modelo HDM	18
2.2. Regresión Lineal	19
2.3. Redes Neuronales Artificiales	22
2.4. Random Forest	26
2.5. Discusión.....	28
3. Marco Teórico.....	31
3.1. Aspectos Generales.....	31
3.2. Sistema de Gestión de pavimentos.....	32
3.2.1. Niveles de gestión de pavimentos.....	33
3.2.2. Gestión a nivel red.....	36
3.2.3. Gestión a nivel proyecto.....	37
3.3. Auscultación.....	38
3.3.1. Índice de Regularidad Internacional (IRI).....	42
3.3.1.1. Definición.....	42
3.3.1.2. Cálculo del Índice de Regularidad Internacional (IRI).....	42
3.3.2. Condiciones superficiales obtenidos de la auscultación.....	44
3.3.2.1. Profundidad de Roderas (PR).....	44
3.3.2.2. Macrotextura (MAC).....	45
3.3.2.3. Deterioros Superficiales (DET).....	46
3.4. Aprendizaje Automático.....	46
3.4.1. Proceso del aprendizaje automático.....	47
3.4.2. Preprocesamiento.....	49
3.4.2.1. Valores faltantes.....	49
3.4.2.2. Valores atípicos.....	50
3.4.2.3. Transformación de atributos. Normalización.....	50



3.4.3. Agrupamiento de los datos.	51
3.4.3.1. K-Medias.	51
3.4.3.2. Método del codo.	52
3.4.4. Modelos de estimación del deterioro de pavimentos	53
3.4.4.1. Regresión lineal.....	53
3.4.4.2. Redes Neuronales Artificiales.....	54
3.4.4.3. Random Tree.....	55
3.4.4.4. Random Forest.....	56
3.4.4.5. Reglas de asociación. M5 Rules.	57
4. Propuesta de Solución.	58
5. Resultados.	63
5.1. Aspectos generales del tramo de estudio.	63
5.6. Selección de las variables.....	65
5.2. Preprocesamiento de los datos.....	67
5.3.1. Eliminación de los datos perdidos.....	67
5.3.2. Detección y eliminación de los datos atípicos.....	69
5.3.2.1. Rango Intercuartil (IQR).....	69
5.3.2.2. Normas de la Secretaría de Comunicaciones y Transportes SCT .	73
5.3.4. Normalización.	76
5.3. Estadística Descriptiva.	78
5.4. Análisis de Correlación.	83
5.5.1. Coeficiente de Correlación de Pearson.	84
5.6. Análisis de los distintos algoritmos para el modelo de estimación.	86
5.6.1. Enfoque con datos particionados.....	86
5.6.1.1. Método del codo.	86
5.6.1.2. Método K-Means.	87
5.6.1.3. Parámetros de los algoritmos.	89
5.6.1.4. Resultados de los modelos para los tres grupos.	92
5.6.1.5. Resultados de los modelos de la clase 1.....	95
5.6.1.6. Resultados de los modelos de la clase 2.....	96
5.6.1.7. Resultados de los modelos de la clase 3.....	97
5.6.1.8. Resultados de los modelos de la clase 4.....	99



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



5.6.1.9. Resultados de los modelos de la clase 5.....	100
5.6.2. Enfoque sin datos particionados.....	102
5.6.2.1. Regresión Lineal.....	102
5.6.2.2. Redes Neuronales Artificiales.....	103
5.6.2.3. Support Vector Machine.....	104
5.6.2.4. Random Forest.....	105
5.6.2.5. M5 Rules.....	106
5.6.2.6. Resultados de los diferentes modelos de estimación.....	107
5.6.3. Selección del modelo de estimación.....	109
6. Conclusiones.....	112
Bibliografía.....	114



Índice de Tablas.

Tabla 1.	Clasificación de los modelos de deterioro.	16
Tabla 2.	Algoritmos, combinación de variables y resultados de los distintos modelos de predicción del Índice de Regularidad Internacional.....	29
Tabla 3.	Continuación, Algoritmos, combinación de variables y resultados de los distintos modelos de predicción del Índice de Regularidad Internacional.	30
Tabla 4.	Ejemplos de algunos datos perdidos de todas las variables.	68
Tabla 5.	Ejemplos de algunos datos perdidos de una variable.....	68
Tabla 6.	Valores mínimos y máximos del IRI.....	74
Tabla 7.	Intervalos de profundidad de roderas para la clasificación de los tramos. Fuente: Norma de la SCT N-CSV-CAR-1-03-009/16	74
Tabla 8.	Valores mínimos y máximos de la PR	75
Tabla 9.	Intervalos de la macrotextura para la clasificación de los tramos. Fuente: Norma de la SCT N-CSV-CAR-1-03-006/16	75
Tabla 10.	Valores mínimos y máximos de la macrotextura.	75
Tabla 11.	Ejemplo de mediciones de las variables antes del proceso de normalización.	77
Tabla 12.	Ejemplo de mediciones de las variables después del proceso de normalización.	78
Tabla 13.	Estadística descriptiva. Medidas de tendencia central y dispersión ..	79
Tabla 14.	Medidas de forma y posición.	80
Tabla 15.	Estadístico descriptivo de los grupos formados por medio del algoritmo K-Means.....	88
Tabla 16.	Parámetros de cada algoritmo.....	89
Tabla 17.	Continuación tabla 17. Parámetros de cada algoritmo.	90
Tabla 18.	Resultados de las métricas de desempeño de los algoritmos Regresión Lineal, Redes Neuronales y Support Vector Machine.	93
Tabla 19.	Resultados de las métricas de desempeño de los algoritmos Random Forest y M5 Rules.	94



Índice de Gráficas.

Gráfica 1. Evolución del Índice de Regularidad Internacional (IRI) del año 2019.....	70
Gráfica 2. Evolución del Índice de Regularidad Internacional (IRI) del año 2020.....	72
Gráfica 3. Evolución del Índice de Regularidad Internacional de los años 2019 y 2020.....	81
Gráfica 4. Evolución de la Profundidad de Rodera de los años 2019 y 2020. ...	82
Gráfica 5. Evolución de la macrotextura de los años 2019 y 2020.....	82
Gráfica 6. Evolución del Agrietamiento de los años 2019 y 2020.	83
Gráfica 7. Comportamiento del Índice de Regularidad Internacional de los años 2019 y 2020.....	84
Gráfica 8. Análisis de Correlación de las variables seleccionadas.....	85
Gráfica 9. Resultado del Método del Codo.....	87
Gráfica 10. Coeficiente de Correlación de la Clase 1.....	95
Gráfica 11. Error cuadrático medio de la Clase 1.....	96
Gráfica 12. Coeficiente de correlación de la Clase 2.	96
Gráfica 13. Error cuadrático medio de la Clase 2.....	97
Gráfica 14. Coeficiente de Correlación de la clase 3.	98
Gráfica 15. Error cuadrático medio de la clase 3.	98
Gráfica 16. Coeficiente de Correlación de la clase 4.	99
Gráfica 17. Error cuadrático medio de la clase 4.	100
Gráfica 18. Coeficiente de Correlación del Clase 5.....	101
Gráfica 19. Error cuadrático medio de la clase 5.	101
Gráfica 20. Coeficiente de correlación de los modelos para el enfoque sin datos particionados.....	107
Gráfica 21. Error Cuadrático Medio de los modelos para el enfoque sin datos particionados.....	108
Gráfica 22. Coeficiente de correlación de ambos enfoques.....	109
Gráfica 23. Error cuadrático medio de ambos enfoques.	110
Gráfica 24. Validación del modelo de estimación con Random Forest.	111



Gráfica 25. Error Cuadrático Medio de la validación del modelo con Random Forest.
111

Índice de Figuras.

Figura 1. Niveles operativos básicos de gestión de pavimentos y actividades componentes principales (Fuente: Haas, R, Hudson, R & Zaniewski, J, 1994). ...	35
Figura 2. Procedimiento para auscultar una red de carreteras (Fuente: Guía de procedimientos y técnicas para la conservación de carreteras en México, 2014).	40
Figura 3. Acciones para evaluar tramos específicos de carreteras (Guía de procedimientos y técnicas para la conservación de carreteras en México, 2014).	41
Figura 4. Modelo cuarto de carro (Badilla 2009).....	43
Figura 5. Esquema de una rodera (Determinación de la profundidad de rodera, N-CSV-CAR-1-03-009/16).....	45
Figura 6. Macrotextura (Determinación de la macrotextura, N-CSV-CAR-1-03-006/16).....	46
Figura 7. Etapas del aprendizaje automático (Fuente: Pineda. C, (2021).)	47
Figura 8. Gráfica de regresión lineal. (Fuente: Elaboración propia.).....	54
Figura 9. Red neuronal simple (Fausett ed., al 1994).....	55
Figura 10. Estructura general de un árbol de decisión (Jehad, Ali 2012).	56
Figura 11. Estructura de un bosque aleatorio (Medina, Rosa 2017).....	57
Figura 12. Macro localización del tramo de estudio.....	64
Figura 13. Micro localización del tramo de estudio.	65
Figura 14. Escala original del Banco Mundial para el IRI. Fuente: Norma de la SCT N-CSV-CAR-1-03-004/16	73
Figura 15. Modelo de Redes Neuronales Artificiales.....	104



1. Introducción.

La carretera se define como una vía de transporte, que es proyectada y construida fundamentalmente para la circulación de vehículos, actualmente en México es el principal factor que contribuye al desarrollo socioeconómico.

El sistema carretero constituye el modo de transporte más importante en nuestro país favoreciendo a la actividad nacional en diferentes aspectos, por ejemplo, en el sector comercial asistiendo a la entrega oportuna de bienes y servicios, en el sector industrial contando con nuevas vías de transportes para las mercancías o productos.

El tener en buenas condiciones las carreteras permite a los usuarios transitar de manera segura, rápida y cómoda, aumentando el confort que perciben los usuarios al circular por ellas acortando los tiempos de traslados y disminuyendo los costos de operación para los vehículos.

El conocer el comportamiento de los pavimentos, así como el estado en que se encuentra es importante, ya que contribuye al comportamiento social y económico del país. Su importancia es tal ya que, desplaza el 55.6% de la carga de productos o mercancías y al 95.7% de los pasajeros. Es por ello que existe la necesidad de la construcción y conservación de este activo carretero.

Los deterioros son fallas que se presentan en la superficie de rodamiento de una carretera que disminuyen el confort de los usuarios cuando transitan por ellas. Los deterioros de las carreteras son producidos por la combinación de factores climáticos y de tráfico. Generalmente los deterioros se ven reflejadas en pérdidas de tiempo en los traslados de pasajeros o de mercancía y en el peor de los casos ocasionan accidentes con pérdidas humanas y materiales.

El conocer el comportamiento a futuro de las carreteras en México permite contar o mejorar la planeación en los trabajos de conservación de las carreteras, además de conocer el estado actual de las carreteras.



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



Es por ello que las agencias dedicadas a la ingeniería de pavimentos de todo el mundo han realizado estudios e investigaciones para desarrollar modelos de predicción del comportamiento de los parámetros de desempeño de las carreteras con la finalidad de ayudar a realizar una correcta gestión de los pavimentos. Estos modelos se utilizan a nivel red para predecir el comportamiento a futuro con base en la medición de deterioros superficiales de una sección de pavimento e identificar las necesidades de conservación, también se utilizan para estimar las condiciones de la red después de la aplicación de diversos trabajos de mantenimiento. Existen parámetros de desempeño que representan el estado actual de un pavimento, como el Índice de Servicio Actual (ISA), Índice de Regularidad Internacional (IRI). El Índice de Regularidad Internacional (IRI), es un indicador desarrollado por el Banco Mundial en 1986, es ampliamente utilizado en los modelos de rendimiento del pavimento.

Los modelos de rendimiento de la condición del pavimento es el estudio del deterioro del pavimento a lo largo de su ciclo de vida. Evalúa el proceso de los deterioros del pavimento e identifica los principales factores y brindan pronósticos a lo largo del tiempo. Uno de los primeros modelos de rendimiento del pavimento es el Highway Design and Maintenance Standards Model (HDM), propuesto por el Banco Mundial para ayudar a los países a planear y mejorar las condiciones de la infraestructura carretera. El modelo predice el comportamiento de los pavimentos en el futuro y a la vez realiza estimaciones de costos, comparativas y análisis económicos de diferentes opciones de inversión. Con el avance de la tecnología y a la vez de las computadoras han surgido nuevos modelos de predicción que hacen uso de algoritmos de inteligencia artificial, por ejemplo, las redes neuronales artificiales, arboles de decisión y bosques aleatorios.

El objetivo de la presente investigación es realizar un modelo de estimación del Índice de Regularidad Internacional (IRI) a partir de variables conocidas como profundidad de rodera, macrotexturas y agrietamiento, haciendo uso de algoritmos de inteligencia artificial.



1.1. Planteamiento del Problema.

En la actualidad existe una problemática del comportamiento a futuro de los pavimentos, ya que el deterioro del pavimento es producido por varios factores como su edad, condiciones climatológicas, condiciones de tráfico y propiedades propias de la estructura del pavimento. Los pavimentos con el paso del tiempo sufren una serie de deterioros que al presentarse en la capa de rodamiento van disminuyendo la capacidad de proporcionar un tránsito rápido, seguro y cómodo al usuario. Por lo que es necesario que la infraestructura vial opere en rangos aceptables de los parámetros de desempeño mediante la planificación de la conservación.

El usuario al ir circulando por una carretera, lo que observa es el estado en que se encuentra el pavimento, ya que los deterioros en la superficie de rodamiento y las fallas en el señalamiento vial es lo primero que percibe, percatándose si la carretera por la cual transita le ofrece la seguridad y la comodidad.

La condición superficial es un factor que influye de manera directa en los costos de operación de los vehículos, el Índice de Regularidad Internacional es un parámetro que representa la regularidad superficial, a la vez refleja las condiciones superficiales en las que se encuentra la infraestructura vial y detecta anomalías en algunos de sus tramos.

La influencia que tiene el sistema carretero es de gran importancia para la población del país por su impacto socioeconómico, motivo por el cual es necesario conocer su comportamiento de los parámetros de desempeño a futuro. El conocer el comportamiento de los pavimentos flexibles permitirá mejorar la planeación en los trabajos de conservación y que estos se realicen de manera proactiva, es decir, de manera temprana y disminuyendo los costos en los trabajos. Por lo anterior, es necesario desarrollar un modelo de estimación del Índice de Regularidad Internacional a nivel red.



1.2. Justificación.

Actualmente existe una incertidumbre sobre el comportamiento a futuro de los pavimentos flexibles en México. En México, la expansión no se ha visto acompañada por un aumento de los presupuestos de mantenimiento. En cambio, el crecimiento del tránsito ha sido mayor que el esperado y las cargas de los vehículos pesados han excedido la capacidad de soporte de los pavimentos flexibles. La combinación de estos factores ha producido un aumento en el deterioro de las carreteras.

Los deterioros en los pavimentos flexibles provocan diversas dificultades a los usuarios al transitar en estos, como es la inseguridad, la alta incomodidad y los altos costos de operación de los vehículos. La infraestructura vial es de importancia para los usuarios ya que las utilizan diariamente para trasladarse y realizar sus diversas actividades económicas, comerciales y sociales, aumentando el nivel socioeconómico de México.

La superficie de rodamiento es la parte que se ve y que siente el usuario de un vehículo al transitar, ya que las fallas de tipo funcional de un pavimento flexibles y el mal señalamiento es lo primero que percibe, por lo cual el usuario se percata inmediatamente si la carretera se encuentra en condiciones de ofrecer un buen servicio.

Los parámetros funcionales son los que determinan las condiciones de seguridad y confort de los usuarios. Por lo anterior, es importante conocer el comportamiento de la regularidad superficial, desde el inicio de su operación y así como en el futuro, para establecer la planeación en los trabajos de conservación. El Índice de Regularidad Internacional es un parámetro utilizado para la evaluación de la regularidad de los pavimentos, el cual refleja el nivel de comodidad al transitar.

Dada la importancia que tienen las carreteras para México, es necesario conocer su estado, así como su comportamiento a futuro para implementar o mejorar la planeación en los trabajos de conservación y estos se realicen en el tiempo



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



adecuado y así evitar los sobrecostos de los trabajos de conservación y de la operación de los vehículos.

La finalidad de la presente investigación es desarrollar un modelo de estimación del parámetro funcional Índice de Regularidad Internacional (IRI) que permita conocer el comportamiento a futuro de los pavimentos flexibles en base del aprendizaje automático y de algoritmos de inteligencia artificial.



1.3. Objetivos

Objetivo General.

- Desarrollar un modelo para la estimación de la evolución del IRI basado en variables explicativas de la carretera Hermosillo – Santa Ana.

Objetivos Particulares.

- Seleccionar las variables que representen la funcionalidad del estado del pavimento para describir el comportamiento del IRI.
- Realizar e interpretar un análisis estadístico descriptivo para conocer el comportamiento funcional de las variables explicativas del IRI.
- Desarrollar un modelo de estimación para determinar las condiciones futuras del IRI basado en algoritmos de aprendizaje automático y las variables seleccionadas anteriormente.



2. Estado del Arte.

Para conocer los niveles de servicio y evaluar el estado de las carreteras se tienen los indicadores de rendimiento, los cuales son, superficiales, estructurales y de seguridad vial. El IRI es un parámetro superficial que se emplea en los modelos de deterioro del pavimento para predecir el rendimiento del pavimento. Estos modelos son fundamentales para cualquier agencia de carreteras ya que se utilizan a nivel red para predecir el rendimiento futuro de una carretera e identificar las necesidades de mantenimiento y rehabilitación. Así como para estimar las condiciones de la red después de los trabajos de mantenimiento o rehabilitación y determinar la eficacia de estos trabajos.

Los modelos de deterioro generalmente se clasifican de acuerdo a la tabla 1.

1.- Enfoque	2.- Tipo de pronóstico	3.- Nivel de gestión
Mecanicista	Determinista	Proyecto
Empírico - mecanicista	Probabilístico	Red
Empírico		

Tabla 1. Clasificación de los modelos de deterioro.

De acuerdo con la tabla 1, se observa que existen diversas formas de clasificar a los modelos de deterioro. El primer grupo “Enfoque” está representado por los modelos:

- Mecanicistas, basados en algún parámetro de respuesta primaria (comportamiento), deformación o desviación (Haas, R, Hudson, R & Zaniewski, J, 1994).
- Empírico-Mecanicista, basada en la combinación de enfoques mecánicos y empíricos, esto permite establecer una forma apropiada de relación utilizando el conocimiento de los procesos mecanicistas involucrados y luego



usar técnicas empíricas para ajusta estas relaciones a los datos disponible (Robinson, R, Danielson, U & Snaith, M, 1998).

- Empírico, tienen su origen en bases de datos reales conformadas a partir de pavimentos existentes. En dichas bases de datos se registra gran cantidad de información referida a diversos aspectos, tales como información general, datos de diseño, tránsito, condiciones climáticas y de deterioros a lo largo de su vida útil. Posteriormente, se determinan las variables más relevantes y se realiza con ellas un análisis estadístico que da origen al modelo.

El segundo grupo “Tipo de Pronóstico”, está clasificado de acuerdo a su metodología de cálculo:

- Modelos Probabilísticos, predicen típicamente la probabilidad de que una condición particular de la carretera prevalezca en un momento fijo en el futuro, cuyo valor depende solo de la condición actual. Los niveles de probabilidad a veces se asignan a los posibles resultados futuros mediante el juicio de ingeniería, y estos a menudo se determinan mediante el análisis de estimaciones realizadas por un panel de ingenieros expertos (Robinson, R, Danielson, U & Snaith, M, 1998).
- Modelos Deterministas, Son aquellos donde la condición se predice como un valor preciso sobre la base de funciones matemáticas de deterioro observado o medido, estos modelos incluyen mecanicistas, de regresión y mecanicista-empírico (Robinson, R, Danielson, U & Snaith, M, 1998).

Por último, se tiene al tercer grupo “Nivel de Gestión”. Este grupo se determina por el tipo de enfoque o alcance del modelo: el cual se clasifica en nivel proyecto y red.

1. Nivel proyecto. Es el proceso de observación de un proyecto o pavimento en particular, con el propósito de determinar el momento en que se debe realizar una rehabilitación. Usa datos específicos de cada proyecto y otorga varias



opciones de acuerdo a los objetivos, los modelos usados a este nivel requieren de información detallada en secciones individuales de un camino.

2. Nivel Red. Incluye fundamentalmente un proceso de observación de un conjunto de pavimentos que conforman una red de caminos, para planificar decisiones para grandes grupos de proyectos o una red de caminos completa a fin de optimizar la asignación de recursos.

En los últimos años, investigadores como Signal, Gharied, Nassiri, Hossein entre otros han realizado estudios para comprender los modelos de deterioro, así como los mecanismos cambiantes del IRI, que luego se han utilizado para analizar las tendencias de deterioro del pavimento. Estos estudios describen las condiciones funcionales y estructurales del pavimento.

La literatura es rica en modelos para predecir el deterioro del pavimento. Los modelos de predicción se describen a continuación, comenzando con uno de los primeros modelos, el HDM, posteriormente se tienen modelos de predicción basados en algoritmos de inteligencia artificial.

2.1. Modelo HDM

El Highway Design and Maintenance Standards Model HDM, es uno de los modelos más difundidos patrocinado por el Banco Mundial, al igual es uno de los modelos tradicionalmente usados, la metodología utilizada fue fundamentalmente empírica-mecanicista (Solminihaç, H, 2005). Jain (2005) menciona que el modelo de Gestión y Desarrollo de Carreteras HDM es su cuarta versión, es un sistema de apoyo a la toma de decisiones para administradores e ingenieros de carreteras. Pero es importante recalcar que los pavimentos se comportan, desarrollan y progresan a diferentes ritmos en diferentes entornos. Es por ello que el modelo HDM-4 se debe de calibrar para reflejar las condiciones locales y asegurar su relevancia para el análisis técnico-económico de las alternativas de mantenimiento y rehabilitación para una red vial. Por lo tanto, Jain calibra las ecuaciones predeterminadas en HDM-4, con el objeto de mejorar la precisión del modelo y reflejar las tasas observadas de deterioro del pavimento en las carreteras donde se aplica el modelo. Jain define



la calibración en tres niveles: nivel 1, aplicación, basada en un estudio de los datos disponibles y la experiencia de ingeniería del desempeño del pavimento, nivel 2, verificación, basada en datos medidos de la condición del pavimento recopilados de una gran cantidad de tramos de camino y nivel 3, adaptación, recopilación de datos experimentales requeridos para monitorear el desempeño a largo plazo de los pavimentos dentro del área de estudio. Una vez calibrado el modelo, es importante validar el modelo antes de ponerlo en uso para comprobar su calibración. La validez del calibrado se realiza seleccionando una red de carreteras con diferentes tipos de suelo, tipo de terreno, condiciones climatológicas y volumen de tráfico. La validez del modelo se realizó por medio de la calibración usando los datos de condición del pavimento recopilados en las secciones del pavimento y se han obtenido factores de calibración para varios modelos de deterioro, tales como modelos de agrietamiento, desprendimiento, formación de baches y rugosidad. Los valores R^2 obtenidos para la progresión de grietas es 0.98, progresión de desmoronamiento de 0.79, progresión de baches 0.88 y progresión de rugosidad de 0.97.

2.2. Regresión Lineal

Dalla (2017), es un autor que se centra en el desarrollo y validación de un modelo empírico para predecir el Índice de Regularidad Internacional (IRI) a lo largo del tiempo. La rugosidad prevista del pavimento se modela en función del IRI y la edad del pavimento. El modelo tiene en cuenta los efectos del clima, subrasante, el tipo de tratamiento, el tipo de pavimento, la carga de tráfico y el sistema funcional. Los datos de carga de tráfico se dividieron en tres categorías (baja, media y pesada). La categoría sistema funcional describe si la sección está ubicada en un área urbana o rural. Se utiliza un método de mínimos cuadrados no lineal para determinar los coeficientes de calibración para cada combinación de tipo de pavimento, nivel de carga de tráfico, subrasante, clima y sistema funcional. En el proceso de calibración se aplicaron restricciones de límite superior e inferior para cada coeficiente de calibración, se aplicó una técnica Bootstrap no paramétrica para estimar los intervalos de confianza y los errores estándar del modelo. Los datos del IRI recopilados en 2015 se utilizaron para validar el modelo de regresión calibrado, se



compararon las predicciones del IRI de 2015 con los datos observados del IRI de 2005. La distribución del error indica que el RMSE de 0.21 asociado con la validación del modelo proporciona una precisión razonable para predecir el IRI a nivel red.

Para Signal (2021), menciona que un pavimento de nueva construcción puede tener cierta rugosidad inicial, que aumenta con el tiempo a medida que el pavimento se deteriora debido al movimiento de los vehículos y factores ambientales. Se reduce la eficiencia de los vehículos y aumenta el consumo de combustible, los costos de mantenimiento y reparación. Los valores de rugosidad son indicadores integrales de evaluación del pavimento que tienen en cuenta no solo la comodidad y la calidad de conducción sino también la presencia de deterioro. Motivo por el cual Signal se centra en el desarrollo de un modelo de rendimiento del pavimento como IRI como parámetro de índice y su relación con los principales factores que causan el deterioro del pavimento, que en este caso se definen cinco variables independientes, IRI inicial, tráfico, precipitación, días acumulados de baja y alta temperatura. Para el desarrollo del modelo de predicción IRI, el conjunto de datos se clasificó aleatoriamente y se dividió en dos conjuntos, denominados datos “dentro de la muestra” y “fuera de la muestra”. El grupo de la muestra compuesta por el 85% del conjunto de datos se utilizó para desarrollar el modelo de regresión. El 15% restante del conjunto de datos, denominado “fuera de la muestra”, se utilizó para evaluar la eficiencia de predicción del modelo de regresión. El valor de R cuadrado ajustado del modelo de regresión es de 76%, lo que implica que la variación del IRI puede explicarse por el IRI inicial, el tráfico, y los factores climáticos.

Gharieb (2021), desarrolla un modelo predictivo de la rugosidad del pavimento tomando como variables al propio IRI, la edad del pavimento y la carga acumulada equivalente sobre un solo eje, la base de datos histórica fue en un lapso de años del 2001 al 2015 tomados de la Red Nacional de Carreteras de Laos. Gharieb aplica el enfoque de la familia de pavimentos, en la que las secciones de pavimento con propiedades superficiales idénticas se agrupan en familias, las cuales fueron pavimento de Hormigón Asfáltico (CA), pavimento de Doble Tratamiento Superficial



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



Bituminoso (DBST) y Concreto de Cemento (CC). Las secciones que demostraron una disminución del IRI con el tiempo fueron excluidas del estudio, esto por debido a que dichas secciones se realizaron trabajos de mantenimiento, por lo cual el IRI disminuyo con el tiempo. Contrario a las expectativas de aumento de los valores de IRI debido al efecto de las cargas de tráfico. Antes de analizar los datos, primero se estudió la importancia de cada variable utilizando el coeficiente de correlación de Pearson, observando que existe correlación entre el IRI y las variables independientes. Después la base de datos se clasifico aleatoriamente y se dividió en dos conjuntos, denominados datos de aprendizaje y datos de prueba, componiendo el 80% y 20% respectivamente de la base de datos, que se utilizó para desarrollar los modelos de regresión. Los datos de prueba se utilizaron para evaluar la eficiencia de predicción de las ecuaciones de regresión, los modelos desarrollados representarían el comportamiento medio de todas las secciones en una familia de pavimento particular, además, los modelos indican la relación directa del IRI con la edad como con el tráfico, lo anterior por los resultados del R2 de los modelos de regresión que fueron para la familia DBST 0.892 y para la familia CA de 0.847. Gharied menciona que de acuerdo con los resultados los modelos desarrollados se pueden incorporar al Sistema de Gestión de Carreteras de Laos.

Además del análisis de regresión tradicional y con el avance de las computadoras, se han empleado otros algoritmos de inteligencia artificial que funcionan para el análisis y desarrollo de nuevos modelos de predicción del desempeño de pavimentos. En la literatura se encuentran investigaciones que desarrollan modelos de predicción a través de Redes Neuronales Artificiales (ANN) y de bosques aleatorios (Random Forest), algoritmos que se describen a continuación.

Otros autores como Abdelaziz (2018), realizaron la comparación de dos algoritmos para desarrollar un modelo de predicción IRI para pavimentos flexibles utilizando análisis de regresión lineal múltiple y redes neuronales artificiales. Los modelos propuestos predicen el IRI en función de la edad del pavimento, el IRI inicial, las grietas transversales, agrietamiento de cocodrilo y la desviación estándar de la profundidad del surco, datos extraídos de la base de datos LTPP para 6 categorías



distintas de pavimento. Los datos recopilados incluyeron varias variables que se examinaron para comprender el comportamiento de cada una y como puede afectar el valor de IRI a lo largo del tiempo. Primero se observó la relación entre el IRI y la edad del pavimento, la cual reveló que algunas secciones de LTPP estaban sujetas a diferentes actividades de mantenimiento en diferentes momentos de la edad del pavimento. Se observó que algunas actividades de mantenimiento tienen influencia inmediata en el IRI, mientras que otras actividades no lo reflejan de inmediato. Por lo tanto, es importante identificar las acciones de mantenimiento efectivas para identificar los puntos de corte de datos que se incluirán en el modelo, es decir para aquellas secciones con tratamientos efectivos, solo las mediciones de IRI hasta la fecha de mantenimiento efectivo se incluyeron en la base de datos utilizada para el modelado. Algunas secciones cuentan con el problema de datos faltantes, por lo tanto, Abdelaziz siguió un procedimiento mediante el desarrollo de un modelo de regresión para cada tramo. Para desarrollar el modelo de predicción de IRI, primero se estudió el significado de cada variable recopilada en la base de datos mediante el coeficiente de correlación de Pearson, el análisis muestra que algunas de las variables tienen correlación baja con el IRI, las variables con mayor correlación con el IRI son, las fisuras transversales y la desviación estándar de la rodera. Posteriormente se realiza el modelo de regresión que se evalúa con un análisis de varianza ANOVA sobre los resultados de la regresión. Este análisis se realizó con la hipótesis nula de que el IRI no está relacionado con el IRI inicial, edad, fisuras transversales y la desviación estándar de la rodera. Mientras que la hipótesis alterna es que el IRI está relacionado con las variables antes mencionadas, comprobando que existe una relación entre el IRI y las variables independientes que contribuyen al modelo y los coeficientes de regresión son significativos.

2.3. Redes Neuronales Artificiales

Hossain (2017) desarrolla un modelo de predicción del Índice de Rugosidad Internacional (IRI) para pavimentos flexibles utilizando datos climáticos y de tráfico mediante el empleo de modelos de redes neuronales artificiales (ANN). Hossain elige el algoritmo de redes neuronales artificiales ya que tienen la capacidad de



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



ejecutar tareas complejas que está respaldada por tres componentes cruciales. El primero es el patrón de conexión entre las neuronas, también conocido como neuronas. El segundo es el método para determinar los pesos, que se conoce como algoritmo de aprendizaje. El último por las funciones de transferencia, también denominadas funciones de activación neuronal, estos tres componentes deben determinarse primero para construir una red neuronal. La base de datos de una sección de pavimento flexible en Peoria, Illinois fue proporcionada por Long-Term Pavement Performance (LTPP) y cuenta con información histórica de 30 años de mediciones de datos climáticos relacionados con la temperatura promedio anual, el índice de congelación promedio anual, la humedad mínima promedio anual, la humedad máxima promedio anual y la precipitación promedio anual. Para el mismo pavimento, los datos de tráfico con respecto al tráfico diario promedio anual y el tráfico diario promedio anual de camiones se recopilan al igual de la base de datos LTPP. Todos los datos recopilados están dentro de la duración de un ciclo de rehabilitación, cuando se realiza mantenimiento o rehabilitación, el valor de IRI disminuye, el modelo tiene como objetivo predecir el IRI del próximo ciclo de rehabilitación utilizando el IRI recopilado del ciclo de rehabilitación existente. Todos los datos se integran para obtener dos conjuntos de datos con todas las variables extraídas y dispuestos en un orden específico con fines de entrenamiento y prueba. El 50% de los datos climáticos y de tráfico se usa para entrenar la red ANN y el 50% restante se usa para probar la red. La red ANN predice los valores de IRI en función de los datos de prueba y estos valores se comparan con los valores de IRI proporcionados en la base de datos LTPP. Así mismo se utiliza el algoritmo de retropropagación Feed – Forward para predecir el IRI ya que se entrenan múltiples entradas. Se hace uso de tres funciones de transferencia para relacionar todas las neuronas y conectar los parámetros para predecir una salida, las tres funciones utilizadas en la estructura de la red neuronal fueron TANSIG, LOGSIG y PURELIN. Para comparar las predicciones de todas las permutaciones y combinaciones, todas las funciones de transferencia se probaron en la red y los resultados se verificaron a través de Root Mean Square Error (RMSE). La arquitectura se determinó mediante



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



el método de prueba y error para seleccionar el número de capas ocultas y el tipo de función de transferencia en función del valor RMSE más bajo siendo la arquitectura 7-9-9-1. La función de transferencia no lineal TANSIG generó un modelo mejor, el RMSE observado fue de 0.012 utilizando la arquitectura 7-9-9-1.

Kargah (2010) busca realizar un modelo para predecir los cambios en las condiciones de la red y así proponer recomendaciones de mantenimiento y rehabilitación. También, las predicciones se utilizan para estimar la mejoría de la red después de los trabajos de mantenimiento y rehabilitación. El modelo que se desarrolló para predecir los cambios en el Índice de Rugosidad Internacional IRI fue el de Redes Neuronales Artificiales ANN. Kargah menciona que la selección de las variables de entrada para el sistema se debe de realizar con certeza ya que la inclusión de gran cantidad de variables hace el proceso de adquisición de datos y predicción sea complicado, es por ello que solo consideró las variables que tienen una influencia considerable como es la rugosidad inicial, la edad del pavimento, el tráfico, las condiciones climáticas (precipitación e índice de congelación), propiedades estructurales del pavimento (contenido de humedad y porcentaje que pasa el tamiz no.200). Los datos de las variables antes mencionadas fueron proporcionados por el programa LTPP. En el desarrollo del modelo de ANN, el 80% de la base de datos de las variables de entrada se seleccionaron al azar para el entrenamiento de la red neuronal, el 20% restante se utilizó para como conjunto de datos de prueba. Para encontrar la arquitectura de la red neuronal (número de capas y neuronas), se realizó mediante la prueba y error, es decir, se realizaron numerosas pruebas de las cuales se seleccionó la que mejor se adecuaba al modelo, la cual fue de 3 capas (entrada, oculta y salida). Con ello se puede utilizar el modelo para determinar cuándo un pavimento requiere tratamientos y a la vez, como dicho pavimento se va a comportar después de cada tipo de mantenimiento y de rehabilitación.

Hossain (2020) se centra en desarrollar modelos de predicción del Índice de Rugosidad Internacional IRI a futuro de los pavimentos rígidos, Hossain hace uso del programa LTPP para obtener los de datos de las variables de entrada, tales



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



como los factores climáticos (temperatura, precipitación, índice de congelación), tráfico (tráfico diario promedio anual y tráfico diario promedio anual de camiones) e IRI. En el modelo se realizan 4 zonas de estudios de acuerdo con su clima (congelación húmeda, sin congelación húmeda, congelación seca y sin congelación seca), de las cuales se seleccionaron 10 tramos en total de pavimento rígido. El método de aprendizaje artificial que se usa es el de Redes Neuronales Artificiales (ANN), en el cual el 70% de los datos de clima y tráfico se utilizan para entrenar la red neuronal, el 15% se usa para probar la red y el 15% restante es utilizada para validar la red. La red neuronal artificial predice los valores del IRI basándose en los datos probados, y estos valores se comparan con los valores de IRI extraídos del programa LTPP. Para encontrar la mejor arquitectura del modelo ANN se realizaron pruebas con distintos números de capas ocultas, la cual arrojó que el modelo 7-9-9-1 ANN predijo los valores más cercanos de IRI comparándolos con los valores de IRI medidos por el programa LTPP, esto con el uso de variables como las condiciones climatológicas y de tránsito.

Ziari (2015) se centran en construir un modelo de predicción de desempeño en los pavimentos, en el cual se enfocan en utilizar algoritmos matemáticos, como el Método Grupal de Manejo de Datos (GMDH) y la Red Neuronal Artificial (ANN). Durante el modelado se clasifican las variables de entrada en tres grupos donde el primer grupo es la estructura del pavimento (espesor de la capa de rodamiento, base y subsabe, contenido de asfalto, etc.), el segundo grupo cambio climático (máximos y mínimos de temperatura y precipitación) y el último grupo es el de tráfico. La base de datos de dichas variables fue proporcionada por el programa LTPP. El desempeño del pavimento conoce a través de las predicciones del IRI producidos por los GDMH y el ANN, dichos valores del IRI se compararon con el IRI real para evaluar el rendimiento de los modelos. Las predicciones de las condiciones del pavimento se realizaron en tres etapas, 1 y 2 años (corto plazo) y a lo largo de su ciclo de vida (largo plazo), donde a cada etapa le pertenece un conjunto de datos de capacitación, pruebas y validación. Ziari concluye que los modelos GDMH con las nueve variables de entrada no es capaz de predecir el comportamiento del



pavimento, en cambio los modelos ANN son mejores que los GDMH para predecir la condición del pavimento, esto cuando se tiene datos homogéneos disponibles.

2.4. Random Forest

Los autores anteriores se centran en utilizar modelos de redes neuronales artificiales que se entrenan para tener una alta precisión, sin tener en cuenta otros algoritmos de aprendizaje automático adecuados y descuidando la importancia de la capacidad de generalización de los modelos para aplicaciones de Ingeniería de Pavimentos (Marcelino, 2019). Marcelino (2019), propone un enfoque general de aprendizaje automático para el desarrollo de un modelo de predicción en sistemas de gestión de pavimentos, el enfoque propuesto es la predicción del Índice de Regularidad Internacional (IRI) para 5 y 10 años, el modelo está basado en un algoritmo de bosque aleatorio, utilizando conjuntos de datos extraídos de LTPP que comprende mediciones anteriores del IRI, datos estructurales, climáticos y de tráfico. Después del recopilado de datos, es necesario procesar los datos, para obtener un conjunto de datos limpio y adecuado para el proceso de aprendizaje, las técnicas que utilizó Marcelino fue el de detectar y eliminar errores o inconsistencias de los datos, además de analizar los datos faltantes de las distintas variables. Una vez procesado la base de datos se empieza con el modelado. Los bosques aleatorios proporcionan una forma de reducir la varianza de un modelo de aprendizaje automático mediante la combinación de diferentes modelos. Los bosques aleatorios usan Bootstrap para generar múltiples conjuntos de datos de entrenamiento, esta técnica emula el proceso de obtención de nuevos conjuntos de datos mediante el muestreo repetido del conjunto de datos original. Una vez que varios conjuntos de entrenamiento están disponibles, el algoritmo entrena a un alumno base en cada uno de esos conjuntos. Los modelos resultantes se promedian y los distintos resultados se fusionan en una sola predicción. Para obtener el modelo del bosque aleatorio se deben definir los hiperparámetros correspondientes al algoritmo, como es el número de árboles en el bosque, profundidad máxima del árbol, número mínimo de muestras requerida para dividir un nodo interno, número mínimo de muestras requeridas para estar en un nodo hoja, número de funciones a



tener en cuenta al buscar la mejor división. Se realizan diferentes modelos con diferentes valores de los hiperparámetros, y se elige el modelo cuyo valor de MSE sea el más bajo. Para el caso del modelo a 5 años se obtuvo un valor de 0.064, para el caso del modelo de predicción de 10 años el valor más bajo de MSE fue de 0.104. Los modelos de predicción de 5 años son más precisos y generalizan mejor que los modelos de predicción de 10 años, esto se debe al tamaño del conjunto de datos. Los resultados indican que el modelo de predicción de 5 años tiene una capacidad predictiva aceptable ya que el coeficiente de determinación R^2 es de 0.98 mientras que el modelo de predicción de 10 años el resultado de R^2 es de 0.93.

Gong (2018), realiza un modelo de regresión de bosques aleatorios para estimar el índice de rugosidad internacional (IRI) de pavimentos flexibles. Los árboles de clasificación y regresión han demostrado ser exitosos debido a su excelencia en la interpretación, visualización. Sin embargo, también adolece de problemas como el sobreajuste, la falta de solidez frente a los valores atípicos y la escasa capacidad predictiva en comparación con otros métodos. Para abordar estos problemas, un método común es agregar los resultados de muchos árboles. Los bosques aleatorios son uno de los métodos con buen desempeño predictivo, funcionan construyendo un grupo de árboles de decisión creados aleatoriamente y pronosticando la clase que es la moda de las clases (clasificación) o la media (regresión) de los árboles individuales. Las variables a utilizar fueron IRI inicial, mediciones de deterioro, tráfico, datos climáticos, propiedades de la estructura del pavimento, al realizar los análisis de correlación se observa que el IRI está altamente correlacionado con la rugosidad inicial, mientras que las fallas como el agrietamiento transversal, el agrietamiento por fatiga, el agrietamiento por bloques, el agrietamiento longitudinal y la formación de surcos tiene un fuerte impacto en el IRI. Una vez establecidas las variables se desarrolla el modelo. El procedimiento para construir un Random Forest es el siguiente, se supone que hay B árboles en el bosque, para cada árbol se extrae una muestra de arranque de tamaño N del conjunto de entrenamiento (submuestreo de fila), para un conjunto de entrenamiento compuesto por p predictores, m de p predictores se eligen al azar



como candidatos para la división (submuestreo de columna), se elige la mejor variable y punto de división entre los predictores y se divide cada nodo en dos subnodos, se predicen nuevos datos agregando las predicciones de los árboles B, es decir, promediando las predicciones de cada árbol individual del bosque. Para encontrar la mejor combinación de hiperparámetros para el modelo, es decir, el número de árboles (B), el número de predictores (m) y muestras (N), se utilizó una búsqueda en cuadrícula en combinación con una validación cruzada, el número de muestras en cada árbol (N) se determinó primero mediante una búsqueda en cuadrícula separada, que consiste en usar el 50%, 75% y 100% de las muestras, se encuentra que la inclusión de todas las muestras genera el MSE más pequeño, después de elegir la N, se utilizó otra búsqueda en cuadrícula con una validación cruzada de 4 veces para encontrar la mejor combinación de B y m. Se encuentra que el bosque crecido con 500 árboles y 11 predictores produce el MSE más pequeño.

2.5. Discusión.

En los últimos años, como se vio en los párrafos anteriores, diferentes autores de distintas partes del mundo se han dedicado a realizar estudios o investigaciones relacionados a desarrollar un modelo predictivo del Índice de Regularidad Internacional (IRI) en los pavimentos. Los autores desarrollan el modelo realizando diferentes combinaciones de variables independientes o predictoras que representen el comportamiento del IRI en los pavimentos. Los modelos basados en diferentes algoritmos de inteligencia artificial arrojan resultados y para conocer el nivel de precisión de los modelos los autores se basan en las métricas de desempeño. En la tabla 2 y 3 se muestra a los distintos autores, así como la combinación de variables que utilizaron, el algoritmo en que se basaron para realizar el modelo de predicción y por ultimo los resultados basados en las métricas de desempeño que obtuvo cada autor con sus respectivas variables y algoritmo.



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



Autor	Algoritmo	Variabes	Coficiente de Correlación	RMSE
Taranath Sigdel & Rojee Pradhananga (2021)	Regresión lineal	IRI inicial	0.76	
		Transito		
		Precipitación		
		Temperaturas Max y Min		
Gharied	Regresión lineal	IRI	0.847	
		Edad del pavimento		
		Condiciones Climatológicas		
		Trafico		
Somayeh Nassiri, Mohammad Hossein Shafiee & Alireza Bayat (2013)	Regresión Lineal Múltiple	Edad del pavimento	0.39	
		Tránsito		
		Propiedades de la estructura		
		Deterioro		
Nader Abdelaziz, Ragaa T. Abd El-Hakim & Sherif M. El-Badawy & Hafez A. Afify (2017)	Regresión lineal	Edad del pavimento	0.57	
		Grietas Transversales		
El-Badawy & Hafez A. Afify (2017)	Redes Neuronales Artificiales	Grietas de cocodrilo	0.75	
Hossain, MI, Gopisetti, LSP y Miah, MS (2017)	Redes Neuronales Artificiales	Datos Climáticos		0.012
		Tráfico		

Tabla 2. Algoritmos, combinación de variables y resultados de los distintos modelos de predicción del Índice de Regularidad Internacional.



Autor	Algoritmo	Variables	Coefficiente de Correlación	RMSE
Nima Kargah-Ostadi, Shelley M. Stoffels & Nader Tabatabaee (2010)	Redes Neuronales Artificiales	IRI inicial	0.9578	
		Edad del pavimento		
		Tránsito		
		Condiciones Climatológicas		
		Propiedades del pavimento		
Pedro Marcelino, Maria de Lurdes Antunes, Eduardo Fortunato & Marta Castilho Gomes (2019)	Random Forest	IRI	5 años=0.98 10 años=0.93	5 años=0.064 10 años=0.104
		Datos estructurales		
		Condiciones Climatológicas		
		Tráfico		
Hongren Gong, Yiren Sol, Xiang Shu & Baoshan Huang (2018)	Random Forest	IRI	0.998	0.0005
		Deterioro		
		Tránsito		
		Condiciones Climatológicas		
		Propiedades de la estructura		

Tabla 3. Continuación, Algoritmos, combinación de variables y resultados de los distintos modelos de predicción del Índice de Regularidad Internacional.



3. Marco Teórico.

En este capítulo se desarrollarán los conceptos necesarios para la realización de un modelo de estimación del Índice de Regularidad Internacional a nivel red de la República Mexicana. En primera instancia se hablará de lo más general como la definición y función del pavimento, así como su clasificación, continuando con el desarrollo del tema, los sistemas de gestión, su definición, importancia en las distintas dependencias de gobierno, seguido de las diferencias entre los dos principales niveles, red y proyecto. En los diferentes sistemas de gestión de pavimentos es necesario contar con una base de datos de diferentes parámetros superficiales y estructurales, el medio para obtener dichos datos es a través de la auscultación de carreteras, por ello se explicará su procedimiento. Por último, se expondrán los diferentes algoritmos de inteligencia artificial a utilizar.

3.1. Aspectos Generales.

De acuerdo con Tapia (2004), el pavimento es una estructura que está caracterizado por las propiedades, espesores y acomodo de los distintos materiales que conforman un conjunto de capas colocadas y apoyadas sobre otra, denominada subrasante, con el propósito de recibir en forma directa las cargas del tránsito y transmitir las a los estratos inferiores en forma disipada y distribuyéndolas con uniformidad. Este conjunto de capas proporciona la superficie de rodamiento y permite por un periodo determinado la circulación de vehículos en condiciones de comodidad y seguridad bajo las diversas condiciones que se presenten.

Así mismo Tapia (2004) clasifica a los pavimentos en:

- Pavimentos Flexibles. Estos cuentan con una capa de rodamiento constituida por mezcla asfáltica, resultan más económicos en su construcción inicial, pero tienen la desventaja de requerir mantenimiento constante para cumplir con su vida útil.
- Pavimentos Rígidos. Una de las características principales es que la superficie de rodamiento es proporcionada por las losas de concreto



hidráulico que en algunas ocasiones presentan un armado de acero. Por su mayor rigidez distribuyen las cargas de los vehículos hacia las capas inferiores por medio de toda la superficie de la losa y de las losas adyacentes.

- Pavimentos Mixtos o Compuestos. Están conformados por una capa de concreto hidráulico, cubierta por una carpeta asfáltica, se emplean en calles y su justificación es debido a que su posición impide efectuar excavaciones a mayor profundidad para alojar una estructura del pavimento flexible convencional. Así mismo, pueden tener una mayor capacidad estructural y por consiguiente un mejor desempeño.

3.2. Sistema de Gestión de pavimentos.

La mayoría de los países consideran que un sistema de transporte por carretera eficiente es una condición previa esencial para el desarrollo económico general, por lo que dedican considerables recursos a la construcción y mejora de las carreteras (Robinson, R, Danielson, U & Snaith, M, 1998).

Un sistema en general consiste de un conjunto de componentes que interactúan. En la estructura del pavimento, los componentes que interactúan entre si suelen ser la superficie, incluyendo los carriles, la base, subbase y la subrasante. Los factores exógenos que afectan al pavimento son el medio ambiente, el tráfico y el mantenimiento. A la vez el sistema de gestión de pavimentos consta de componentes que interactúan entre sí, como la planificación, la programación, el diseño, la construcción, y el mantenimiento. Un sistema de gestión de pavimentos ideal permitiría obtener el mejor valor posible de los fondos disponibles y, al mismo tiempo, proporcionar pavimentos sin problemas, seguros y económicos. No existe un sistema de gestión de pavimentos ideal que sea el mejor para todas las agencias. Cada agencia representa una situación única con necesidades específicas. Por lo tanto, cada organismo debe definir cuidadosamente lo que quiere de un sistema de gestión de pavimentos (Haas, R, Hudson, R & Zaniewski, J, 1994)

La utilización de un adecuado sistema de gestión sobre los caminos permitirá obtener un adecuado rendimiento de los recursos invertidos, valorando para tal



efecto a los diversos costos involucrados. Para conseguir un adecuado sistema de gestión es útil conocer algunos de los requerimientos esenciales (Haas, R, Hudson, R & Zaniewski, J, 1994):

- Capacidad de ser fácilmente utilizado, posibilitando agregar y actualizar datos y modificarlo con nueva información sin mayor complicación.
- Capacidad de considerar estrategias alternativas dentro de la evaluación.
- Capacidad de identificar la estrategia o alternativa óptima.
- Capacidad de basar sus decisiones en procedimientos racionales, con atributos, criterios y restricciones cuantificables.
- Capacidad de usar información de retroalimentación para conocer las consecuencias de las decisiones.

Los pavimentos son estructuras complejas que involucran variables tales como combinaciones de cargas que soportan, solicitudes de medio ambiente, materiales y formas de construcción, entre otras diversas. Son importantes los factores técnicos y económicos que involucran su construcción, explotación y mantención para poder hacer una apropiada gestión de los pavimentos.

La auscultación de carreteras es una herramienta fundamental para la gestión de la infraestructura carretera influye en su etapa de construcción y en la etapa de operación, ya que permite obtener información relevante sobre el estado y condición de sus parámetros superficiales (IRI, PR, MAC, CF), estructurales (DEF), así como parámetros de seguridad vial. Los datos arrojados por la auscultación generan bases de datos permitiendo clasificar a las carreteras según su estado físico.

3.2.1. Niveles de gestión de pavimentos.

Los problemas de la gestión de pavimentos operan en dos niveles principales, los que se conocen como nivel de proyecto y nivel de red, que se consideran como las instancias más importantes en la toma de decisiones, decisiones globales que afectan a la red carretera como un todo y las decisiones más específicas afectan a los proyectos individuales (Solminihac, H, 2005).



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



Para Hernan Solminihac (2005), un sistema de gestión de pavimentos opera a todos los niveles de gestión, pero cada nivel tiene sus necesidades particulares en cuanto a tipo y cantidad de información. Un sistema completo y eficiente debe de producir la información necesaria para apoyar a la toma de decisiones, ya sea a nivel red y de proyecto.

La figura 1 enumera las principales actividades que ocurren en los dos niveles de trabajo u operación en la gestión de pavimentos. Se observa que la gestión a nivel red tiene como objetivo principal el desarrollo de un programa prioritario y un cronograma de trabajo, dentro de las restricciones presupuestarias. El trabajo a nivel de proyecto, se pone en marcha en el momento adecuado del cronograma y representa la implementación física real de las decisiones de la red (Haas, R, Hudson, R & Zaniewski, J, 1994).

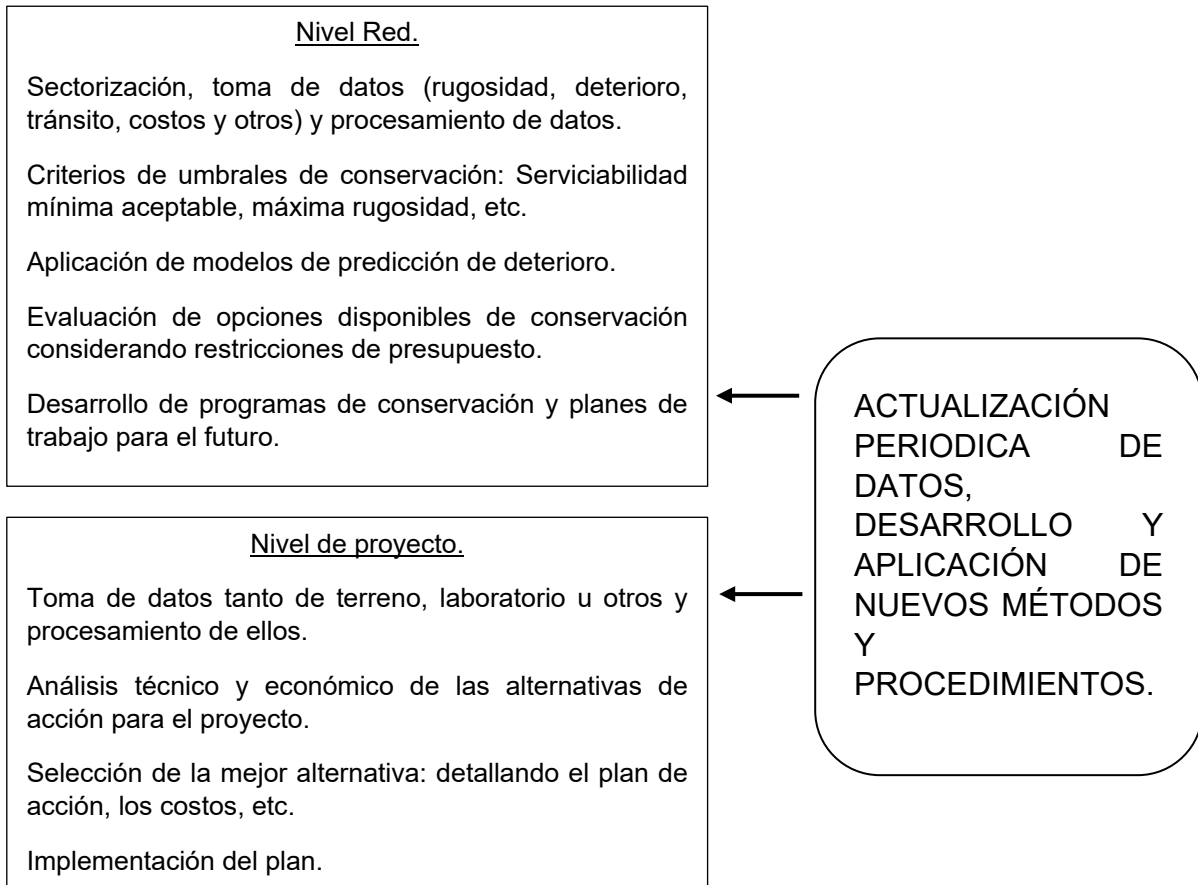


Figura 1. Niveles operativos básicos de gestión de pavimentos y actividades componentes principales (Fuente: Haas, R, Hudson, R & Zaniewski, J, 1994).



A continuación, se describirá las características, los datos necesarios y procedimientos de cada tipo de nivel, así como los beneficios.

3.2.2. Gestión a nivel red.

El nivel de red incluye fundamentalmente un proceso de observación de un conjunto de pavimentos que conforman una red de caminos, para planificar decisiones para grandes grupos de proyectos o una red de caminos a fin de optimizar la asignación de recursos (Solminihaç, H, 2005).

En la gestión a nivel red se contemplan las siguientes actividades (Haas, R, Hudson, R & Zaniewski, J 1994)

- Identificación de necesidades y “vías candidatas” a ser mejoradas dentro de la red de caminos. Dentro de la identificación futura se contempla la aplicación de modelos de comportamiento y de deterioro de pavimentos.
- Generación de alternativas para cada proyecto candidato o sección a mantener.
- Selección de período de análisis, tasa de descuento, niveles de calidad mínimos del pavimento, etc., para el análisis técnico-económico e identificación de las bases para la decisión.
- Análisis técnico de cada alternativa en función del comportamiento esperado en el pavimento.
- Análisis económico de cada alternativa en función de los costos y beneficios esperados para el ciclo de vida del pavimento.
- Desarrollo de un programa para nuevas construcciones, mantención y rehabilitación de los pavimentos de la red en estudio.

El nivel de red tiene como propósito el desarrollo de un programa prioritario y organizado de rehabilitación, mantenimiento o construcción de nuevos pavimentos teniendo en cuenta la restricción de presupuestos correspondientes (Solminihaç, H, 2005).



Para Robinson (1998) hay una diferencia entre los beneficios que otorga cada nivel, los beneficios a nivel red suelen ser menos tangibles. Sin embargo, es a nivel red, cuando se contempla las funciones de planificación y programación, para conocer cuando una carretera se volverá deficiente, mostrando diferentes alternativas como nueva construcción o conservación.

3.2.3. Gestión a nivel proyecto.

El sistema de gestión a nivel de proyecto contempla básicamente una decisión detallada para un proyecto individual, como tal este requiere información detallada de secciones específicas del pavimento.

Los datos utilizados para los estudios a nivel proyecto incluyen los siguientes (Haas, R, Hudson, R & Zaniewski, J, 1994):

- Cargas que recibe el pavimento.
- Factores ambientales que lo afectan.
- Propiedades de su base, subbase y subrasante.
- Variables de construcción y mantención.
- Costos.

Una típica salida de los modelos a nivel de proyecto corresponde a un conjunto de estrategias que minimicen los costos totales de ciclo de vida del pavimento, en los que se incluyen los de costos de construcción y/o mantenimiento, además de los usuarios (Solminihaç, H, 2005).

En gestión a nivel de proyecto se contempla las siguientes actividades (Haas, R, Hudson, R & Zaniewski, J, 1994).

- Generación de alternativas de tratamientos de conservación de pavimentos.
- Selección del periodo de análisis, tasa de descuento, niveles de calidad mínimos del pavimento, entre otros, para el análisis técnico económico de los pavimentos.
- Análisis técnico de cada alternativa en función del comportamiento esperado en el pavimento.



- Análisis económico de cada alternativa en función de los costos y beneficios esperados para el ciclo de vida de pavimento.
- Selección de la alternativa adecuada, con base en criterios cuantitativos y cualitativos.

En general, el nivel de proyecto corresponde a las decisiones de conservación, reconstrucción o construcción de uno nuevo.

Para Robinson (1998), dependiendo de la actividad de gestión de proyecto en particular, es posible ser más específico acerca de los beneficios que probablemente se obtendrán al considerar las actividades a nivel proyecto. Por ejemplo, reducir los costos de operación de los vehículos, los caminos necesitarán menos insumos de mantenimiento, ahorro en tiempo de viaje y reducción de la siniestralidad.

3.3. Auscultación.

La auscultación de carreteras es una herramienta para la gestión de la infraestructura carretera en su etapa de construcción y en fase de operación, ya que permite obtener información relevante sobre el estado y condición de sus parámetros superficiales.

La auscultación es aplicable a una red de carreteras a cargo de una dependencia, para que a través de un sistema de gestión se determinen las inversiones requeridas con el fin de conservarla en buen estado de operación, se definan las prioridades de atención y se realice la programación de los trabajos por ejecutar. La auscultación se realiza con equipos de tecnología de alto rendimiento para obtener diversa información sobre las condiciones de servicio y estructurales de los pavimentos. La auscultación de una red de carreteras en operación tiene como objetivo fundamental medir sus condiciones de servicio actuales de los siguientes elementos (Guía de procedimientos y técnicas para la conservación de carreteras en México, 2014):



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



- ✓ Pavimento.
- ✓ Seguridad vial.
- ✓ Obras de drenaje.
- ✓ Derecho de vía.
- ✓ Señalamiento.
- ✓ Cortes y taludes.
- ✓ Estructuras.

Los datos obtenidos durante la auscultación sirven para calificar y evaluar las características de la red en términos de sus condiciones superficiales, estructurales y de seguridad vial. En la figura 2 se muestra el procedimiento de auscultación, donde comienza con la medición de los parámetros superficiales, estructurales y de seguridad vial, dichos parámetros formaran una base de datos, que con la ayuda de un análisis estadístico se conocerá el estado físico de las carreteras, dependiendo de su deterioro se propondrán diferentes líneas de acción, una vez elegida la línea de acción se realizará su estudio y proyecto, para comenzar de nuevo con el ciclo de las mediciones superficiales, estructurales y de seguridad vial.

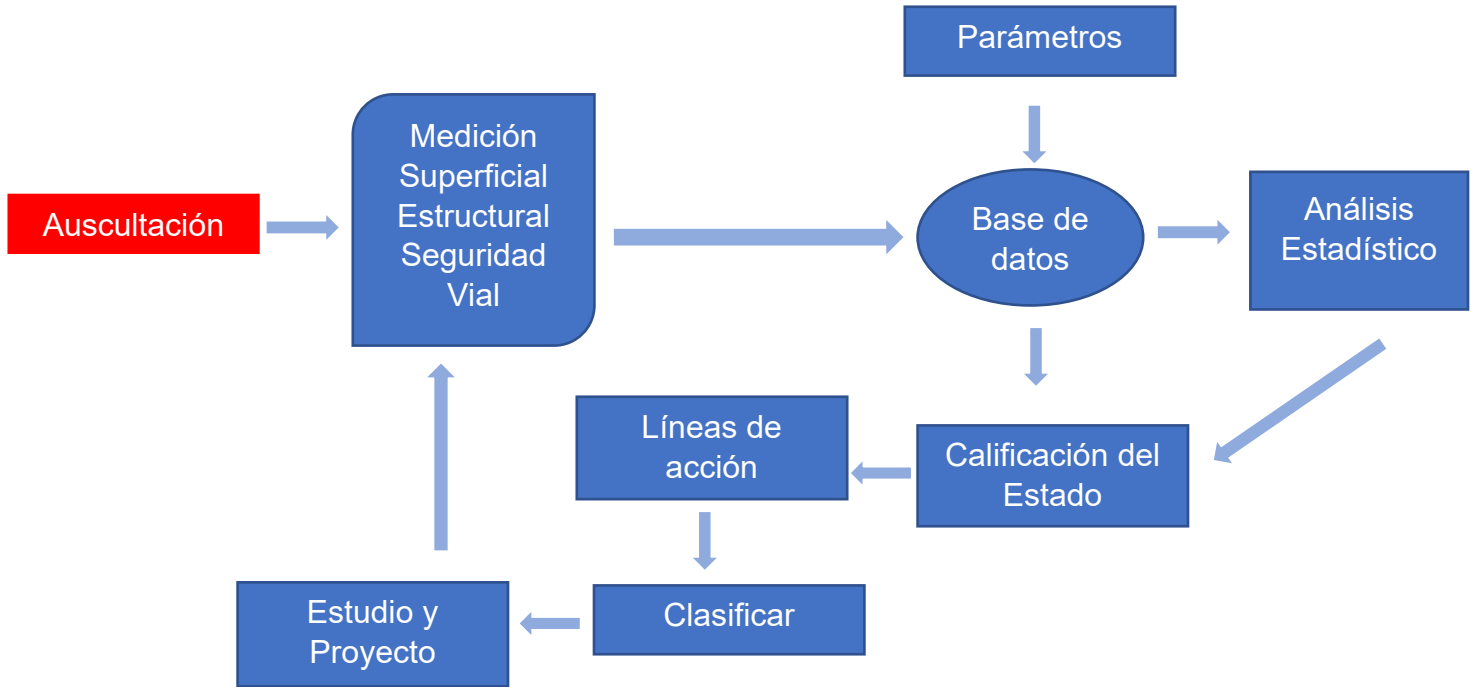


Figura 2. Procedimiento para auscultar una red de carreteras (Fuente: Guía de procedimientos y técnicas para la conservación de carreteras en México, 2014)

Como se mencionó en el párrafo anterior, las condiciones a medir en una auscultación de carreteras son las superficiales, estructurales y de seguridad vial. Los parámetros de cada una se muestran en la figura 3. Por ejemplo, la condición superficial está dada por parámetros como el Índice de Regularidad Internacional, Profundidad de Rodera, Macrotextura, entre otros. Mientras que la condición estructural está representada por las deflexiones y la seguridad vial es evaluada a través de 62 atributos de la infraestructura. Dichos parámetros son medidos a través de equipos de alto rendimiento.

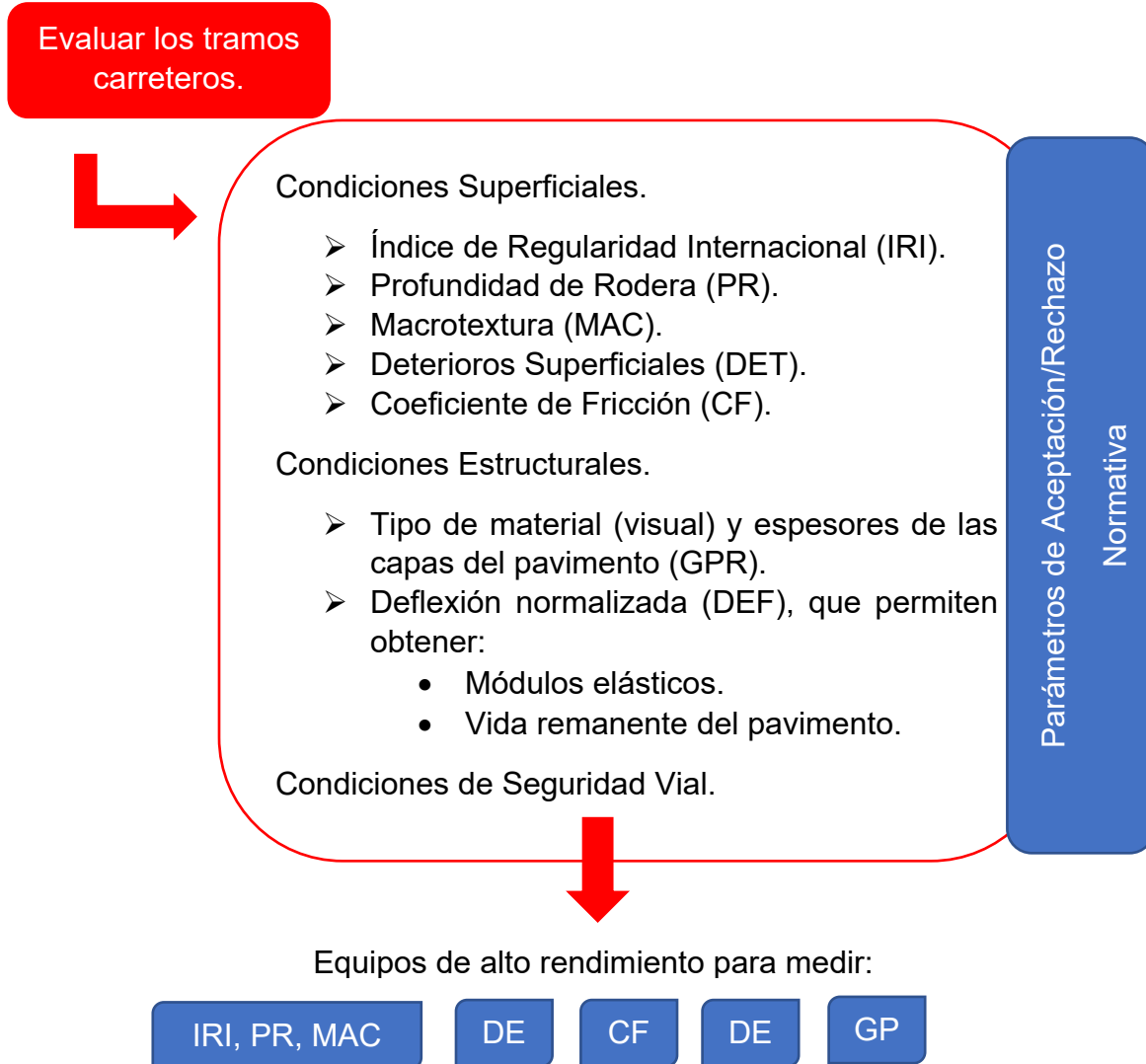


Figura 3. Acciones para evaluar tramos específicos de carreteras (Guía de procedimientos y técnicas para la conservación de carreteras en México, 2014).



3.3.1. Índice de Regularidad Internacional (IRI).

3.3.1.1. Definición.

La regularidad es el conjunto de desviaciones de la superficie de un pavimento con respecto a una superficie perfectamente plana, con dimensiones características que afectan la dinámica vehicular, la calidad del viaje, las cargas dinámicas y el drenaje del pavimento (Determinación del Índice de Regularidad Internacional (IRI), SCT).

El Índice de Regularidad Internacional es un indicador estadístico de la irregularidad superficial del pavimento, al igual que otros indicadores representa la diferencia entre el perfil longitudinal teórico (recta o parábola continua perfecta, $IRI=0$) y el perfil longitudinal real existente en el instante de la medida. El perfil real de una carretera recién construida tiene un estado cero, definido por su IRI inicial > 0 , debido a condiciones constructivas. Una vez puesta en servicio, la geometría del pavimento se modifica lentamente en función del paso del tránsito, evolucionando hacia valores más elevados del IRI (mayores irregularidades) (Solminihac, H, 2005).

3.3.1.2. Cálculo del Índice de Regularidad Internacional (IRI).

El cálculo del IRI involucra la utilización de herramientas matemáticas, estadísticas y computacionales que permiten derivar la medida de regularidad asociada al camino. El primer paso del procedimiento para el cálculo del IRI, y el más importante de todos consiste en medir las cotas o elevaciones de terreno que permitan representar el perfil real de camino. Esto significa que el IRI es independiente de la técnica o equipo utilizado para obtener el perfil, y dependerá únicamente de la calidad del perfil longitudinal (Badilla, 2009).

Estos datos son sometidos a un primer filtro, en el cual se realiza un análisis estadístico (media móvil) y adecuaciones matemáticas para generar un nuevo perfil posible de ser analizado desde el punto de vista de las irregularidades que se pudieran observar. Al nuevo perfil generado se le aplica un segundo filtro, el cual consiste en la aplicación de un modelo de cuarto de carro que se desplaza a una velocidad de 80 km/h, a través de este se registran las características asociadas al camino basadas en los desplazamientos verticales inducidos a un vehículo

estándar, el cual es modelado de forma simplificada como un conjunto de masas ligadas entre sí y con la superficie de la carretera mediante resortes y amortiguadores (Badilla 2009).

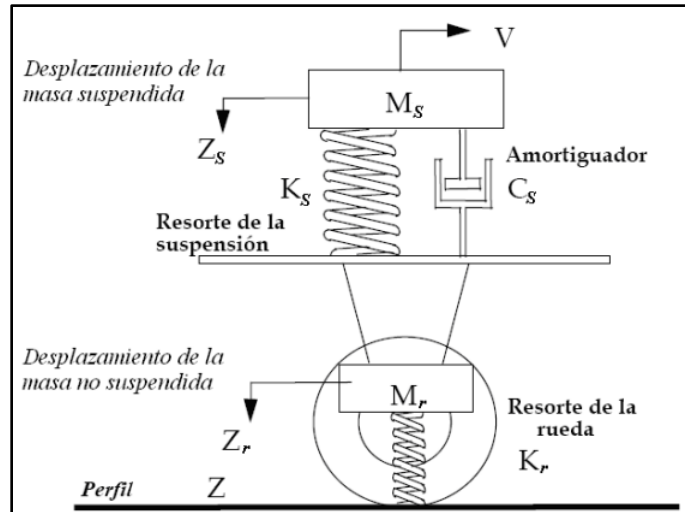


Figura 4. Modelo cuarto de carro (Badilla 2009).

El modelo cuarto de carro, como se observa en la figura 4, realiza una simulación a partir de una masa amortiguada o suspendida (masa de un cuarto de carro ideal) conectada a una masa no amortiguada (eje neumático), a través de un resorte y un amortiguador lineal (suspensión), y por último el neumático es representado por otro resorte lineal (Badilla 2009).

El modelo de cuarto de carro emplea los parámetros de lo que se denomina como el Carro de Oro, los cuales se muestran a continuación:

$$k_2 = \frac{K_s}{M_s} = 63.3s^{-2} \quad k_1 = \frac{K_r}{M_r} = 653s^{-2} \quad c = \frac{C_s}{M_s} = 6s^{-1} \quad u = \frac{M_r}{M_s} = 0.15$$

Donde:

K_s =Constante del resorte de la suspensión kg/s².

K_r =Constante del resorte de la rueda kg/s².

M_s =Masa suspendida, kg.



M_r =Masa no suspendida kg.

C_s =Amortiguador kg/s.

Las ecuaciones dinámicas presentes en el modelo, forman un sistema de ecuaciones que utilizan como dato de entrada el perfil de la carretera (en la parte inferior del resorte del neumático). El movimiento vertical del eje respecto a la masa suspendida se calcula y acumula. El valor en m/km (metros acumulados por kilómetro viajado) es la medida final de la regularidad del camino (Badilla 2009).

3.3.2. Condiciones superficiales obtenidos de la auscultación.

3.3.2.1. Profundidad de Roderas (PR).

La profundidad de rodera son los surcos o huellas que se presentan en la superficie de una carretera pavimentada y son el resultado de la densificación o movimiento lateral de los materiales que la constituyen por efectos del tránsito. Así mismo, es la deformación vertical permanente del pavimento que se refleja en el perfil longitudinal y que se presenta a lo largo del camino bajo las huellas de rodamiento. Geométricamente se define como la máxima depresión por huella en el sentido perpendicular al eje del camino, como se aprecia en la figura 5. La presencia de roderas en el pavimento afecta no solo la condición estructural del pavimento, sino también, en los niveles extremos, afecta su condición funcional dificultando las condiciones de manejo y la seguridad de los usuarios (Determinación de la profundidad de roderas. N-CSV-CAR-1-03-009/16).

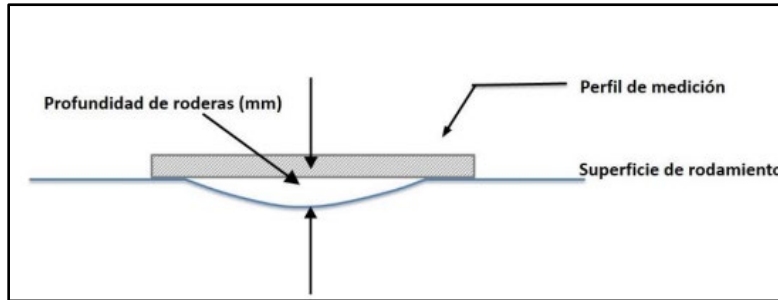


Figura 5. Esquema de una rodera (Determinación de la profundidad de rodera, N-CSV-CAR-1-03-009/16)

3.3.2.2. Macrotextura (MAC).

Las macrotexturas se refieren a las irregularidades de un pavimento con respecto a una superficie plana verdadera, las longitudes de onda de la macrotextura se encuentran en un intervalo de 0.5 a 50 mm, la amplitud de pico a pico varía normalmente entre 0.01 y 20 mm. (Determinación de la macrotextura. N-CSV-CAR-1-03-006/16, normativa SCT)

La macrotextura, como se aprecia en la figura 6, se refiere a la textura superficial del pavimento proveniente del efecto conjunto de las partículas de los agregados pétreos que sobresalen de la superficie. La macrotextura es importante para proporcionar canales de salida de agua en la interacción neumático-pavimento, evitando de esta forma que cause el efecto hidropneumático.

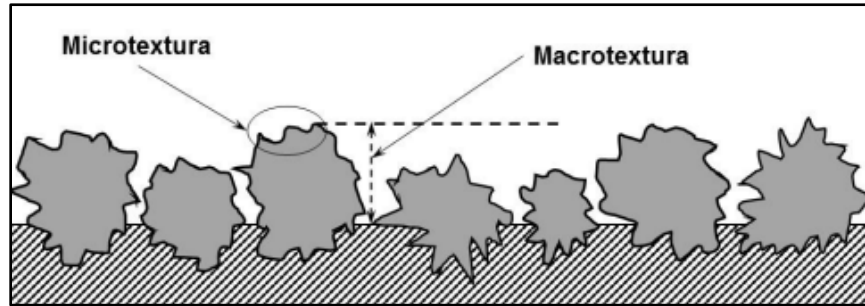


Figura 6. Macrotextura (Determinación de la macrotextura, N-CSV-CAR-1-03-006/16)

3.3.2.3. Deterioros Superficiales (DET).

La determinación de los deterioros es el conjunto de trabajos de campo y su procesamiento, necesarios para conocer los daños del pavimento de un tramo carretero o un conjunto de ellos que permita definir su estado físico. Los deterioros superficiales son los defectos visibles en la superficie del pavimento que evidencian su degradación por efecto de las cargas del tránsito, los agentes medioambientales, las características de los materiales o la interacción entre ellos. El tipo de deterioro, su extensión y su nivel de severidad de acuerdo con el estado físico que presenten, son indicativos de su capacidad estructural y su capacidad funcional. El origen y tipo de los deterioros se determinan en función del tipo de pavimento que se esté analizando, ya sea asfáltico o de concreto hidráulico (Determinación de los deterioros superficiales de los pavimentos. N-CSV-CAR-1-03-008/18, normativa SCT).

3.4. Aprendizaje Automático.

El aprendizaje automático es la rama de la inteligencia artificial que busca que un programa de computador aprenda de un conjunto de datos con los cuales se entrena, y buscará identificar un patrón con el que pueda realizar predicciones sobre nuevos datos.

Pineda (2021) menciona que en el aprendizaje automático o machine learning los datos y las salidas son los datos iniciales que mediante un proceso de entrenamiento producen reglas, las cuales comúnmente reciben el nombre de



modelo, siendo dicho modelo el resultado de detectar en los datos patrones o tendencias que se pueden usar para hacer predicciones sobre datos nunca vistos. Este proceso en mención es realizado por algoritmos, que debidamente ajustados, a final son los que van a permitir que puedan lograrse buenos resultados predictivos. Tales algoritmos se clasifican dependiendo el tipo de problema de aprendizaje que se esté abordando, en donde deben de afinarse sus parámetros, en la mayoría de los casos mediante prueba y error, a fin que se consigna el mejor desempeño del modelo. Por esta razón, es que muy difícilmente una misma solución se pueda replicar en diferentes escenarios, puesto que cada proceso está supeditado tanto a la naturaleza de los datos que utiliza como insumo, como a los ajustes que en materia de parámetros se realizan sobre el algoritmo seleccionado para una determinada solución.

En el capítulo siguiente se abordará el debido proceso que se debe de seguir en el aprendizaje automático.

3.4.1. Proceso del aprendizaje automático.

El aprendizaje automático implica la realización de una serie de etapas que se muestran en la figura 7.



Figura 7. Etapas del aprendizaje automático (Fuente: Pineda. C, (2021).)



Las etapas del aprendizaje automático de acuerdo con Pineda (2021) se describen a continuación:

- Preprocesamiento de los datos. Es una de las etapas más importantes, dado que en ella se realizan tareas como limpieza y transformación de los datos para que queden de una forma adecuada y puedan ser utilizados por el algoritmo de aprendizaje automático.
- Separación en conjunto de entrenamiento y pruebas. En aprendizaje automático el conjunto de datos suele separarse en dos subconjuntos llamados entrenamientos y prueba. El primero se destina para entrenar y estimar los parámetros del modelo. Por otro lado, el segundo se usa para hacer predicciones y probar el modelo con datos diferentes a los de entrenamiento para ver si arroja los resultados esperados.
- Configuración del algoritmo. En esta etapa, básicamente se crea una instancia del algoritmo a utilizar y se define los llamados hiperparámetros para ese algoritmo con valores apropiados que el científico de datos o programador debe de ir ajustando. Los hiperparámetros se diferencian de los parámetros en que estos últimos los define el modelo internamente durante el entrenamiento, no teniendo el usuario ninguna participación directa en la generación de sus valores. Algunos hiperparámetros empleados con regularidad de machine learning son el número de épocas y la tasa de aprendizaje.
- Entrenamiento del modelo. Consiste en proporcionarle al objeto o instancia del algoritmo de aprendizaje automático un conjunto de datos de entrenamiento para que pueda aprender, logrando así estimar los parámetros del modelo de aprendizaje.
- Predicción. Una vez generado el modelo se puede probar su nivel de predicción pesándole muestras del conjunto de pruebas. El resultado de la predicción sobre una muestra es un valor continuo o discreto que debe de ser lo más cercano posible al valor esperado.



- Evaluación. La evaluación es el proceso de determinar numéricamente que tan efectivo fue nuestro modelo de aprendizaje automático. Esta efectividad en el rendimiento parte del supuesto de que a menor diferencia entre la salida esperada y la salida predicha mejor es la evaluación.

3.4.2. Preprocesamiento.

La recopilación de los datos debe de ir acompañada de una limpieza de los mismos, para que estos estén en condiciones para su análisis. Los beneficios del análisis y de la extracción del conocimiento a partir de datos dependen, en gran medida, de la calidad de los datos recopilados. Además, generalmente, es necesario realizar una transformación de los datos para obtener una materia prima que sea adecuada para el propósito concreto y las técnicas que se quieren emplear. En definitiva, el éxito de un proceso de aprendizaje automático depende, no solo de tener todos los datos necesarios (una buena recopilación), sino de que éstos estén íntegros, completos y consistentes.

En los siguientes temas se describirán una serie de técnicas para la limpieza de los datos, como son histogramas, detección de valores anómalos, y transformaciones como la normalización.

3.4.2.1. Valores faltantes.

La detección de los valores faltantes, perdidos o ausentes (missing values) puede parecer sencilla, pero a la vez importante, si los datos proceden de una base de datos, basta mirar en la tabla resumen de atributos/características y ver la cantidad de datos nulos que tiene cada atributo.

El tratamiento o las acciones de los datos faltantes pueden ser las siguientes:

- Ignorar: algunos algoritmos son robustos a datos faltantes (por ejemplo, los árboles de decisión).
- Eliminar toda la columna: solución extrema, pero a veces la proporción de nulos es tan alta que la columna no tiene arreglos. Otras veces, existe otra columna dependiente con datos de mayor calidad.



- Reemplazar el valor: se puede intentar reemplazar manualmente o automáticamente por un valor que preserve la media o la varianza, en el caso de valores numéricos, o por el valor moda, en el caso de valores nominales.
- Modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

Quizás una de las soluciones anteriores más frecuentes cuando el algoritmo a utilizar no maneja bien los nulos sea reemplazar el valor. Si sustituimos un dato faltante por un dato estimado, hemos de tener en cuenta que, en primer lugar, perdemos información, ya que ya no se sabe que el dato era faltante y, en segundo lugar, la información introducida puede ser errónea o provocar sesgo.

3.4.2.2. Valores atípicos.

La detección de valores erróneos en atributos numéricos suele empezar por buscar valores anómalos, atípicos o extremos (outliers), también llamados datos aislados, exteriores o periféricos. Es importante destacar que un valor erróneo y un valor anómalo no son lo mismo. Existen casos en los que los valores extremos se categorizan como anómalos estadísticamente, pero son correctos, es decir, representan un dato fidedigno de la realidad. No obstante, así y todo, pueden ser un inconveniente para algunos métodos que se basan en el ajuste de pesos, por ejemplo, las redes neuronales.

El no detectar un valor anómalo puede ser un problema si el atributo se normaliza posteriormente, ya que la mayoría de datos estarán en un rango muy pequeño y puede haber poca precisión o sensibilidad para algunos métodos de aprendizaje automático.

3.4.2.3. Transformación de atributos. Normalización.

La transformación de datos engloba, en realidad, cualquier proceso que modifique la forma de los datos.

Es necesario normalizar todos los atributos al mismo rango, ya que la mayoría de los algoritmos están basados en distancias, ya que las distancias debidas a



diferencias de un atributo que van entre 0 y 100 serán mucho mayores que las distancias debidas a diferencias de un atributo que va entre 0 y 10.

La normalización más común es la normalización lineal uniforme y se normaliza a una escala entre cero y uno utilizando la siguiente formula:

$$v' = \frac{v - \min}{\max - \min}$$

El resultado de esta normalización es que la relación (el cociente) entre los valores se mantiene. Para realizar esta normalización sólo es necesario conocer el máximo y mínimo de los valores dados para ese atributo.

3.4.3. Agrupamiento de los datos.

Para Pineda (2021) el agrupamiento (clustering) es la tarea descriptiva y consiste en obtener grupos naturales a partir de los datos. Hablamos de grupos y no de clases, porque, a diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo.

Un algoritmo muy común el agrupamiento de los datos es a través de K-medias, tema que se hablara en el siguiente tema.

3.4.3.1. K-Medias.

En el algoritmo K-medias no es necesario etiquetar los datos, ya que el algoritmo es el encargado de encontrar las características en común que estos tienen y crea k grupos donde va separando las muestras que comparten esos rasgos similares.

Los pasos del algoritmo son básicamente los siguientes (Pineda, 2021):

1. Especificar la cantidad k grupos o clúster a generar.
2. Cada grupo tendrá un punto central llamado centroide, los cuales inicialmente se seleccionan aleatoriamente.



3. Una vez se tengan k centroides, se agrupan en un clúster las muestras del conjunto de datos que comparten el centroide más cercano formando un grupo. Dicha cercanía es medida usualmente mediante la distancia euclidiana.
4. Mover los centroides hacia el centro de cada clúster actualizando su posición a partir de las medias de todas las muestras del clúster al que pertenece. Por esta razón a este algoritmo se le llama K-medias.
5. Se repiten los pasos 3 y 4 hasta que el modelo converja, que los centroides alcancen un movimiento mínimo o se alcance un número máximo de iteraciones establecido por el programador.

Con los pasos anteriores se busca minimizar las distancias de las muestras dentro del clúster y maximizar la distancia entre clústeres. Un paso importante es el de asignar el número o la cantidad k grupos, por lo cual existe un método que nos arroja el número k óptima de acuerdo a la base de datos.

3.4.3.2. Método del codo.

Para Pineda (2021) el aspecto más crítico de usar el algoritmo k -medias es encontrar el valor óptimo para el hiperparámetro k . El método del codo usa los valores de la inercia una vez se ha aplicado k -medias con un número n de clústeres. La inercia es la suma de las distancias al cuadrado de cada punto del clúster a su centroide. Matemáticamente se expresa de la siguiente forma:

$$\text{Inercia} = \sum_{i=1}^n \|x_i - \mu\|^2$$

El agrupamiento de los datos tiene el objetivo de facilitar el aprendizaje de los diversos algoritmos, ya que se crean diferentes grupos que comparten similitudes entre sí. El aprendizaje automático tiene la finalidad de realizar diferentes tareas, cada de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo del aprendizaje automático. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir



mucho de la obtenida con otra. Uno de las tareas del aprendizaje automático es el desarrollo de modelos de estimación, tema que a continuación se detalla.

3.4.4. Modelos de estimación del deterioro de pavimentos

De acuerdo con Leiva (2004) los modelos de deterioro generalmente, corresponde a expresiones matemáticas que permiten predecir la posible evolución del estado del pavimento en el tiempo, con base en el conocimiento de las condiciones al momento de la puesta en servicio y al momento de la realización del análisis.

Existen modelos de deterioro de pavimentos uno de los más conocidos y utilizados es el propuesto por el Banco Mundial, denominado Highway Design and Maintenance Standards Model HDM, cuya última versión HDM-4 ha sido llevada a cabo por grupos de investigadores situados en diversos países. El HDM-4 cual tiene dos objetivos, el primero es la realización de los estudios necesarios para desarrollar los modelos que permiten evaluar a los pavimentos, el segundo es el desarrollo de un Software para evaluar proyectos viales a nivel proyecto, a fin de optimizar la inversión en construcción y mantenimiento de un conjunto de caminos (Solminiach, H, 2005).

Además del HDM-4 también se tiene al Sistema mexicano de Administración de Pavimentos (SIMAP), desarrollado por el Instituto Mexicano del Transporte, el cual es un conjunto de actividades relacionadas con los procesos de organización, coordinación y control que afecten la funcionalidad, economía y vida útil de los pavimentos y que permitan una utilización adecuada de los recursos presupuestales disponibles (Rodríguez, R, Orozco y Orozco, Gutiérrez, T, & García, P)

Actualmente existen modelos de deterioro basados en algoritmos de inteligencia artificial, dichos algoritmos son regresión lineal, redes neuronales artificiales, arboles de decisión y modelos basados en reglas.

3.4.4.1. Regresión lineal.

La regresión lineal es el proceso en el que una variable dependiente de la condición observada o medida se relaciona con una o más variables independientes, como aplicaciones de carga por eje o tráfico, espesores y propiedades de la capa de



pavimento, resistencia de la subrasante, factores ambientales y similares. Los modelos de regresión son particularmente aplicables cuando se ha adquirido una buena base de datos histórica (Robinson, R, Danielson, U & Snaith, M 1998).

Para Jiawei (2011) los datos de una regresión lineal se modelan utilizando una línea recta. La regresión lineal modela una variable aleatoria Y (llamada variable de respuesta), en una función lineal de otra variable aleatoria X (llamada variable predictora) es decir:

$$Y = \alpha + \beta X$$

Donde se supone que la varianza de Y es constante, mientras “ α ” y “ β ” son coeficientes de regresión que especifican la intersección con Y y la pendiente de la línea, respectivamente. Estos coeficientes se pueden resolver mediante el método de mínimos cuadrados, que minimiza el error entre la línea real que separa los datos y la estimación de la línea (Jiawei Han, 2011).

En la figura 8 representa un modelo de regresión lineal, donde la línea negra representa la recta de mejor ajuste dada por la ecuación que permite realizar predicciones, y los puntos rojos representan los datos reales.

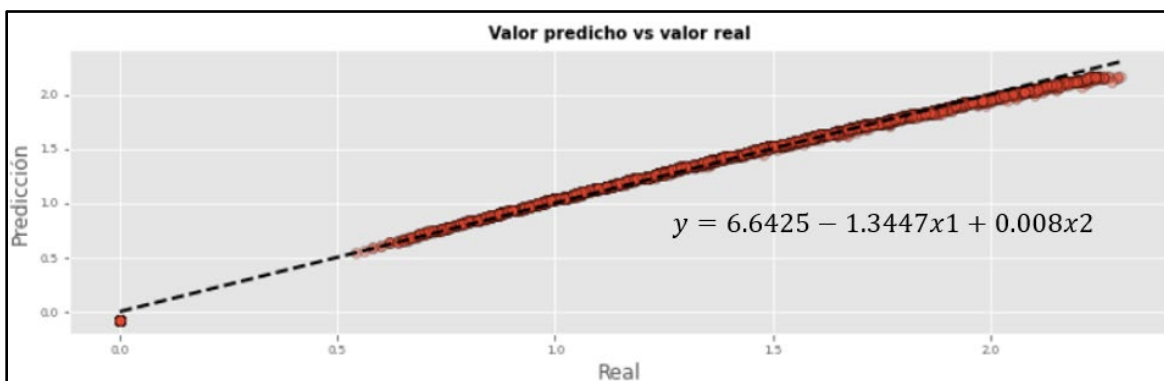


Figura 8. Gráfica de regresión lineal. (Fuente: Elaboración propia.)

3.4.4.2. Redes Neuronales Artificiales.

Una red neuronal artificial es un sistema de procesamiento de información que tiene ciertas características de desempeño en común con las redes neuronales

biológicas. Una red neuronal consta de una gran cantidad de elementos de procesamiento simples llamados neuronas, unidades, células o nodos. Cada neurona está conectada a otras neuronas por medio de enlaces de comunicación dirigidos, cada uno con un peso asociado. Los pesos representan la información que utiliza la red para resolver un problema (Fausett, 1994).

Un ejemplo de una red neuronal simple esta dado por la figura 9 donde se tiene una neurona Y , que recibe entradas de las neuronas X_1, X_2 y X_3 . Las activaciones o señales de salida de estas neuronas son X_1, X_2 y X_3 . Respectivamente. Los pesos de las conexiones de X_1, X_2 y X_3 . A la neurona Y son W_1, W_2 y W_3 respectivamente. Por lo cual la entrada a la neurona Y es la suma de las señales ponderadas de las neuronas X_1, X_2 y X_3 . Además, la neurona Y está conectada a las neuronas Z_1 y Z_2 con pesos V_1 y V_2 respectivamente. La neurona Y envía su señal a cada una de las unidades, los calores recibidos por las neuronas Z_1 y Z_2 serán diferentes (Fausett, 1994).

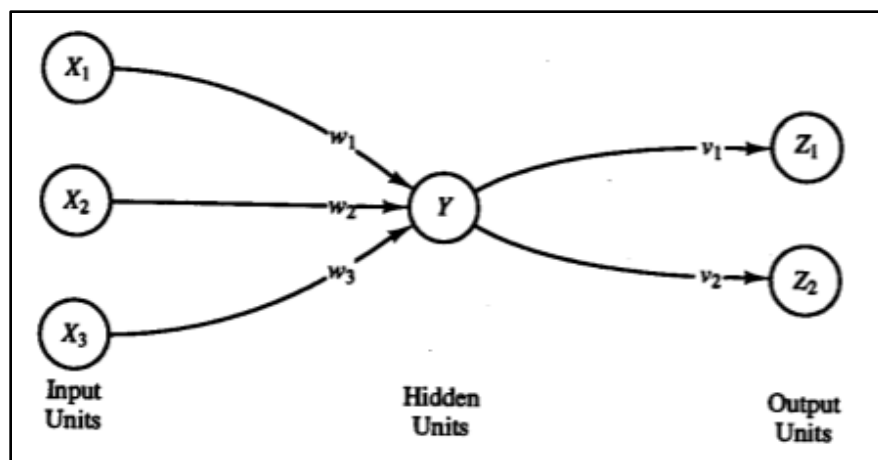


Figura 9. Red neuronal simple (Fausett ed., al 1994).

3.4.4.3. *Random Tree.*

Los árboles de clasificación incorporan un enfoque de clasificación supervisada, su estructura es similar a la de un árbol que se compone de una raíz, nodos (las posiciones donde las ramas se dividen), ramas y hojas, de manera similar, un árbol de clasificación se construye a partir de nodos que representan los círculos y las

ramas son representadas por los segmentos que conectan los nodos. Un árbol de clasificación se inicia desde la raíz, se extiende hacia abajo y generalmente se dibuja de izquierda a derecha. El nodo inicial se llama nodo raíz, mientras los nodos en los extremos de la cadena se les conocen como nodos hoja. Dos o más ramas pueden extenderse desde cada nodo interno, es decir, desde un nodo que no es el nodo hoja (Jehad, Ali, 2012).

La figura 10, describe la estructura general de un árbol de decisión.

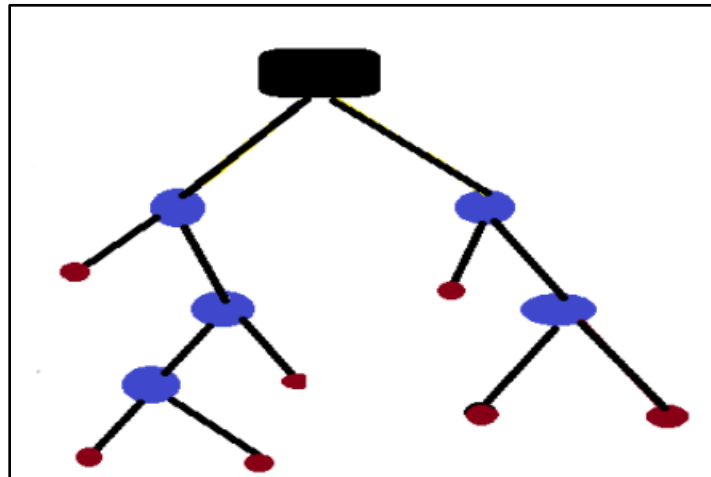


Figura 10. Estructura general de un árbol de decisión (Jehad, Ali 2012).

3.4.4.4. *Random Forest.*

Para Rosa Fátima Medina (2017) el algoritmo Random Forest surge como la agrupación de varios árboles de clasificación, básicamente selecciona de manera aleatoria una cantidad de variables con los cuales se construye cada uno de los árboles individuales, y se realizan predicciones con estas variables que posteriormente serán ponderadas a través del cálculo de la clase más votada de los árboles que se generaron, para finalmente hacer la predicción por Random Forest.

La figura 11, representa la estructura de del Random Forest.

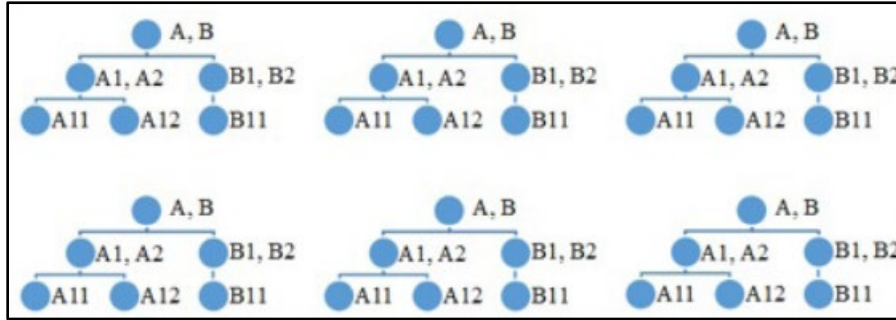


Figura 11. Estructura de un bosque aleatorio (Medina, Rosa 2017).

3.4.4.5. Reglas de asociación. M5 Rules.

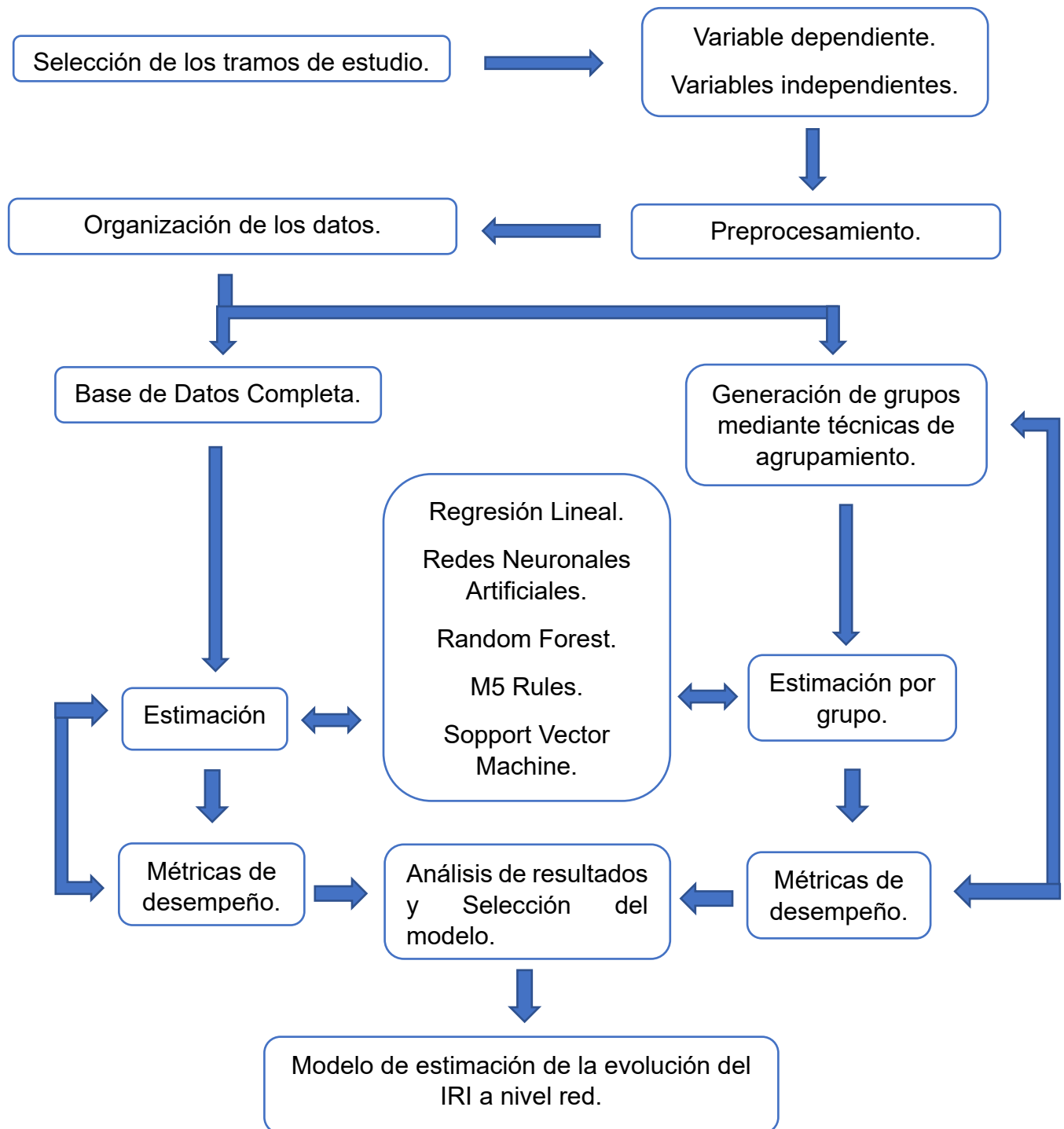
Las reglas de asociación se centran en las técnicas específicas para el aprendizaje de reglas de asociación y dependencias. Estas reglas expresan patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más atributos. Las reglas de asociación son una manera muy popular de expresar patrones de datos de una base de datos. Estos patrones pueden servir para conocer el comportamiento general del problema que genera la base de datos, y de esta manera, se tenga más información que pueda asistir en la toma de decisiones.

Para Hernández, una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertos estados en una base de datos. Por lo cual, una regla de asociación puede ser vista como reglas de la forma SI α ENTONCES β , donde α y β son dos conjuntos de características disjuntos. Otra forma muy utilizada de expresar una regla de asociación es $\leftarrow \alpha$, o también $\alpha \Rightarrow \beta$.



4. Propuesta de Solución.

La metodología a utilizar en el desarrollo de la presente investigación se muestra en el siguiente diagrama.





Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



La selección de los tramos de estudio es el primer paso. A partir de la base de datos de mediciones históricas del Índice de Regularidad Internacional IRI se seleccionan un conjunto de carreteras que cuenten con información histórica del Índice de Regularidad Internacional (variable independiente), así como temperatura máxima y mínima, altura de precipitación condiciones de tránsito, deflexiones, profundidad de rodera y porcentaje de agrietamiento, considerados como variables independientes.

Existen varios factores que se consideran al momento de seleccionar los tramos de estudio, uno es que debe de tener información histórica de la variable dependiente y de las variables independientes. Otro factor es que los tramos seleccionados deben de contar con la mínima cantidad de datos perdidos o faltantes en sus variables.

Una vez que se tiene la base de datos con información de la variable dependiente y de las variables independientes, se realiza el preprocesamiento. Para Jiawei (2011) el preprocesamiento funciona para limpiar los datos, completando valores faltantes, suavizando los datos, identificando o eliminando valores atípicos.

Los valores faltantes son eliminados de la base de datos. Los valores atípicos son identificados utilizando el rango de valores máximos y mínimos que se presenta en la norma de la secretaria de Comunicaciones y Transportes (SCT) N-CSV-CAR-1-03-004/16, estos valores son eliminados de la base de datos. El suavizado de los datos es la transformación de las diversas variables que forman la base de datos a través de ecuaciones matemáticas. La técnica a utilizar será la normalización, proceso donde los datos de las variables se escalan para caer dentro de un pequeño rango especificado de 1 a 0 (Jiawei, 2011).

La organización de los datos consiste en dividir la base de datos en condiciones del sentido de la carretera (sentido 1 o 2), dependiendo si el kilometraje asciende o desciende, así mismo se divide en la cantidad de carriles 1 o 2, dependiendo de la geometría de la carretera.



Posterior al preprocesamiento y la organización de los datos, comienza la predicción del Índice de Regularidad Internacional IRI (variable dependiente), la cual se realiza a través de la base de datos completa o aplicando un Clustering (agrupamiento).

Utilizando la base de datos completa, se le aplican diferentes algoritmos de aprendizaje artificial, los cuales son:

- Regresión lineal.
- Redes Neuronales Artificiales.
- Radom Forest.
- M5 Rules.
- Soppot Vector Machine.

Una vez generados los diferentes modelos con sus respectivos algoritmos se analizan los resultados y se observa si el modelo se comporta de manera correcta, para este paso se utiliza las diferentes medidas de bondad de ajuste que cuenta el aprendizaje automático como lo son:

- ✓ Coeficiente de Correlación. Es la medida que representa la intensidad de la relación entre la variable dependiente con las variables dependientes, un resultado cercano a la unidad represente fuerte relación entre las variables de estudio, mientras que un resultado cercano a 0 representa una nula relación entre variables.
- ✓ Error absoluto medio (MAE). Representa la diferencia entre las mediciones reales con respecto a las predicciones, por lo cual se busca un valor cercano a cero. La fórmula que denomina al MAE es la siguiente:

$$MAE = \frac{\sum_{i=1}^N |Y_i - X_i|}{N}$$

- ✓ Error cuadrático medio (RMSE). Al igual que el MAE representa la diferencia entre las mediciones reales con las predicciones, pero esa diferencia es elevada al cuadrado, por lo cual este error es más sensible a los valores que



el algoritmo detecte como atípicos, de igual manera se busca un valor cercano a cero. La fórmula que denomina al RMSE es la siguiente:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N |Y_i - X_i|^2}{N}}$$

- ✓ Error absoluto relativo (RAE). Medida que permite observar la precisión del modelo, así como el porcentaje de las veces que se equivoca a la hora de predecir. La fórmula que denomina al RAE es la siguiente:

$$RAE = \frac{|P_1 - a_1| + \dots + |P_n - a_n|}{|a_1 - a| + \dots + |a_n - a|}$$

Una vez finalizada el análisis para la base de datos completa, se realiza el análisis a través de clustering con el objetivo de mejorar los resultados del modelo anterior, esta etapa se realiza a través de algoritmos de agrupamiento de tramos, es decir, formar grupos donde los tramos tengan un comportamiento del Índice de Regularidad Internacional similar.

El agrupamiento de los tramos se realiza a través del algoritmo de clustering, Jiawei menciona que el agrupamiento divide a los objetos en grupos, los grupos se agrupan según el principio maximizar la similitud intraclase y minimizar la similitud interclase.

Lo que se busca es formar grupos donde los objetos o mediciones dentro de un grupo tengan una gran similitud entre sí, pero a la vez sean diferentes a los objetos de otros grupos.

A través del método del codo se conoce el número K de grupos que son necesarios para el tamaño y características de la base de datos. Una vez obtenida el número K de grupos se prosigue a describir las características de los diferentes grupos y se observa el comportamiento del Índice de Regularidad Internacional, la descripción se realiza a través de las distintas variables como son temperatura máxima y



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



mínima, altura de precipitación, condiciones de tránsito, deflexiones, profundidad de rodera y porcentaje de agrietamiento. Es decir, se observa los rangos del Índice de Regularidad Internacional con respecto a los rangos de las diferentes variables. Por ejemplo, para un Índice de Regularidad Internacional bajo, que rangos de clima, transito, deflexiones, profundidad de rodera y porcentaje de agrietamiento le corresponden.

Una vez formados los grupos se realiza la predicción para cada grupo formado en el paso anterior, cada grupo es analizado por los mismos 5 algoritmos de aprendizaje artificial que se utilizaron para la base completa, los cuales son, regresión lineal, redes neuronales artificiales, random tree, random foresty m5 rules.

Posteriormente, se analizan sus resultados a través de las medidas de bondad de ajuste, si los resultados no son los deseados, se tendrá que realizar de nuevo la agrupación, pero ahora disminuyendo y aumentado el número de K grupos, hasta encontrar el K grupos con los satisfagan las medidas de bondad.

Finalizado la predicción para la base de datos completa, así como la predicción por grupo se comparan los resultados de las medidas de bondad, y se toma la decisión de predicción se acerca a los datos reales y así tener un Modelo de Predicción de la Evolución del Índice de Regularidad Internacional a Nivel Red.



5. Resultados.

En este capítulo, se especifica el proceso metodológico que se utilizó para obtener la información que se necesita para la elaboración de la presente investigación. Se especifican los elementos necesarios, así como el procedimiento para el desarrollo del modelo de estimación del Índice de Regularidad Internacional (IRI).

5.1. Aspectos generales del tramo de estudio.

La selección de la carretera se realizó de acuerdo a la información disponible de diversas carreteras de la República Mexicana. La elección de los datos se obtuvo con base en la información histórica de la variable dependiente, así como de las variables independientes. Al momento de la elección de la carretera se formó la base de datos y se comprobó que la base de datos de los años 2019 y 2020 contara con la menor cantidad de datos faltantes o perdidos, y de datos atípicos que causaran ruido al momento del aprendizaje de los algoritmos de inteligencia artificial.

La carretera seleccionada fue la Hermosillo – Santa Ana ubicada en el estado de Sonora.

Sonora es uno de los treinta y uno estados que conforman México. Está ubicado en la región noroeste del país, limitado al norte con Arizona (Estados Unidos) y con Nuevo México (Estados Unidos), al este con Chihuahua, al sur con Sinaloa y al oeste con el golfo de California y con Baja California. En la figura 12 se muestra la ubicación del estado de Sonora, así como la red de caminos.

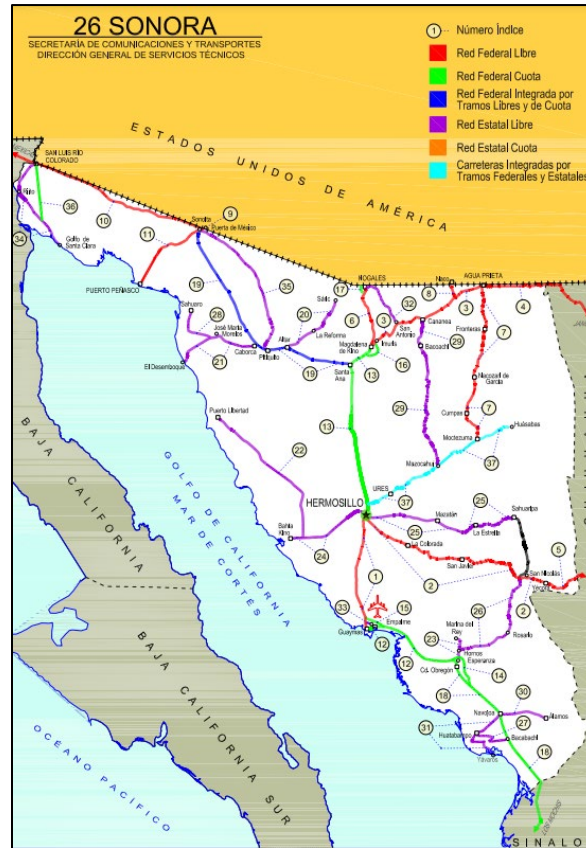


Figura 12. Macro localización del tramo de estudio.

El objeto de la presente investigación, es la realización de un modelo de estimación del Índice de Regularidad Internacional con la aplicación de varios algoritmos de inteligencia artificial, de una carretera piloto, de la cual se muestran los siguientes aspectos generales:

- Ubicada en el Estado de Sonora.
- Carretera Hermosillo – Santa Ana.
- Ruta MEX-015D
- Longitud 163.1 km
- Carretera Federal de Cuota de 4 carriles (2 por sentido).
- Coordenadas



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



	Latitud	Longitud
Coordenadas Inicio	29.097416	-110.930946
Coordenadas Fin	30.618561	-110.995628

La ubicación de dicha carretera se muestra en la figura 13.

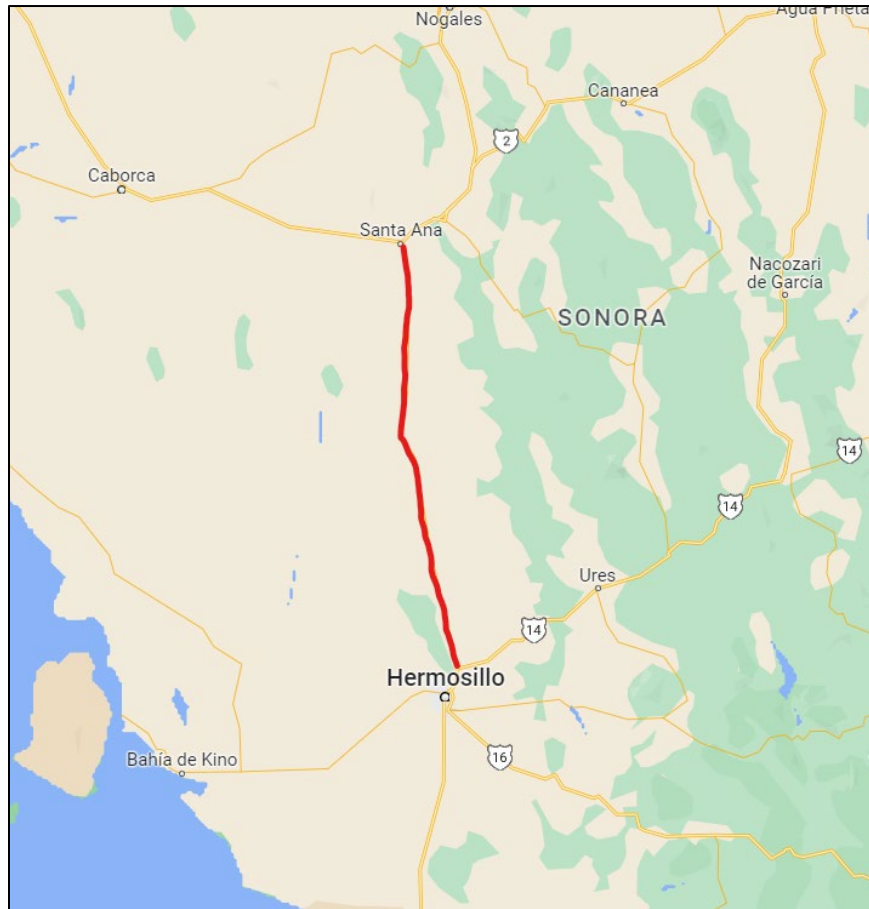


Figura 13. Micro localización del tramo de estudio.

5.6. Selección de las variables.

Los factores que afectan la rugosidad de la superficie del pavimento son Profundidad de Rodera, Macrotextura, Porcentaje de Agrietamiento. Las variables se utilizaron como variables de entrada para desarrollar un modelo de estimación de la variable dependiente (Índice de Regularidad Internacional). Para Kargah (2010), existen muchos factores que afectan el Índice de Regularidad Internacional,



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



aunque la mayoría de estos factores deben tenerse en cuenta en el modelo para que sea completo, la gran cantidad de variables hace que el proceso de adquisición de datos y estimación del Índice de Regularidad internacional sea complicado y costoso.

La información de las variables del tramo de estudio que se utilizaron para desarrollar la presente investigación fue proporcionada por el Instituto Mexicano del Transporte. Las mediciones de las variables fueron extraídas del paquete de información producido por la aplicación anual del “Programa de auscultación de la Red Carretera Federal” por parte de la Dirección General de Servicios Técnicos de la Secretaría de Comunicaciones y Transportes (SCT).

De acuerdo con la información disponible se recopilaron las mediciones las variables dependientes e independientes que representan el comportamiento del Índice de Regularidad Internacional a lo largo del tiempo, por lo cual la base de datos que conforma la carretera Hermosillo – Santa Ana está dada por dos años consecutivos 2019 y 2020, de los cuales contienen información de las variables que a continuación se mencionan:

- Índice de Regularidad Internacional (IRI) = Variable Dependiente.
- Profundidad de Rodera (PR) = Variable Independiente 1.
- Macrotextura (MAC) = Variable Independiente 2.
- Porcentaje de Agrietamientos = Variable Independiente 3.

Por lo cual, para cada lectura del IRI en cada punto específico del pavimento, le corresponde una lectura de cada variable independiente para ese momento en específico, es decir, cada punto específico del pavimento le corresponde una lectura del IRI, así como una lectura del PR, MAC y Agrietamientos, medidos en la misma fecha. Por lo tanto, el número de registros de datos del IRI fue el mismo que el número de mediciones de la PR, MAC y del porcentaje de agrietamientos. Lo cual el número total de mediciones que cuenta la base de datos de la carretera Hermosillo – Santa Ana es de 3138 datos.



5.2. Preprocesamiento de los datos.

Terminada la recopilación de los datos, es necesario realizar un preprocesamiento de los datos. Para García (2016) el preprocesamiento de datos es una etapa esencial del proceso de descubrimiento de información, esta etapa se encarga de la limpieza de datos, su integración, transformación y reducción.

El preprocesamiento de los datos es un paso importante ya que el uso de datos de baja calidad implica un proceso con resultados poco confiables, es por ellos que se hace necesaria la aplicación de diversas técnicas de preprocesamiento. Esto con el objeto de obtener una nueva base de datos adecuada para el proceso de aprendizaje de los algoritmos de inteligencia artificial.

El preprocesamiento de los datos está formado por una serie de técnicas que tienen el objetivo de preparar correctamente los datos que servirán al proceso de aprendizaje para los algoritmos. En esta área se incluye la transformación de datos y normalización, integración, limpieza de ruido e imputación de valores perdidos. (García, S, Ramírez, S, Luengo, J & Herrera, F (2016). Big Data: Preprocesamiento y calidad de datos.)

5.3.1. Eliminación de los datos perdidos.

Como se mencionó anteriormente, para cada punto específico de la carretera le corresponde un grupo de mediciones del IRI, PR, MAC, Agrietamiento. En la base de datos se encontró puntos específicos de la carretera que no contaban con mediciones de alguna de las variables, las cuales se les denomina datos perdidos o faltantes, dichos datos perdidos fueron eliminados de la base de datos original.



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



Cadenamiento	Año	IRI (m/km)	PR (mm)	MAC (mm)	Agrietamiento (%)
15+000	2019	1.58	6.56	1.12	0
15+100	2019	1.46	4.21	1.08	0.02
15+200	2019	2.01	4.22	1.13	0.22
15+300	2019				
15+400	2019				
15+500	2019				
15+600	2019	1.2	2.79	1.09	0
15+700	2019	1.39	3	1.2	0.13

Tabla 4. Ejemplos de algunos datos perdidos de todas las variables.

En la tabla 4, se observa que en los cadenamientos 15+300 al 15+500 no cuenta con ninguna medición de las variables, los cuales se consideran como datos perdidos y son eliminados de la base de datos.

En algunas ocasiones se encontró que solo faltaba la medición de una sola variable, como es el caso de la tabla 5, se observa que a partir del cadenamiento 110+600 al 11+800 no se tiene lecturas de la Profundidad de Roderas (PR), por lo cual se considera como datos faltantes y es eliminada de la base de datos original.

Cadenamiento	Año	IRI (m/km)	PR (mm)	MAC (mm)	Agrietamiento (%)
110+400	2019	2.06	5.43	1.1	36.52
110+500	2019	2.41	5.59	0.92	23.87
110+600	2019	2		0.9	0.48
110+700	2019	2.77		0.86	0.05
110+800	2019	2.42		0.87	0.25
110+900	2019	3.01	6.88	1.03	0.07
111+000	2019	2.51	8.45	1.14	40.18

Tabla 5. Ejemplos de algunos datos perdidos de una variable.



La base de datos original contaba con un total de 3138 datos, de los cuales se encontraron 9 datos faltantes, correspondiente al 0.3% de la base de datos original, dichos datos faltantes fueron eliminados.

5.3.2. Detección y eliminación de los datos atípicos.

La detección de valores atípicos es un problema, ya que se deben de encontrar patrones que no están en el rango de comportamiento normal de la base de datos. Estos patrones anómalos se denominan valores atípicos.

Se pueden seguir varios métodos para la detección de estos datos atípicos, los métodos que se utilizaron fueron el de rango intercuartil y la detección de datos atípicos utilizando las normas de la secretaria de Comunicaciones y Transportes.

5.3.2.1. Rango Intercuartil (IQR)

El primer procedimiento que se siguió para la detección y eliminación de los valores atípicos fue por la técnica del rango intercuartil (IQR).

Para la técnica del intercuartil se necesitaron el primer y tercer cuartil, para cada año diferente y para las distintas variables que compone la base de datos.

Primero se analizó la variable Índice de Regularidad Internacional para el año 2019, la cual tiene los siguientes datos:

$$\text{Cuartil 1} = 1.03$$

$$\text{Cuartil 3} = 1.54$$

Conociendo los cuartiles, se calculó el rango intercuartil que es la diferencia del cuartil 3 con el cuartil 1, como se muestra:

$$\text{IQR} = 1.54 - 1.03 = 0.51$$

Ahora cálculos los límites superiores e inferiores:

$$\text{Límite Superior} = 1.54 + 1.5(0.51) = 2.30$$

$$\text{Límite Inferior} = 1.03 - 1.5(0.51) = 0.265$$



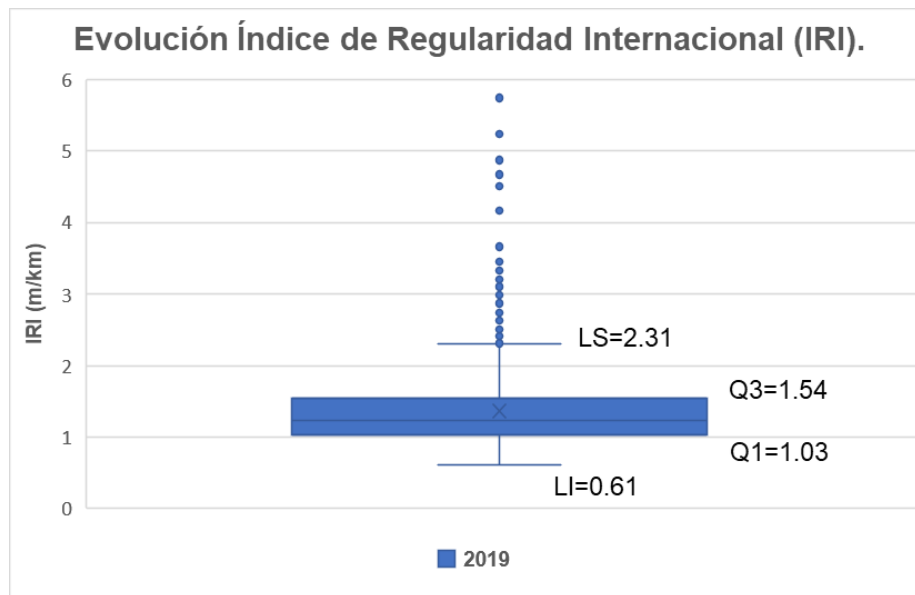
Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



Los valores extremos son todas las mediciones del IRI menores a 0.265 m/km o mayores a 2.30 m/km.

El valor mínimo se consideró como la menor medición del conjunto que sea mayor o igual al límite inferior, en este caso el valor mínimo es de 0.61 m/km. El valor máximo es el mayor elemento que sea menor o igual al límite superior, en este caso el valor máximo es de 2.31 m/km.

El brazo inferior irá desde el primer cuartil hasta el mínimo, desde 1.03 hasta 0.61 m/km, el brazo superior abarcará desde el tercer cuartil hasta el máximo, desde 1.54 hasta 2.31 m/km. Conociendo los puntos se construyó la gráfica de caja como se muestra en la gráfica 1.



Gráfica 1. Evolución del Índice de Regularidad Internacional (IRI) del año 2019.

En la gráfica 1 observamos que no se tienen valores menores al límite superior, pero se observa que se tienen valores superiores al límite superior (2.31 m/km) por los que estos valores son considerados como atípicos y están representados por puntos y se tienen que eliminar de la base de datos.

Después se analizó la misma variable (IRI) pero ahora para el año 2020, para lo cual se tienen los siguientes datos:



$$\text{Cuartil 1} = 1.07$$

$$\text{Cuartil 3} = 1.62$$

Conociendo los cuartiles, se calculó el rango intercuartil que es la diferencia del cuartil 3 con el cuartil 1, como se muestra:

$$\text{IQR} = 1.62 - 1.07 = 0.55$$

Ahora cálculos los límites superiores e inferiores:

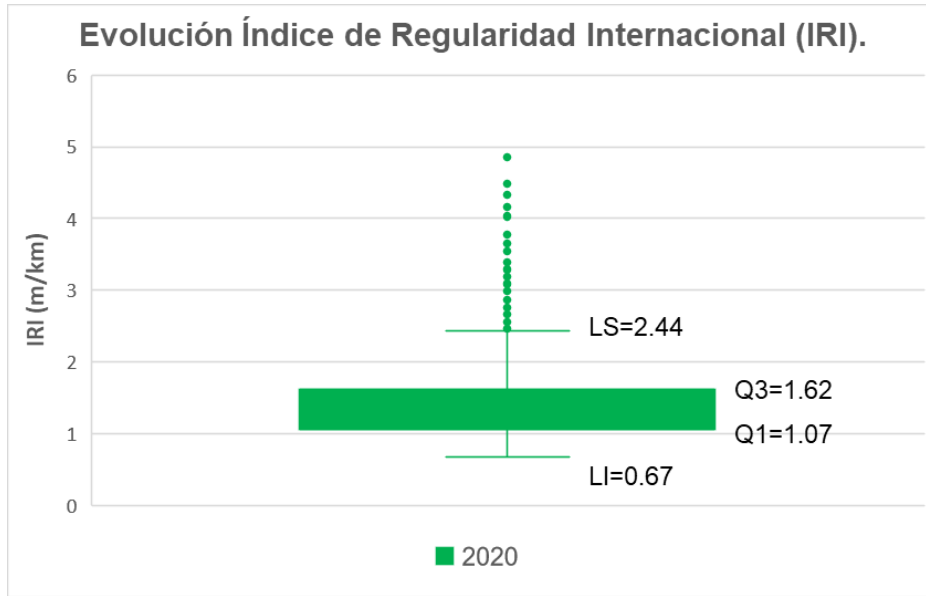
$$\text{Límite Superior} = 1.62 + 1.5(0.55) = 2.44$$

$$\text{Límite Inferior} = 1.07 - 1.5(0.55) = 0.24$$

Los valores extremos son todas las mediciones del IRI menores a 0.24 m/km o mayores a 2.44 m/km.

El valor mínimo se consideró como la menor medición del conjunto que sea mayor o igual al límite inferior, en este caso el valor mínimo es de 0.67 m/km. El valor máximo es el mayor elemento que sea menor o igual al límite superior, en este caso el valor máximo es de 2.44 m/km.

El brazo inferior irá desde el primer cuartil hasta el mínimo, desde 1.07 hasta 0.67 m/km, el brazo superior abarcará desde el tercer cuartil hasta el máximo, desde 1.62 hasta 2.44 m/km. Conociendo los puntos se construyó la gráfica de caja como se muestra en la gráfica 2.



Gráfica 2. Evolución del Índice de Regularidad Internacional (IRI) del año 2020

En la gráfica 2 observamos que no se tienen valores menores al límite superior, pero se observa que se tienen valores superiores al límite superior (2.44 m/km) por los que estos valores son considerados como atípicos y están representados por puntos y se tienen que eliminar de la base de datos.

Como se observa en las dos gráficas 1 y 2, se consideran como datos atípicos a los valores que son superiores a 2.31 y 2.44 m/km respectivamente, esto lo comparamos con los rangos que se presentan en la norma N-CSV-CAR-1-03-004/16.

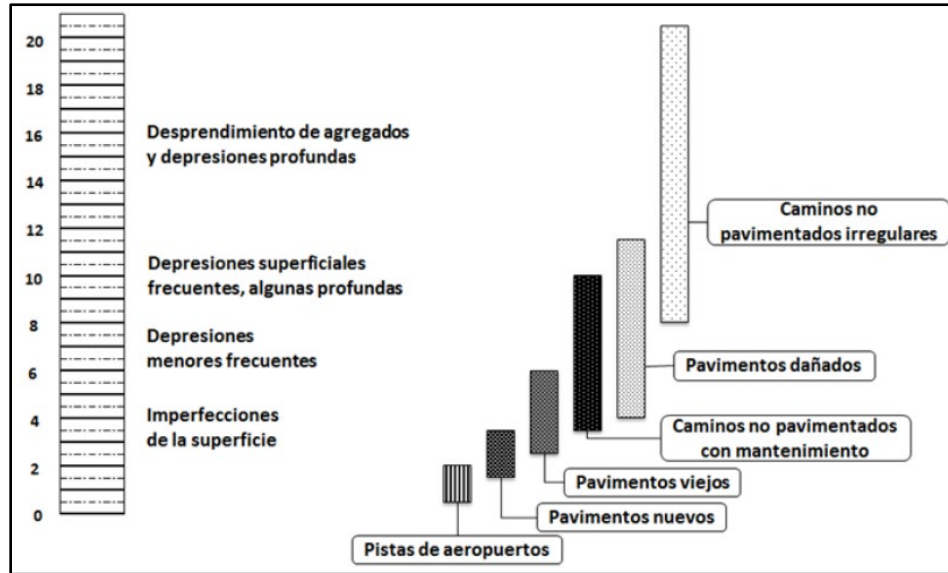


Figura 14. Escala original del Banco Mundial para el IRI. Fuente: Norma de la SCT N-CSV-CAR-1-03-004/16

En la figura 14, observamos que el IRI de una carretera puede variar entre 0 para una superficie perfectamente plana hasta 12 m/km, para pavimentos dañados con depresiones superficiales frecuentes con algunas depresiones profundas, por lo cual, con el método del rango intercuartil (IQR) se están considerando valores atípicos a mediciones que son necesarias para el aprendizaje de los diferentes algoritmos de inteligencia artificial. Motivo por el cual el método del IQR no es factible y no se seguirá analizando las demás variables, razón por la cual se eligió establecer los límites superiores e inferiores de las 4 variables con la ayuda de las normas de la Secretaría de Comunicaciones y Transportes (SCT).

5.3.2.2. Normas de la Secretaría de Comunicaciones y Transportes SCT

Como se vio en la figura 14, de la escala del IRI, se observa que el valor máximo para pavimentos dañados es de 12 m/km, razón por el cual dicho valor se tomará como límite superior, es decir las mediciones mayores serán consideradas como valores atípicos.



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



En la tabla 6 se observan los valores mínimos y máximos del IRI para los dos años de estudio, se observa que no se tienen valores al límite superior que se estableció de 12 m/km, por lo cual para esta variable no se tienen valores atípicos.

Año	Mínimo	Máximo
2019	0.61	5.74
2020	0.67	4.86

Tabla 6. Valores mínimos y máximos del IRI

En seguida se consideró el límite superior de la variable Profundidad de Rodera. La figura 8 se muestra los rangos de valores establecidos en la norma de STC.

Estado	Intervalos de PR mm	
	Autopistas y Corredores Carreteros	Red Básica Libre y Red Secundaria
Bueno	< 5	< 7
Regular	5,1 a 8	7,1 a 9
Malo	> 8	> 9

Tabla 7. Intervalos de profundidad de roderas para la clasificación de los tramos. Fuente: Norma de la SCT N-CSV-CAR-1-03-009/16

En la tabla 7 se observa que una carretera puede estar dentro de los valores de 0 a mayores de 8 mm, mientras mayor sea la medición mayor será el desgaste de la carretera. Por lo cual se estableció como límite superior el valor de 15 mm, es decir, los datos mayores a 15 mm de la variable Profundidad de Rodera (PR) se consideran como datos atípicos.

Para los años 2019 y 2020 se encontraron 28 datos superiores a 15 mm, lo cual son considerados como datos atípicos por lo tanto se eliminaron de la base de datos. Por lo cual el rango de valores de la PR se muestra en la tabla 8.



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



Año	Mínimo	Máximo
2019	1.46	14.35
2020	1.88	14.98

Tabla 8. Valores mínimos y máximos de la PR

Por último, se consideraron los límites de la variable Macrotextura. En la tabla 9 se muestra el rango de valores establecidos en la norma N-CSV-CAR-1-03-006/16.

Estado	Intervalos de la PMT mm	
	Autopistas de Cuota y Corredores Carreteros	Red Básica Libre y Red Secundaria
Bueno	> 0,90	> 0,80
Regular	0,75 a 0,90	0,65 a 0,80
Malo	< 0,75	< 0,65

Tabla 9. Intervalos de la macrotextura para la clasificación de los tramos.
Fuente: Norma de la SCT N-CSV-CAR-1-03-006/16

En la tabla 9 se muestra que una carretera puede estar en el rango de valores de 0 hasta mayores de 0.90 mm, en este caso mientras menor sea la medición mayor será el desgaste que presenta la carretera. En este caso se estableció como límite superior 2 mm, es decir, para datos mayores a 2 mm son considerados como datos atípicos.

Año	Mínimo	Máximo
2019	0.38	1.57
2020	0.43	1.66

Tabla 10. Valores mínimos y máximos de la macrotextura.

Como se observa en la tabla 10, los valores máximos para los años 2019 y 2020 no son mayores al límite superior establecido anteriormente, por lo cual para la variable MAC no existen datos atípicos.



Terminado el análisis de los datos atípicos de las variables, se encontraron y eliminaron un total de 28 datos, correspondiente al 0.9% de la base de datos.

Con la limpieza de la base de datos se eliminaron un total de 38 datos, 9 datos perdidos y 28 datos atípicos, por lo cual la nueva base de datos limpia es de 3000 datos y es utilizada para el aprendizaje de los algoritmos de inteligencia artificial.

5.3.4. Normalización.

La normalización es el proceso por medio del cual los datos de la base de datos son transformados a un rango en específico, ya que, al tener varias variables, cada variable tiene su propio espacio o intervalo de valores, por lo que dificulta el aprendizaje de los modelos de inteligencia artificial, motivo por el cual es necesario la normalización, que en esta investigación se realizó por la técnica mínimos-máximos.

La Normalización Min-Max, transforma los datos de las variables con el objetivo de contar con un espacio o intervalo de 0-1 en todas las variables, esto se llevó a cabo a través de la siguiente fórmula:

$$Dato\ Nuevo = \frac{(Dato\ Original - Valor\ Mínimo)}{(Valor\ Máximo - Valor\ Mínimo)}$$



Cadenamiento	Año	IRI (m/km)	PR (mm)	MAC (mm)	Agrietamiento (%)
9+420	2019	4.67	6.6	1.32	0.3345
9+500	2019	1.51	2.78	0.99	0.3345
9+600	2019	1.27	2.07	1.16	0.0007
9+700	2019	1.15	2.47	1.04	0.0013
9+800	2019	1.18	2.7	1.04	0.0036
9+900	2019	1.39	4.58	0.99	0.0023
10+000	2019	1.42	3.69	1.05	0.0018
10+100	2019	1.61	3.4	0.96	0.0017
10+200	2019	1.66	2.5	1	0.002
10+300	2019	1.37	3.18	0.92	0.0022
10+400	2019	1.27	2.01	0.98	0.0014
10+500	2019	1.4	1.82	0.98	0.0009

Tabla 11. Ejemplo de mediciones de las variables antes del proceso de normalización.

En la tabla 11 se muestra algunos ejemplos de las mediciones de las variables antes de ser normalizados.



Cadenamiento	Año	IRI (m/km)	PR (mm)	MAC (mm)	Agrietamiento (%)
9+420	2019	0.791	0.380	0.734	0.335
9+500	2019	0.175	0.098	0.477	0.335
9+600	2019	0.129	0.045	0.609	0.001
9+700	2019	0.105	0.075	0.516	0.001
9+800	2019	0.111	0.092	0.516	0.004
9+900	2019	0.152	0.231	0.477	0.002
10+000	2019	0.158	0.165	0.523	0.002
10+100	2019	0.195	0.143	0.453	0.002
10+200	2019	0.205	0.077	0.484	0.002
10+300	2019	0.148	0.127	0.422	0.002
10+400	2019	0.129	0.041	0.469	0.001
10+500	2019	0.154	0.027	0.469	0.001

Tabla 12. Ejemplo de mediciones de las variables después del proceso de normalización.

En la tabla 12 observamos los mismos datos que la tabla 11, pero ahora están normalizados utilizando la formula anterior estableciendo un intervalo 0-1 para las cuatro variables.

5.3. Estadística Descriptiva.

Concluido el preprocesamiento de los datos se realizó un análisis de estadística descriptiva con el objetivo de explorar los datos a fin de identificar sus principales características que permitieron realizar la descripción e interpretación del conjunto de datos.

Para describir el conjunto de datos en necesario realizar un análisis individual de cada variable, para posteriormente realizar el estudio de las relaciones entre variables. Para representar la distribución de valores de las variables cuantitativa se recurre a determinadas medidas numéricas que permiten resaltar las características



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



más importantes de cada variable, como lo son, medidas de tendencia central, medidas de dispersión, percentiles y medidas de forma.

Medidas de tendencia central y dispersión								
Año		Mínimo	Máximo	Media	Mediana	Moda	Varianza	Desviación
2019	IRI (m/km)	0.610	5.740	1.366	1.230	1.200	0.269	0.519
	PR (mm)	1.460	14.350	4.117	3.760	4.140	3.471	1.863
	MAC (mm)	0.380	1.570	1.110	1.120	0.920	0.046	0.215
	Agrietamiento (%)	0.000	1.000	0.082	0.003	0.000	0.054	0.232
2020	IRI (m/km)	0.670	4.860	1.438	1.270	1.140	0.320	0.566
	PR (mm)	1.880	14.980	4.860	4.390	3.870	4.330	2.081
	MAC (mm)	0.430	1.660	1.134	1.110	0.930	0.062	0.250
	Agrietamiento (%)	0.000	1.000	0.079	0.003	0.000	0.053	0.229

Tabla 13. Estadística descriptiva. Medidas de tendencia central y dispersión

En la tabla 13 se muestra el rango de valores de las medidas de tendencia central y dispersión de cada variable en cada año. En la primera y segunda columna se tienen los valores mínimos y máximos de cada variable. En la tercera columna se tiene a la media o al promedio, para el caso del IRI y de la PR la media tiende a aumentar de un año para otro, por ejemplo, en el 2019 la media del IRI es de 1.366, la cual aumenta en el año 2020 con un valor de 1.438, lo cual refleja un comportamiento normal de una carretera ya que se está deteriorando al pasar el tiempo. Después se tiene a la mediana que representa el valor central de los datos, en el caso de la PR hay un aumento en sus valores ya que el año 2019 la mediana es de 3.760 aumentando a 4.390 para el año 2020. La moda que indica el valor que más se repite para cada variable. Por último, se tiene a la desviación estándar, medida de dispersión que representa la homogeneidad o la dispersión en que se encuentra los datos con respecto a la media calculada. Es decir, a mayor variabilidad o dispersión, se tiene un resultado mayor de la desviación estándar, en



forma contraria, a menor variabilidad o dispersión, se tiene un resultado menor de la desviación estándar, por lo cual las variables IRI y PR presentan un resultado mayor de desviación estándar para los dos años lo que nos indica que los datos se encuentran muy dispersos.

MEDIDAS DE FORMA Y POSICIÓN						
Año		IRI (m/km)	PR (mm)	MAC (mm)	Agrietamiento (%)	
2019	Asimetría		2.35	1.92	-0.34	3.07
	Curtosis		9.88	5.09	-0.28	8.27
	Percentiles	25	1.03	2.85	0.94	0.00
		50	1.23	3.76	1.12	0.00
		75	1.54	4.76	1.29	0.01
2020	Asimetría		1.87	1.97	-0.12	3.14
	Curtosis		4.46	4.86	-0.96	8.70
	Percentiles	25	1.07	3.54	0.93	0.00
		50	1.27	4.39	1.11	0.00
		75	1.62	5.43	1.35	0.01

Tabla 14. Medidas de forma y posición.

Para caracterizar el perfil de una distribución de valores existen dos coeficientes, llamados medidas de forma, útiles para describir la forma de una distribución. El primer coeficiente como se observa en la tabla 14 es la asimetría. La asimetría del IRI, PR y Agrietamiento es mayor a cero y positiva por lo que indica una asimetría positiva y la cola es hacia la derecha, en el caso de la macrotextura su medida es negativa por lo que indica asimetría negativa y la una cola hacia la izquierda. Después se calculó el coeficiente de curtosis, la curtosis mide el grado de apuntamiento de una distribución con respecto a la distribución normal, para el caso de las variables IRI y PR se observa que los valores son altos positivos lo cual no indica una variabilidad de los datos, fenómeno visto en la tabla 14. El problema de

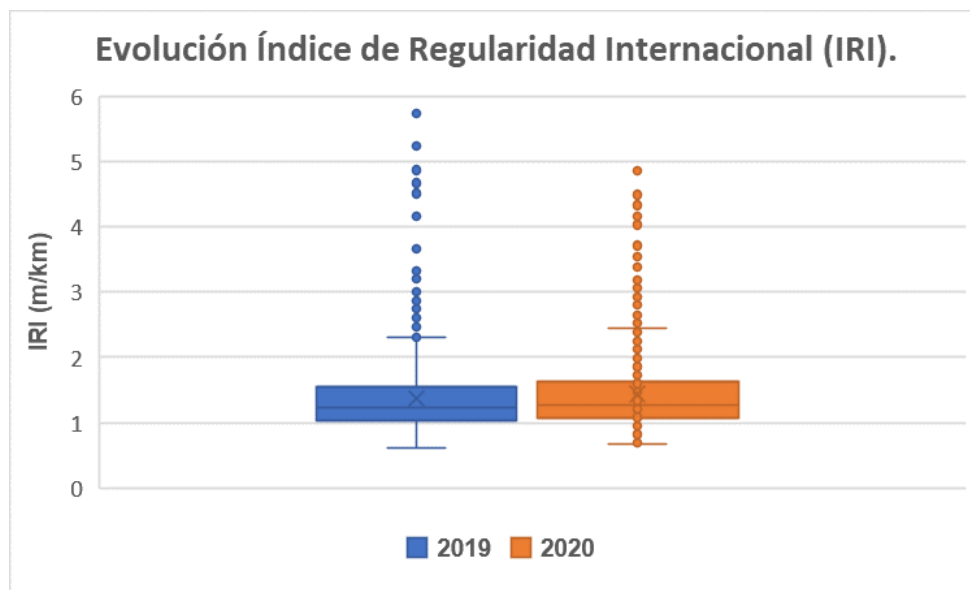


la variabilidad y dispersión de los datos puede ocasionar dificultades al momento del aprendizaje de los modelos de inteligencia artificial.

Por último, se tienen las medidas de posición o percentiles los cuales nos permiten identificar los valores ubicados en diferentes posiciones de un grupo de datos, los percentiles (primero, segundo y tercero) marcan el valor ubicado al 25%, 50% y 75% respectivamente de la totalidad de los datos.

La estadística descriptiva refleja una dispersión en el conjunto de los datos, por lo cual se realizó un análisis de la evolución de las variables.

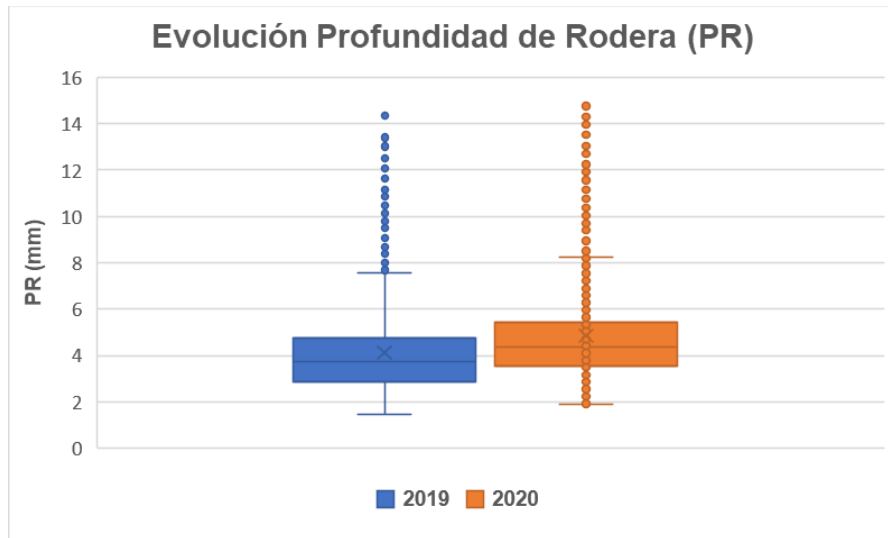
Primero se analizó la evolución del Índice de Regularidad Internacional (IRI), en la gráfica 3 se observa que para el año 2019 la media tiene un valor de 1.36 m/km, pero esta aumenta para el año 2020 teniendo un valor de 1.44 m/km.



Gráfica 3. Evolución del Índice de Regularidad Internacional de los años 2019 y 2020.

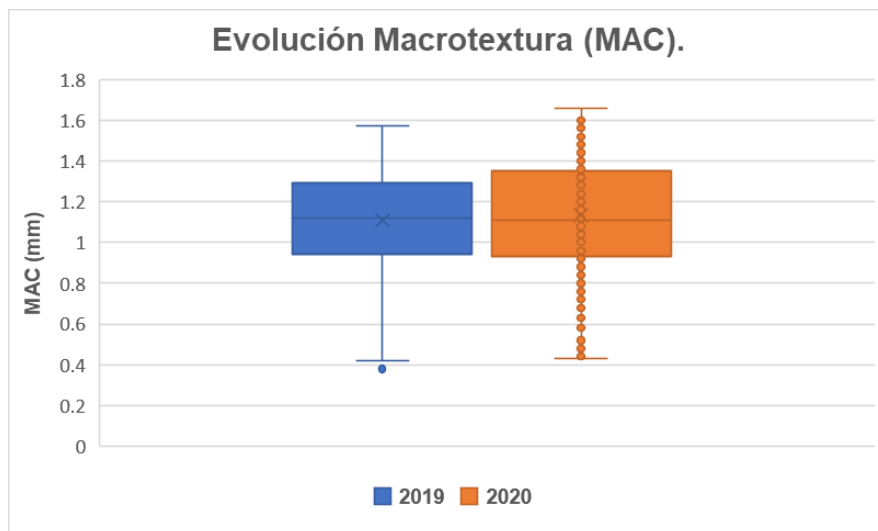


Después se analizó la evolución de la Profundidad de Rodera (PR), en la gráfica 4 se observa que para el año 2019 la media es de 4.12 mm y para el año 2020 la media es de 4.86, lo que indica que hubo un aumento en dicha variable.



Gráfica 4. Evolución de la Profundidad de Rodera de los años 2019 y 2020.

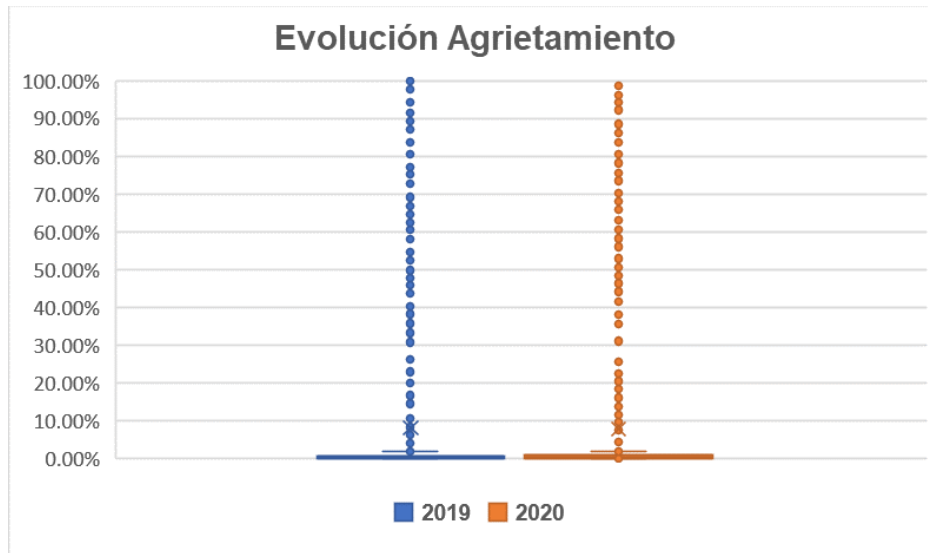
En la gráfica 5 se observa la evolución de la Macrotextura (mm), la media para el año 2019 es de 1.10 mm y para el año 2020 es de 1.13 mm.



Gráfica 5. Evolución de la macrotextura de los años 2019 y 2020.



Por último, se analizó la variable Agrietamiento (%), para el año 2019 la media es de 8.16%, pero para el año 2020 es de 7.87%, lo que indica una disminución en los valores.



Gráfica 6. Evolución del Agrietamiento de los años 2019 y 2020.

Con las gráficas 3 y 4 (IRI y PR), se observa que existe un aumento en los parámetros superficiales de la carretera lo que indica que con el paso del tiempo la carretera se fue deteriorando, esto debido a los diversos factores que afectan el comportamiento y la vida útil de la carretera.

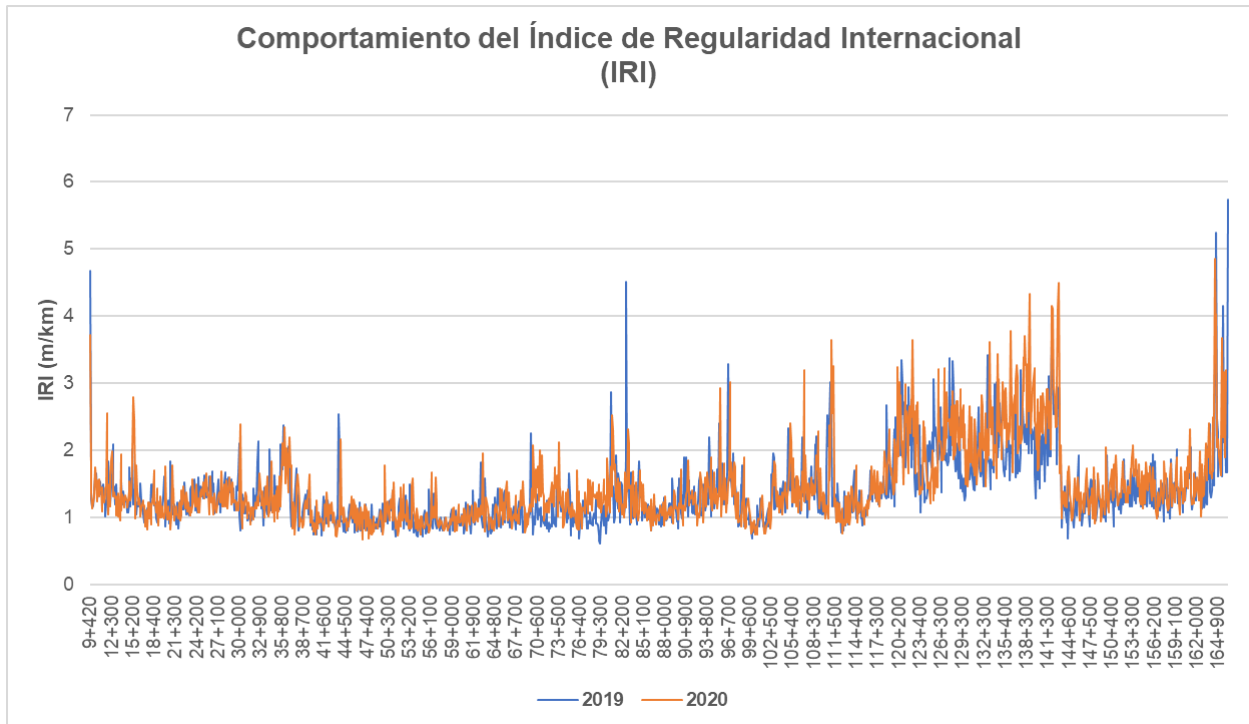
5.4. Análisis de Correlación.

La selección de las variables independientes, como ya se mencionó anteriormente es una tarea importante ya que se tienen que seleccionar aquellas variables que mejor representen el comportamiento del Índice de Regularidad Internacional (IRI), ya que el omitir variables puede llegar a ser perjudicial como favorable para el modelo de estimación, esto dependiendo de la correlación existente de las variables independientes con la variable dependiente.

En la gráfica 7 se observa el comportamiento del Índice de Regularidad Internacional (IRI) en el año 2019 como en el año 2020, el comportamiento del IRI como se observa en la gráfica es caótico, es decir no presenta ningún periodo



constante en su comportamiento, lo cual el encontrar las variables que mejor representen estos cambios es una tarea difícil.

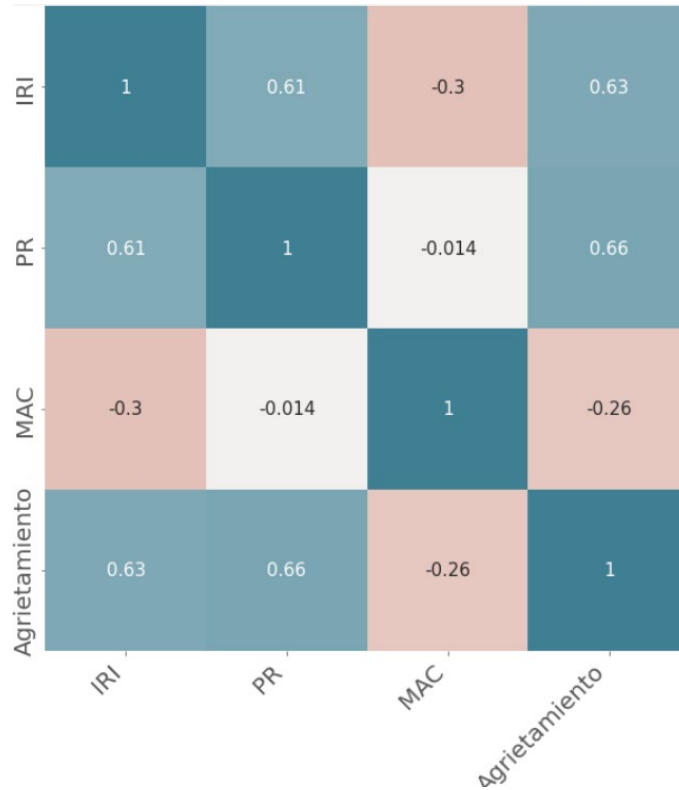


Gráfica 7. Comportamiento del Índice de Regularidad Internacional de los años 2019 y 2020

5.5.1. Coeficiente de Correlación de Pearson.

Los análisis de correlación se usan para examinar el grado de similitud de los valores de dos variables. El coeficiente de correlación de Pearson es un índice de fácil ejecución e, igualmente de fácil interpretación, el cual los valores están comprendidos entre -1 y 1.

El análisis de correlación sirvió para observar si las variables independientes permiten describir el comportamiento de la variable dependiente.



Gráfica 8. Análisis de Correlación de las variables seleccionadas.

En la gráfica 8, se observa la matriz con los respectivos coeficientes de correlación entre la variable dependiente (IRI) y las variables independientes (PR, MAC, Agrietamiento), se observa que el IRI con la PR presentan una correlación de 0.61 (correlación positiva moderada) es decir cuando el IRI aumenta la PR aumenta, el IRI con la MAC la correlación es de -0.3 (correlación negativa débil) cuando el IRI aumenta la MAC disminuye y por último el IRI con el Agrietamiento es de 0.63 (correlación positiva moderada). Por lo tanto, las variables PR y Agrietamiento tienen una correlación existente con el IRI, por lo cual van a ser las variables con mayor peso en el modelo de estimación, mientras que la variable MAC va afectar poco en el modelo.



5.6. Análisis de los distintos algoritmos para el modelo de estimación.

Como se mencionó en la propuesta de solución, la metodología que se utilizó consta de dos alternativas a elegir. La primera considera un enfoque con datos particionados o agrupamiento a través del algoritmo K-Means con la final de formar grupos que compartan características entre sí, pero a la vez que las características entre grupos sean diferentes. La segunda alternativa utiliza un enfoque sin datos particionados, es decir, utilizando la base de datos completa.

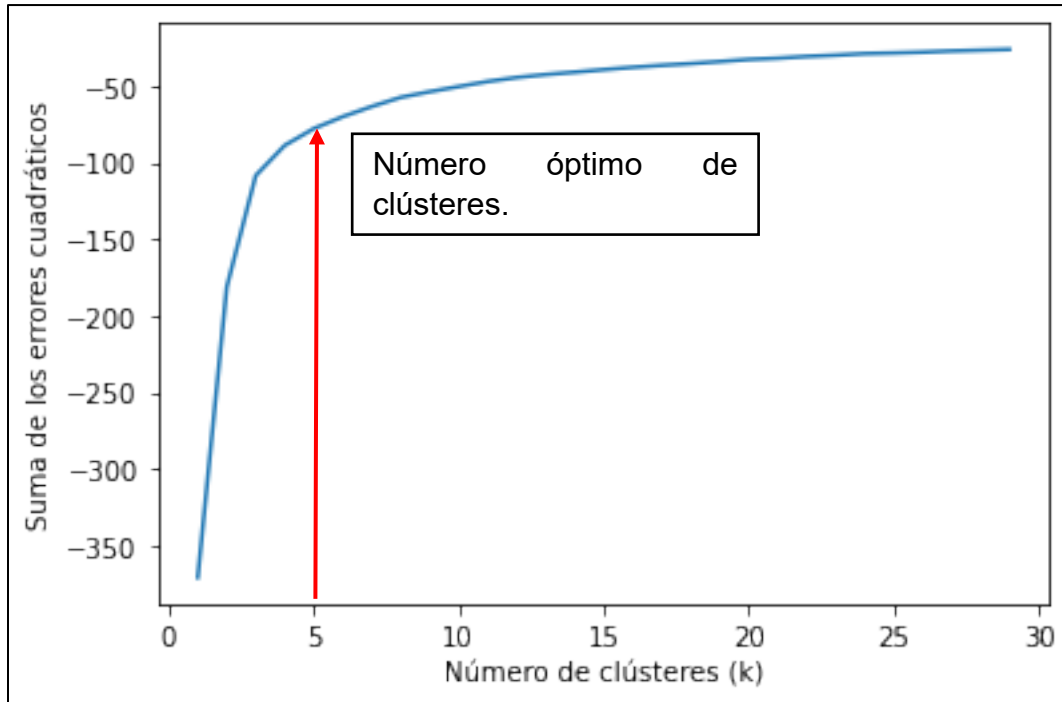
5.6.1. Enfoque con datos particionados.

Para enfoque con datos particionados o agrupamiento se utilizó el algoritmo de K-means, un algoritmo de aprendizaje no supervisado de clasificación que agrupa objetos en k grupos basándose en sus características. En el algoritmo es necesario seleccionar un valor k del número de grupos en los que se agrupan los datos, en general no hay forma exacta de determinar el número de grupos, pero se pueden usar ciertos métodos, como es el método del codo que nos ayuda a estimar el número de grupos para K-means para su posterior análisis de los diferentes algoritmos de inteligencia artificial.

5.6.1.1. Método del codo.

La idea básica del algoritmo de K-means es la minimización de la varianza intracluster y la maximización de la varianza inter-cluster, lo que se desea es que cada medición se encuentre muy cerca a las de su mismo grupo y los más lejos a los demás grupos.

El método del codo utiliza la distancia media de las observaciones a su centroide, es decir, se centra en las distancias intra-cluster, cuanto más grande es el número de clústeres k, la varianza intra-cluster tiende a disminuir, cuanto menor es la distancia intra-cluster mejor, ya que significa que los clústeres son más compactos.



Gráfica 9. Resultado del Método del Codo.

En la gráfica 9 se observa que la suma de las distancias disminuye conforme aumenta el número de clústeres y esa disminución se va haciendo más suave conforme aumentan el número de clústeres. El punto en donde se hace un codo y el cambio en el valor de la suma de las distancias se reduce significativamente es cuando $k=5$, y es el número de clústeres que deberá tener la muestra.

5.6.1.2. Método K-Means.

El clustering es un algoritmo de agrupamiento cuyo objetivo es realizar agrupaciones de datos de acuerdo a sus características. El método de K-means es un algoritmo de clustering que utiliza cada punto de la base de datos para asignarle un grupo o clúster de manera iterativa. El objetivo es asignar cada punto a un grupo con base en la similitud de sus características.

Una vez establecido el número $k=5$, que corresponde al número de clústeres, se asigna las coordenadas de los centroides que marcan el centro de cada agrupación, una vez establecidos los centroides se agrupan cada punto con el centroide más cercano.



Estadísticos descriptivos						
Cluster		N	Mínimo	Máximo	Media	Desv. Desviación
0	IRI	153	.127	1.000	.32173	.122576
	PR	153	.146	.953	.41399	.171596
	MAC	153	.164	.734	.42616	.108009
	Agrietamiento	153	.194	.810	.50729	.134541
1	IRI	1084	.000	.554	.13765	.057293
	PR	1084	.000	.308	.12006	.058120
	MAC	1084	.023	.633	.44758	.075532
	Agrietamiento	1084	.000	.349	.00441	.023631
2	IRI	1479	.012	.520	.10904	.060019
	PR	1479	.047	.663	.22259	.074830
	MAC	1479	.547	1.000	.74281	.081728
	Agrietamiento	1479	.000	.202	.00564	.009095
3	IRI	157	.150	.903	.36660	.120345
	PR	157	.025	.983	.56855	.192711
	MAC	157	.172	.844	.43190	.136122
	Agrietamiento	157	.616	1.000	.93919	.100670
4	IRI	228	.076	.830	.26909	.128732
	PR	228	.089	1.000	.36250	.162719
	MAC	228	.000	.664	.35334	.152925
	Agrietamiento	228	.000	.318	.04581	.073683

Tabla 15. Estadístico descriptivo de los grupos formados por medio del algoritmo K-Means.

En la tabla 15 se observa la estadística descriptiva de los 5 clústeres que ahora serán llamados clases, se observa que en la mayoría de las clases existe un solapamiento en las mediciones de las 4 variables. Este solapamiento entre clases provoca que la base de datos no sea linealmente separable lo que puede ocasionar problemas al momento de modelar con los algoritmos de inteligencia artificial.



5.6.1.3. *Parámetros de los algoritmos.*

Antes de comenzar con los modelos de estimación, primero se establecieron los parámetros para cada algoritmo, mismos que se muestran en la tabla 16 y 17.

Regresión Lineal	
attributeSelectionMethod	M5 method
batchSize	100
debug	False
doNotCheckCapabilities	False
eliminateColinearAttributes	True
minimal	False
numDecimalPlaces	4
outputAdditionalStats	False
ridge	1.08E-08
useQRDecomposition	False
Redes Neuronales Artificiales	
GUI	True
autoBuild	True
batchSize	100
debug	False
decay	False
doNotCheckCapabilities	False
hiddenLayers	4
learningRate	0.2
momentum	0.2
nominalToBinaryFilter	True
normalizeAttributes	False
normalizeNumericClass	False
numDecimalPlaces	2
reset	False
resume	False
seed	0
trainingTime	1000
validationSetSize	0
validationThreshold	20

Tabla 16. Parámetros de cada algoritmo.



Support Vector Machine	
batchSize	100
c	1
debug	False
doNotCheckCapabilities	False
filterType	False
kernel	Puk
numDecimalPlaces	2
regOptimizer	RegSMOImproved
Random Forest	
bagSizePercent	60
batchSize	100
breakTiesRandomly	False
calcOutOfBag	False
computeAttributeImportance	False
debug	False
doNotCheckCapabilities	False
maxDepth	0
numDecimalPlaces	2
numExecutionSlots	1
numFeatures	0
numIterations	200
outputOutOfBagComplexityStatistics	False
printClassifiers	False
seed	1
storeOutOfBagPredictions	False
M5 Rules	
batchSize	100
buildRegressionTree	False
debug	False
doNotCheckCapabilities	False
minNumInstances	4
numDecimalPlaces	4
unpruned	False
useUnsmoothed	False

Tabla 17. Continuación tabla 17. Parámetros de cada algoritmo.



En la tabla 16 y 17 se observan los parámetros que se utilizaron al momento de realizar los modelos de estimación con los distintos algoritmos. Los algoritmos que se utilizaron fueron regresión lineal, redes neuronales artificiales, support vector machine, random forest y M5 rules. Se observa que cada algoritmo cuenta con sus propios parámetros para una configuración adecuada para el tipo de datos de la investigación. Existen dos parámetros presentes en los 5 algoritmos, el primero de ellos es la normalización que se le dio a la base de datos, es por ello que el parámetro debug (depuración) no es necesario aplicarla ya que se cuenta con una normalización previa. El segundo parámetro la validación que se le dio a los modelos, en todos los casos se realizó una validación cruzada por medio del proceso K-Folds.

La técnica K-Folds permite que todas los datos o mediciones de la base de datos tengan la oportunidad de aparecer en la serie de entrenamiento y en la serie de prueba. Primero se empieza separando aproximadamente en grupos del mismo tamaño la serie de datos de manera aleatoria en K folds, el procedimiento tiene un único parámetro llamado "K" que hace referencia al número de grupos en que se dividirá la muestra, K-1 grupos se emplean para entrenar el modelo y uno de los grupos restantes se emplea como validación. Este proceso se repite K veces utilizando un grupo distinto como validación en cada iteración. El proceso genera K estimaciones del error cuyo promedio se emplea como estimación final.

En esta investigación se utilizó un valor de 10 para el parámetro K. Por lo cual teniendo establecidos los parámetros se realizaron los modelos por medio de los diferentes algoritmos.



5.6.1.4. Resultados de los modelos para los tres grupos.

Como se mencionó anteriormente el número de clústeres de acuerdo con el método del codo fue de $k=5$, por lo cual el análisis comenzó con 5 clases. En las tablas 18 y 19 se observan los resultados de coeficiente de correlación y del error cuadrático medio de cada clase del grupo 1 ($k=5$) para los diferentes algoritmos de inteligencia artificial. Los resultados no son favorables ya que se observa en el caso del algoritmo de Random Forest se tienen dos grupos con un coeficiente de correlación baja y solamente dos grupos cuentan con una correlación superior a 0.4.

Motivo por el cual se realizaron dos nuevos análisis, es decir, aumentando el número k , y otro disminuyendo el número k , ahora se tienen dos nuevos grupos uno con $k=6$ y $k=4$. En las tablas 18 y 19 se observan los resultados del coeficiente de correlación, así como del error cuadrático medio de los tres grupos.



		Regresión Lineal		Redes Neuronales		Support Vector Machine	
		Correlación	RMSE	Correlación	RMSE	Correlación	RMSE
1 Grupo (k=4)	Clase 1	0.1681	0.0564	0.0605	0.0601	0.1673	0.0567
	Clase 2	0.231	0.0632	0.1828	0.0698	0.2166	0.0643
	Clase 3	0.4051	0.1222	0.2369	0.155	0.4102	0.1223
	Clase 4	0.7833	0.0666	0.7132	0.0803	0.7781	0.0681
2 Grupo (k=5)	Clase 1	0.5647	0.105	0.4875	0.113	0.5712	0.102
	Clase 2	0.1461	0.1597	0.0023	0.1648	0.1421	0.157
	Clase 3	0.2018	0.1588	0.0966	0.1626	0.1936	0.1599
	Clase 4	0.5341	0.1035	0.3877	0.1194	0.5347	0.1025
	Clase 5	-0.1884	0.1306	0.2624	0.1294	-0.0524	0.1318
3 Grupo (k=6)	Clase 1	-0.1716	0.1302	0.1466	0.1419	-0.0252	0.1319
	Clase 2	0.5203	0.1028	0.3706	0.1206	0.542	0.1024
	Clase 3	0.231	0.0632	0.1828	0.0698	0.2166	0.0643
	Clase 4	0.5728	0.1006	0.4556	0.1139	0.5943	0.0993
	Clase 5	0.235	0.0489	0.1543	0.0537	0.2425	0.0498
	Clase 6	0.1681	0.0564	0.0605	0.0601	0.1673	0.0567

Tabla 18. Resultados de las métricas de desempeño de los algoritmos Regresión Lineal, Redes Neuronales y Support Vector Machine.



		Random Forest		M5Rules	
		Correlación	RMSE	Correlación	RMSE
1 Grupo (k=4)	Clase 1	0.2176	0.0577	0.1685	0.0569
	Clase 2	0.2947	0.0632	0.1205	0.1043
	Clase 3	0.5825	0.1093	0.4469	0.1248
	Clase 4	0.8112	0.0629	0.697	0.0817
2 Grupo (k=5)	Clase 1	0.5332	0.1056	0.215	0.1243
	Clase 2	0.1605	0.156	0.0901	0.1601
	Clase 3	0.2841	0.1548	0.1892	0.1623
	Clase 4	0.4509	0.1111	0.3827	0.1232
	Clase 5	0.4117	0.1194	0.091	0.1285
3 Grupo (k=6)	Clase 1	0.4939	0.1124	0.0261	0.1387
	Clase 2	0.4689	0.1092	0.401	0.1252
	Clase 3	0.2947	0.0632	0.1205	0.1043
	Clase 4	0.5134	0.1079	0.1211	0.1289
	Clase 5	0.2541	0.0504	0.1166	0.056
	Clase 6	0.2176	0.0577	0.1685	0.0569

Tabla 19. Resultados de las métricas de desempeño de los algoritmos Random Forest y M5 Rules.

Como se observa en las tablas 18 y 19, el grupo 1 (k=4), el algoritmo que mejor resultados obtuvo fue Random Forest, donde dos de sus clases obtuvieron un resultado favorable. En el grupo 3 (k=6) de igual manera el algoritmo el que obtuvo mejores resultados fue Random Forest donde tres de sus clases alcanzaron resultados favorables.

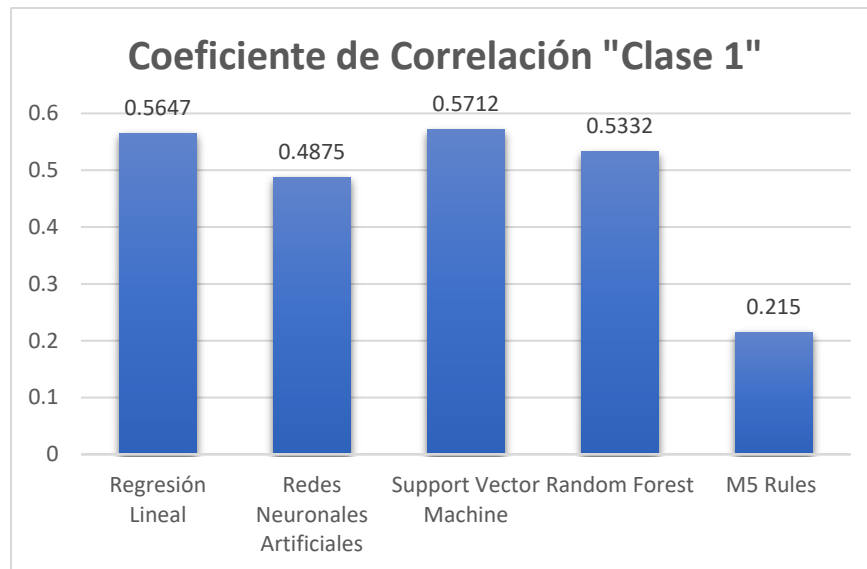
De los tres grupos analizados, el grupo 2 (k=5) fue donde la mayoría de las clases adquirieron resultados favorables ya que en los grupos 1 y 3 la mitad de sus clases se desempeñaron favorablemente y la mitad no. Motivo por el cual se eligió al grupo

2 como modelo de estimación y este se comparará con el modelo utilizando la base de datos completa.

En los capítulos posteriores se realizó un análisis de los resultados del coeficiente de correlación, así como del error cuadrático medio del grupo 2.

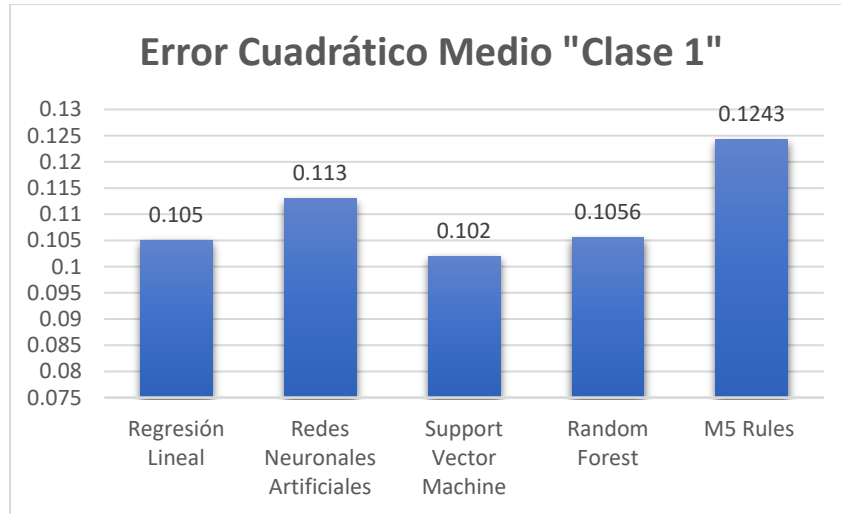
5.6.1.5. Resultados de los modelos de la clase 1.

En la gráfica 10 se observa los resultados del Coeficiente de Correlación de los cinco algoritmos, como ya se mencionó en los capítulos anteriores, el coeficiente de correlación mientras más cercano sea a 1 o -1 nos indica una correlación fuerte de las variables, por lo cual es un resultado que se está buscando. El algoritmo de Support Vector Machine obtuvo el coeficiente más alto.



Gráfica 10. Coeficiente de Correlación de la Clase 1.

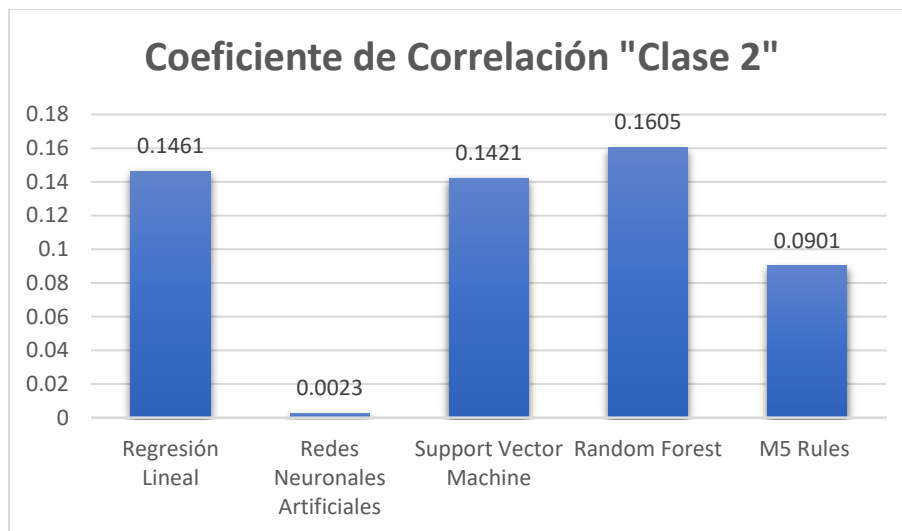
En la gráfica 11 se observa el Error Cuadrático Medio (RMSE), como se mencionó en capítulos anteriores el RMSE es una medida que indica la precisión del modelo a la hora de estimar el Índice de Regularidad Internacional (IRI), el cual no indica que valores cercanos a 0 el modelo es favorable, de forma contraria mientras mayor sea el error nos indica que el modelo no se desempeña de la mejor manera. La gráfica 11 indica que el algoritmo Support Vector Machine es el que obtuvo el resultado más bajo del error cuadrático medio.



Gráfica 11. Error cuadrático medio de la Clase 1.

5.6.1.6. Resultados de los modelos de la clase 2.

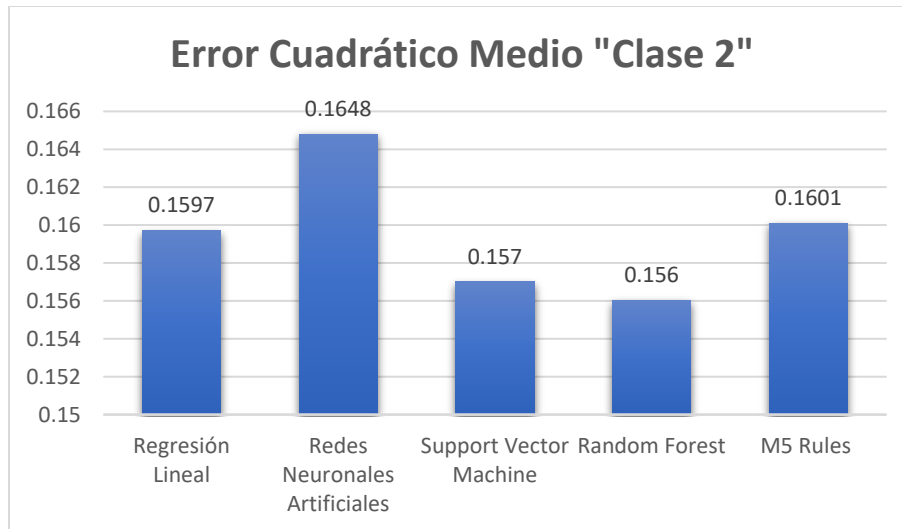
En la gráfica 12 se observa que el algoritmo Random Forest es el que obtuvo mejor resultado de coeficiente de correlación del grupo 2. Mientras que el algoritmo de Redes Neuronales obtuvo un coeficiente muy cercano a cero.



Gráfica 12. Coeficiente de correlación de la Clase 2.



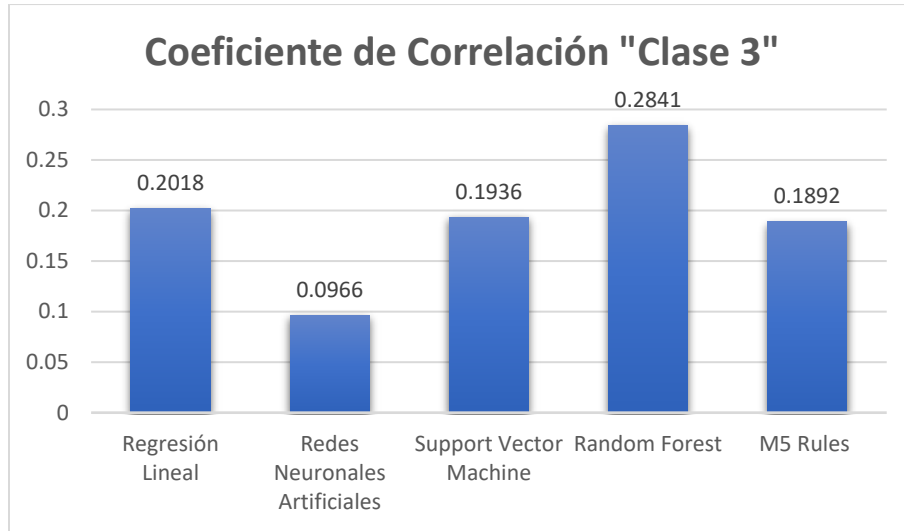
En la gráfica 13 se observa el valor que el Error Cuadrático Medio del algoritmo Random Forest es el más bajo por lo que es el algoritmo que mejor se desempeñó de los cinco para la clase 2.



Gráfica 13. Error cuadrático medio de la Clase 2.

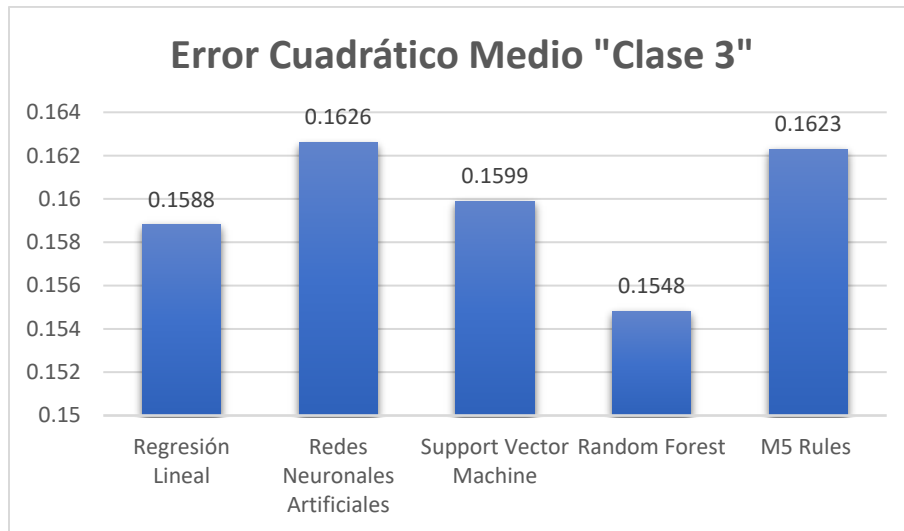
5.6.1.7. Resultados de los modelos de la clase 3.

En la gráfica 14 se observa que el algoritmo Random Forest es el que obtuvo un resultado cercano a 1, por lo cual es el algoritmo que mejor se desempeña para la clase 3. De igual manera que en la clase 2, el algoritmo de Redes Neuronales fue el que obtuvo una correlación baja.



Gráfica 14. Coeficiente de Correlación de la clase 3.

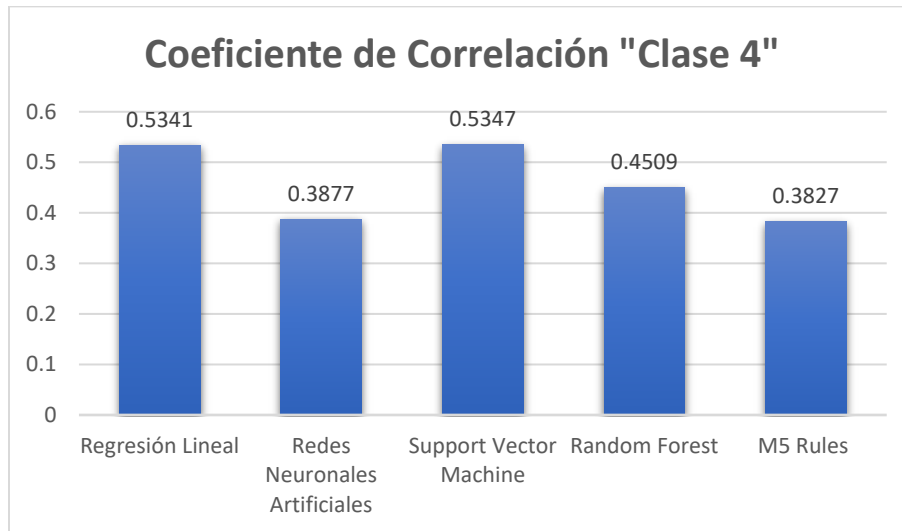
En la gráfica 15 se observa el Error Cuadrático Medio de los algoritmos de la clase 3, el cual Random Forest es que tiene el resultado más bajo, por lo cual es el que mejor se desempeña de los cinco.



Gráfica 15. Error cuadrático medio de la clase 3.

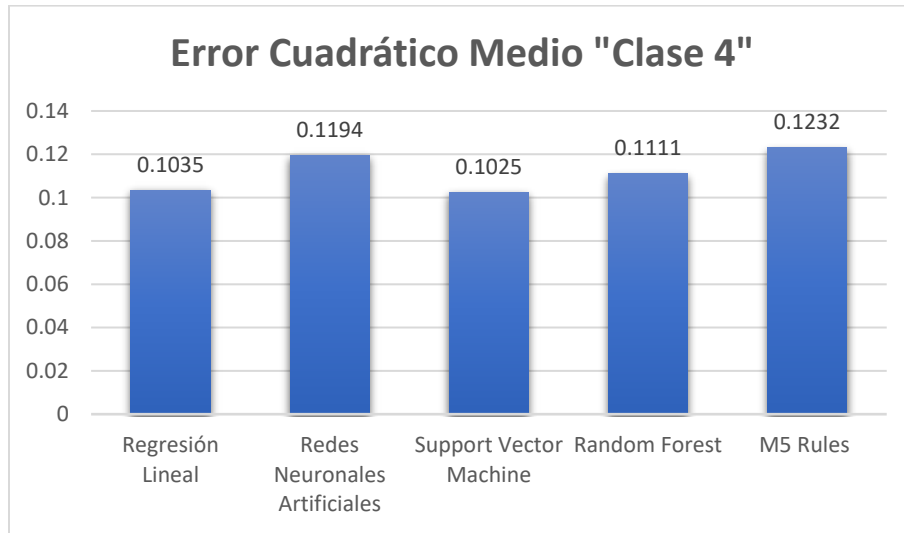
5.6.1.8. Resultados de los modelos de la clase 4.

En la gráfica 16 se observa los coeficientes de correlación para los cinco algoritmos de la clase 4, en este caso el algoritmo Support Vector Machine es el obtuvo el resultado más alto siendo el que mejor se desempeña, de forma contraria el algoritmo Redes Neuronales Artificiales obtuvo el peor resultado.



Gráfica 16. Coeficiente de Correlación de la clase 4.

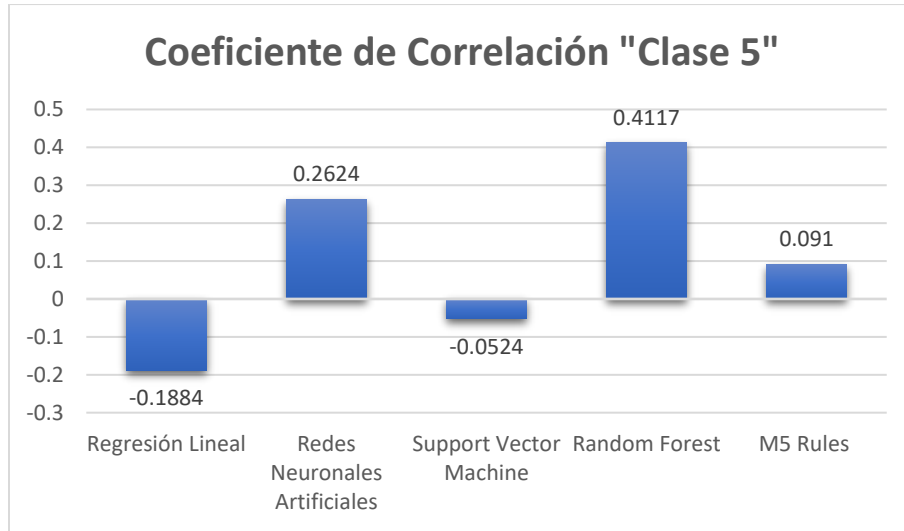
En la gráfica 17 se observa los errores cuadrático medio de los cinco algoritmos, el cual Support Vector Machine es que tiene el error más bajo, siendo el que mejor se desempeña de los cinco.



Gráfica 17. Error cuadrático medio de la clase 4.

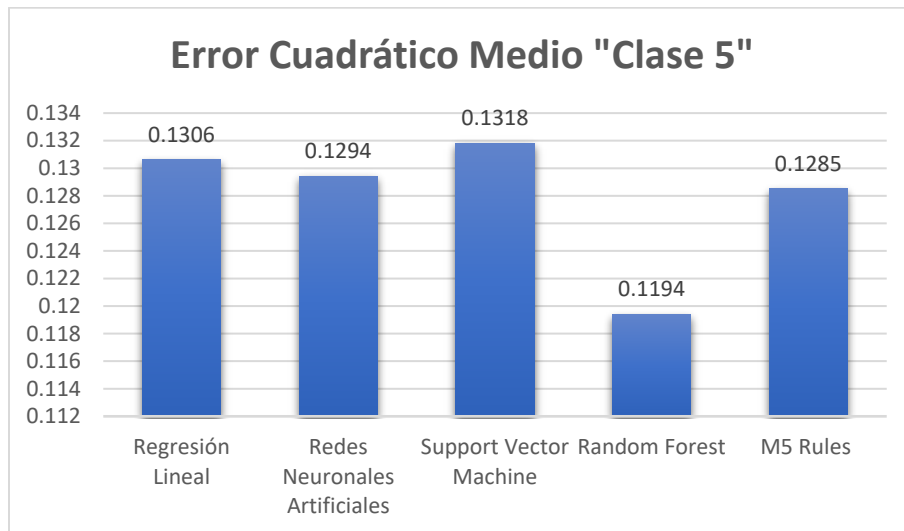
5.6.1.9. Resultados de los modelos de la clase 5.

En la gráfica 18 se observa los resultados de los Coeficientes de Correlación para la clase 5, el cual Random Forest es que él tiene el coeficiente más alto de los cinco, mientras Support Vector Machine es el que peor se desempeñó de los cinco algoritmos.



Gráfica 18. Coeficiente de Correlación del Clase 5.

En la gráfica 19 se observa que el algoritmo Random Forest es el que obtuvo el error más bajo de los cinco por lo cual es que mejor se desempeña para la clase 5.



Gráfica 19. Error cuadrático medio de la clase 5.



Como se vio en las gráficas anteriores los algoritmos que mejor se desempeñaron fueron Support Vector Machine y Random Forest, ya que con dichos algoritmos se logró alcanzar un coeficiente de correlación de 0.57 en el caso del algoritmo de Support Vector Machine, así como un error cuadrático bajo de 0.0548 con Random Forest.

5.6.2. Enfoque sin datos particionados.

Como se mencionó en la metodología, el segundo enfoque consiste en considerar la base de datos completa sin particiones.

El análisis se realizó con los mismos algoritmos de inteligencia artificial que se utilizaron en el enfoque anterior. Al finalizar el análisis con todos los algoritmos se realizó una comparación para elegir el modelo que mejores resultados en sus métricas de desempeño obtuvo.

5.6.2.1. Regresión Lineal.

Como primer algoritmo a utilizar dentro del enfoque sin datos particionados, se encuentra la regresión lineal. La regresión lineal es un algoritmo, como ya se mencionó en los capítulos anteriores trata de buscar la línea de la recta de mejor ajuste, estableciendo una relación entre la variable dependiente y las variables independientes.

En este caso, la variable dependiente es el Índice de Regularidad Internacional (IRI), y las variables independientes son, la Profundidad de Rodera, Macrotextura y los Agrietamientos.

Una vez establecido las variables dependientes como las independientes, se prosigue a modelar y encontrar la ecuación de la reta. La ecuación de la recta dada por la base de datos y por la selección de la variable dependiente como las independientes es la que muestra a continuación.

$$IRI = 0.269 * PR - 0.1506 * MAC + 0.1305 * AGRI + 0.1709$$



La ecuación anterior es la que mejor ajuste tiene de acuerdo con las variables seleccionadas, dicho modelo arroja una precisión de 0.0778 de acuerdo con el Error Cuadrático Medio (RMSE siglas en inglés) y un coeficiente de correlación de 0.6801.

5.6.2.2. Redes Neuronales Artificiales.

El siguiente algoritmo que se utilizó fue el de Redes Neuronales Artificiales. Una Red Neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la información, funciona simulando un número elevado de unidades de procesamiento interconectados que parecen versiones abstractas de neuronas. Las unidades de procesamiento se organizan en capas. Hay tres partes normalmente en una red neuronal, una capa de entrada, con unidades que representan los campos de entrada, una o varias capas ocultas y una capa de salida, con una unidad o unidades que representan el campo o los campos de destino. La red aprende examinando los registros individuales, generando una estimación para cada registro y realizando ajustes a las ponderaciones cuando realiza una estimación incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta haber alcanzado uno o varios criterios de parada.

La arquitectura de la red que mejores resultados obtuvo fue la que se muestra en la figura 14:

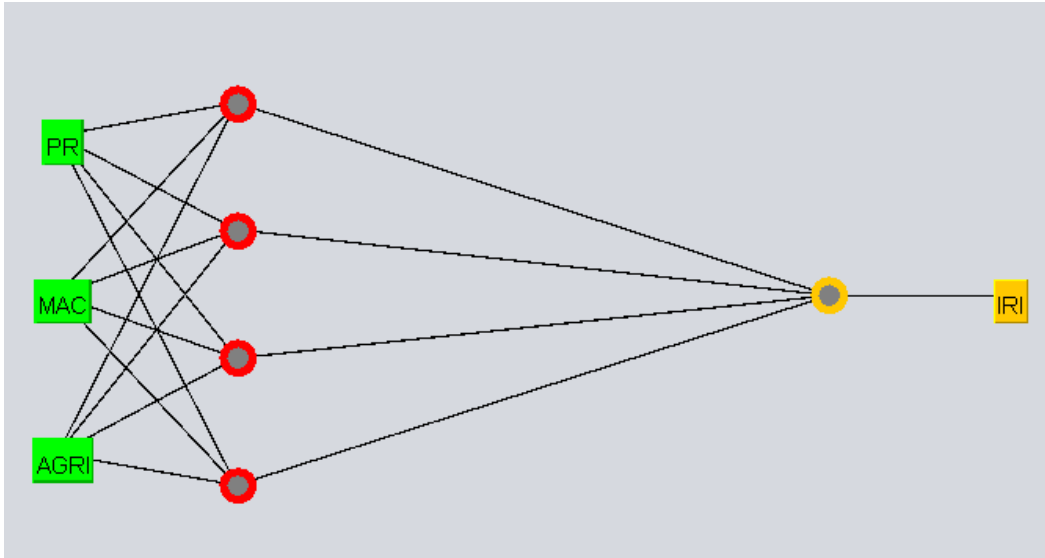


Figura 15. Modelo de Redes Neuronales Artificiales.

Como se observa en la figura, la arquitectura de la red es de 3-4-1, los datos de entrada se presentan en la primera capa (profundidad de rodera, macrotextura y agrietamientos), y los valores se propagan desde cada neurona hasta cada neurona de la capa siguiente. Al final, se envía un resultado desde la capa de salida (IRI).

La validación del modelo se llevó a cabo con el método K-folds. El modelo de Redes Neuronales Artificiales estima el futuro del Índice de Regularidad Internacional con una precisión de 0 .0784, esto acuerdo con el Error Cuadrático Medio (RMSE siglas en ingles).

5.6.2.3. Support Vector Machine.

El siguiente algoritmo que se utilizó fue el de Maquinas de Vectores de Soporte para regresión, el objetivo del algoritmo es encontrar la curva o hiperplano que modele la tendencia de los datos de entrenamiento para estimar cualquier dato en el futuro. El hiperplano es la manera en que se separa a los datos, el hiperplano es margen mas amplio entre las mediciones, el margen se define como la anchura máxima de la región paralela al hiperplano que no tiene puntos de datos interiores.

Los vectores de soporte hacen referencia a un subconjunto de observaciones de entrenamiento que identifican la ubicación del hiperplano de separación, en esta



investigación se hizo uso del método kernel. La función kernel asignan los datos a un espacio dimensional diferente, que suele ser superior, con la expectativa de que resulte más fácil separar las clases después de esta transformación, simplificando potencialmente los límites de decisión complejos no lineales para hacerlos lineales en el espacio dimensional.

Como se mencionó anteriormente el algoritmo de Maquinas de Vectores de Soporte para regresión se basa en estimar valores numéricos, dado que la salida es un número real, se vuelve muy difícil estimar la información disponible, ya que se tiene infinidad de posibilidades, sin embargo, el objetivo es el minimizar el error aumentando la precisión del modelo. En este caso el coeficiente de correlación que obtuvo el modelo fue de 0.7291, con una precisión del modelo de acuerdo con el Error Cuadrático Medio (RMSE siglas en inglés) fue de 0.0736, ligeramente menor en comparación con el modelo de Redes Neuronales Artificiales.

5.6.2.4. Random Forest.

Random Forest fue el cuarto algoritmo que se utilizó para realizar el análisis. Como se mencionó anteriormente, Random Forest está basando en los modelos de árboles de decisión, es uno de los algoritmos que se usa para tareas de clasificación y regresión, combinado con su naturaleza no lineal, lo hace adaptable para el análisis de esta investigación.

Random Forest es un algoritmo de aprendizaje supervisado por lo que crea un bosque y lo hace de manera aleatoria. El algoritmo crea múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable. Mientras que un árbol de decisiones en solitario tiene un resultado y un rango reducido de grupos, el bosque asegura un resultado más preciso con una mayor cantidad de grupos y decisiones. Es este algoritmo se agrega aleatoriedad adicional al modelo, mientras crece los árboles, en lugar de buscar la característica más importante al dividir un nodo (árboles de decisión), busca la mejor característica entre un subconjunto aleatorio de características. En general entre más árboles en el bosque,



más robusto es el bosque, normalmente cuantos más árboles mejor, pero a partir de cierto punto deja de mejorar y solo hace que el modelo vaya más lento.

El modelo arrojó resultados favorables con número de 100 árboles en el bosque, la validación del modelo se llevó a cabo con un número de k-folds de 10, dando como resultado un coeficiente de correlación de 0.7389, con una precisión de acuerdo con el Error Cuadrático Medio (RMSE siglas en inglés) de 0.0716. El modelo basado en Random Forest dio resultados más favorables en el coeficiente de correlación y el Error Cuadrático Medio que la Regresión Lineal, Redes Neuronales Artificiales y Support Vector Machine.

5.6.2.5. M5 Rules

Por último, se probó el algoritmo M5 Rules, este método genera reglas a partir de árboles modelos, se aplica un árbol de aprendizaje al conjunto de datos de entrenamiento completo. A continuación, la mejor hoja se convierte en una regla y el árbol se descarta. Todas las instancias cubiertas por la regla se eliminan del conjunto de datos. El proceso se aplica recursivamente a las instancias restantes y finaliza cuando todas las instancias están cubiertas por una o más reglas. Esta es la estrategia básica de separar y conquistar para aprender reglas, sin embargo, en lugar de construir una sola regla, como se hace normalmente, construimos un árbol modelo completo en cada etapa y se convierte su mejor hoja en una regla. M5 Rules construye árboles completos en lugar de árboles parcialmente explorados. La construcción de árboles parciales conduce a una mayor eficiencia y no afecta el tamaño y la precisión de las reglas resultantes (Holmes, G 1999).

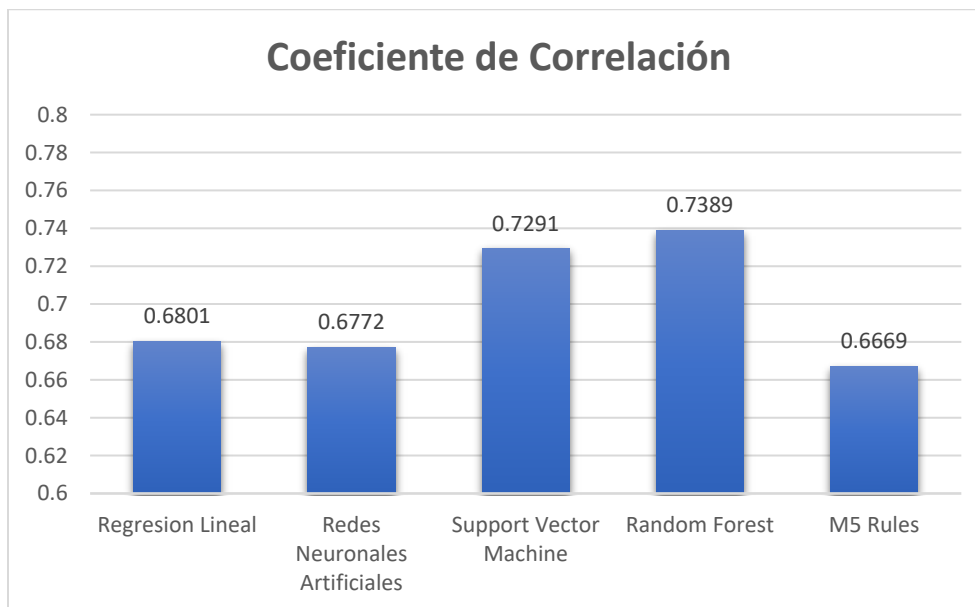
El modelo se probó con 4 instancias que se permiten en un nodo hoja, generó una lista de reglas de decisión en lugar de un árbol, el modelo fue válido con un valor de 10 K-Folds. El resultado del coeficiente de correlación fue de 0.6669 y la precisión del modelo de acuerdo con el Error Cuadrático Medio (RMSE siglas en inglés) de 0.0795. El modelo no mejoró en comparación con el algoritmo de Random Forest, ya que su coeficiente de correlación fue menor y su Error Cuadrático Medio fue mayor.



5.6.2.6. Resultados de los diferentes modelos de estimación.

Las gráficas 20 y 21 representan los resultados de los diversos modelos de cada algoritmo para el enfoque sin datos particionados, las gráficas ayudarán a una mejor visualización del comportamiento de los modelos con respecto a sus resultados, a la vez permitirán la correcta selección del modelo que mejores resultados obtuvo de los 5 modelos analizados anteriormente.

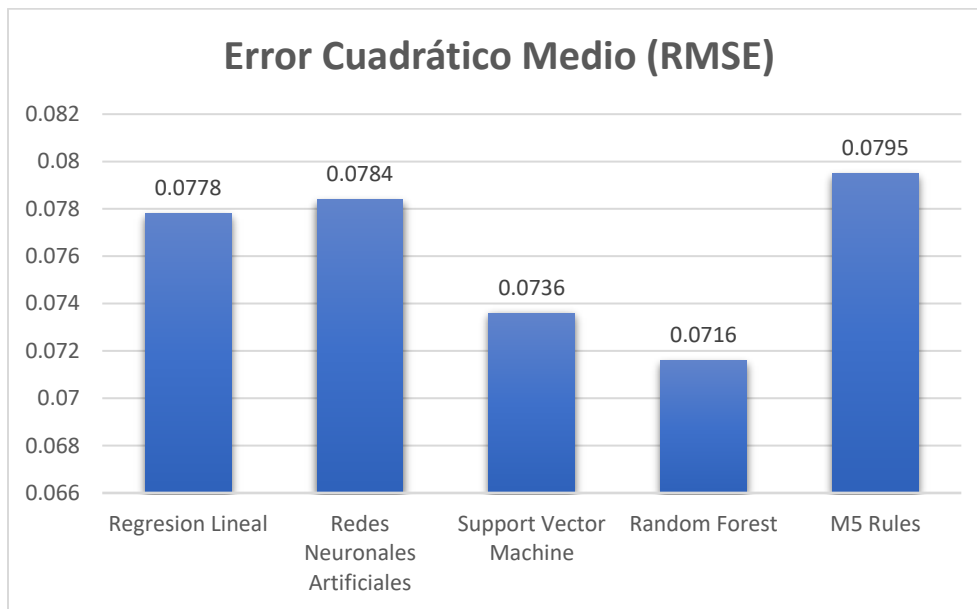
La grafica 21 representa los resultados del coeficiente de correlación para el enfoque sin datos particionados, como se mencionó anteriormente, el modelo que obtenga el resultado más cercano a la unidad representa una alta correlación del modelo, por lo cual es lo que se está buscando. Se observa que el algoritmo Random Forest es el modelo con el resultado más cercano a la unidad con un valor de 0.7389, mientras M5 Rules, es el algoritmo que peor se desempeñó de los cinco algoritmos analizados.



Gráfica 20. Coeficiente de correlación de los modelos para el enfoque sin datos particionados.



En la gráfica 21 se observa los resultados del Error Cuadrático Medio (RMSE) de los modelos para el enfoque sin datos particionados. Como se mencionó anteriormente, mientras más cercano a 0 sea el error, el modelo tendrá más precisión. Se observa que el algoritmo Random Forest es el modelo con el resultado más cercano a 0, por lo cual es el modelo que tiene más precisión con un error de 0.0716. De igual manera se aprecia que el algoritmo M5 Rules es el modelo que tiene la precisión más baja de los cinco algoritmos estudiados.



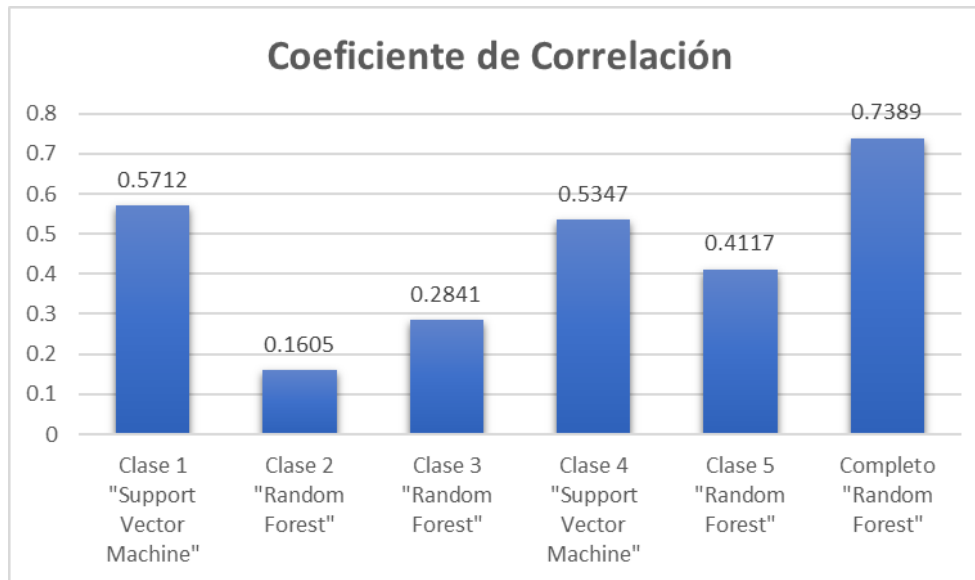
Gráfica 21. Error Cuadrático Medio de los modelos para el enfoque sin datos particionados.

Una vez finalizada el análisis de los modelos con el enfoque datos particionados y el enfoque datos sin particionar, es momento de elegir el enfoque y el modelo que mejores resultados obtuvo con respecto al coeficiente de correlación y al error cuadrático medio.



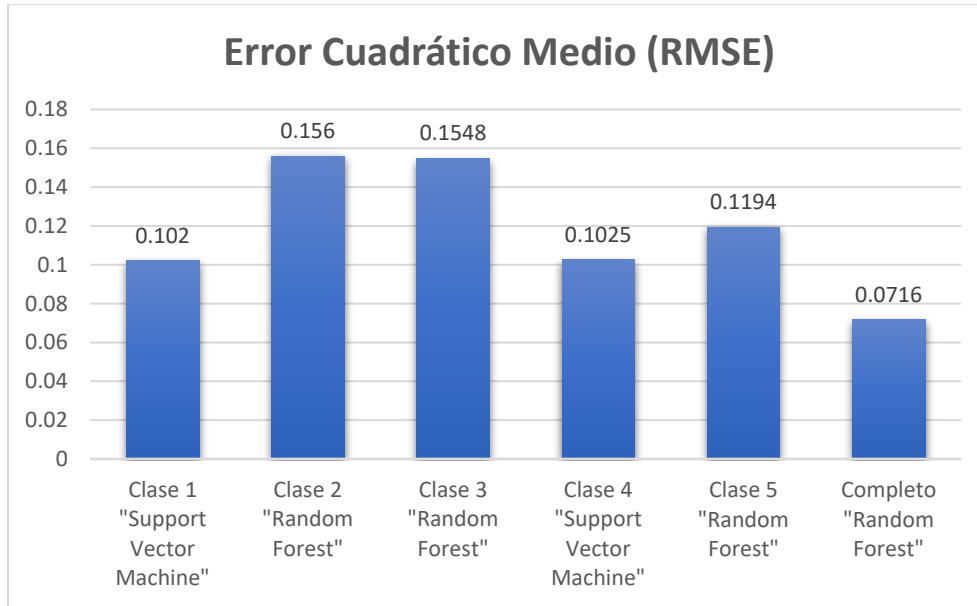
5.6.3. Selección del modelo de estimación.

Finalizados los análisis de ambos enfoques, se realizó una comparación entre los coeficientes de correlación y el error cuadrático medio del mejor grupo del enfoque con datos particionados y el mejor modelo del enfoque sin datos particionados, el cual se observa en la figura 22 y 23.



Gráfica 22. Coeficiente de correlación de ambos enfoques.

En la gráfica 22 se muestran los resultados de las 5 clases que se agruparon por medio del algoritmo K-Means, las cuales fueron las que obtuvieron un mejor resultado en comparación con los otros grupos. Al igual se observa el resultado que se obtuvo al utilizar el enfoque sin datos particionados. De los dos enfoques se aprecia que al utilizar la base de datos completa se obtiene una mejor correlación entre las variables, al obtener un resultado de 0.7389, resultado que es superior a las distintas clases.



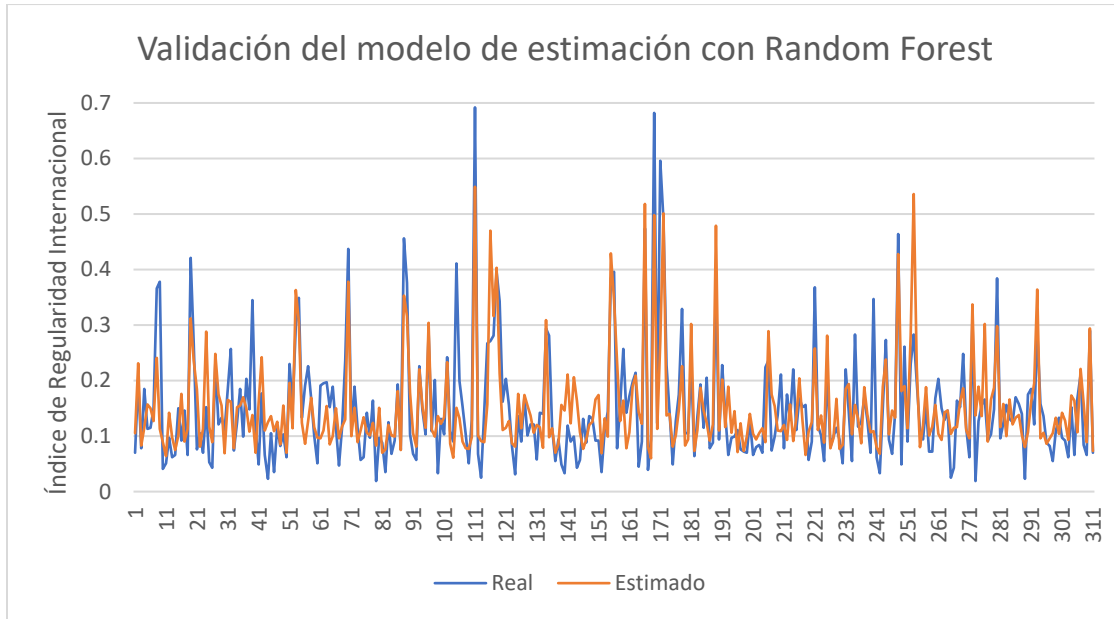
Gráfica 23. Error cuadrático medio de ambos enfoques.

En la gráfica 23 se observan los resultados del error cuadrático medio de los dos enfoques analizados, se aprecia que al utilizar el enfoque sin datos particionados se tiene un mejor resultado ya que el valor de 0.0716 es inferior a los resultados del enfoque con datos particionados.

En base a las graficas 22 y 23, el modelo de estimación del Índice de Regularidad Internacional es a base de utilizar la base de datos completa, además de utilizar el algoritmo Random Forest.

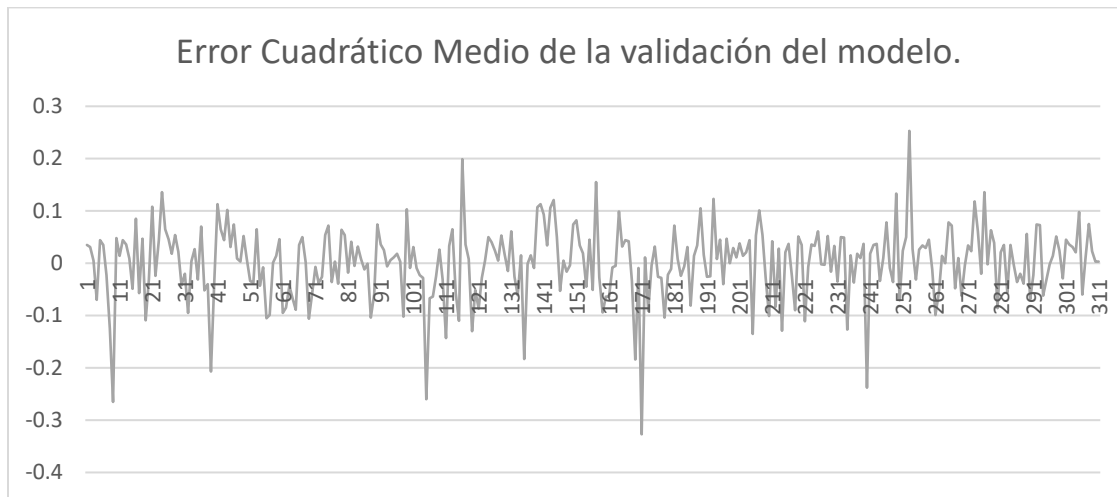
A continuación, se muestran las representaciones graficas del modelo de estimación.

En la gráfica 24 se observa un ejemplo de la estimación de valores que proporciona el modelo siendo la línea azul los valores reales normalizados del Índice de Regularidad Internacional, y la línea naranja como los valores estimados normalizados. Además de que la representación gráfica sirve como validación del modelo.



Gráfica 24. Validación del modelo de estimación con Random Forest.

A continuación, se muestra el error cuadrático medio de los valores de salida o de las estimaciones de la validación del modelo con Random Forest.



Gráfica 25. Error Cuadrático Medio de la validación del modelo con Random Forest.



6. Conclusiones.

El conocer el comportamiento de los pavimentos flexibles con respecto al Índice de Regularidad Internacional (IRI) es una tarea importante, ya que, de la información disponible es necesario encontrar las variables que describan el comportamiento y evolución del IRI. En esta investigación se observó que las variables Profundidad de Rodera, Macrotexturas y Agrietamientos, describen el comportamiento del IRI.

El preprocesamiento de los datos es una tarea que se debe de realizar antes del aprendizaje de los modelos, consiste en la limpieza de los datos, la detección y el tratamiento del ruido en los datos. El preprocesamiento de los datos tiene la finalidad de obtener una nueva base de datos limpia y adecuada para el proceso de aprendizaje automático a través de los algoritmos de inteligencia artificial.

El comportamiento del IRI en los pavimentos flexibles es caótico, es decir, no presenta patrones y no es constante en su comportamiento, esto se vio en el análisis estadístico donde se observó que los datos se encuentran muy dispersos. Además, los trabajos de conservación tienen un papel importante en el comportamiento del IRI ya que algunos trabajos tienen influencia inmediata en el IRI, pero en otras variables no lo refleja de inmediato, mientras que otros trabajos no se refleja de inmediato en el IRI, pero se refleja rápidamente en otras variables. Por lo tanto, es importante identificar los trabajos de conservación para que el modelo aprenda de dichos trabajos.

La presente investigación se basó en el uso de las variables Profundidad de Rodera, Macrotextura y Agrietamiento para estimar el Índice de Regularidad Internacional (IRI). La carretera seleccionada para dicho estudio fue Hermosillo – Santa Ana, la información histórica utilizada fueron de dos años consecutivos (2019 y 2020), solo se seleccionaron dos años para disminuir la dispersión de los datos, además no se contaba con la información de los trabajos de conservación realizados en dicha carretera.

El desarrollo del modelo se llevó a cabo mediante dos enfoques. El primero enfoque con datos particionados se realizó mediante el método del codo para encontrar el



Desarrollo de un modelo de estimación del Índice de Regularidad Internacional (IRI).



número de clases, para después realizar el agrupamiento por medio del algoritmo no supervisado K-Means, donde cada clase es analizada por cinco algoritmos diferentes eligiendo el algoritmo que se desempeñe de mejor manera. Se observó que $K=5$ es el número de clases donde se obtienen mejores resultados.

El segundo enfoque sin datos particionados se desarrolló utilizando la base de datos completa, donde se analizó la base con cinco algoritmos diferentes y comparando sus resultados finales. El algoritmo Random Forest fue el que se desempeño de mejor manera de los cinco algoritmos.

El modelo de estimación final se basó en utilizar información histórica de carreteras donde no tengan trabajos de conservación, las variables utilizadas fueron profundidad de rodera, macrotextura y agrietamientos. El aprendizaje del modelo se llevó a cabo mediante el uso del algoritmo de Random Forest donde se obtienen resultados de 0.7389 para el coeficiente de correlación y de 0.0716 para el error cuadrático medio, lo que nos indica que es el modelo con la mejor precisión a comparación de todos los modelos realizados y analizados.



Bibliografía.

Abdelaziz, A. El-Hakim, A. El-Badawy, S. & Afify H. (2018). Internacional Roughness Index prediction model for flexible pavements.

Badilla, G. (2009). Determinación de la regularidad superficial del pavimento, mediante el cálculo del Índice de Regularidad Internacional (IRI).

Dalla, F. Liu, L. Asce, M. & Gharaibeh, G. (2017). IRI Prediction Model for Use in Network-Level Pavement Management Systems.

Gharied, M. & Nishikawa, T. (2021). Development of Roughness Prediction Models for Laos National Road Network.

Gong, H., Sun, Y., Shu X. & Huang, B. (2018). Use of random forests regression for predicting IRI of asphalt pavements.

Haas, R. Hudson, R. & Zaniewski, J. (1994). Modern Pavement Management.: Krieger.

Hernández, J., Ramírez, M. & Ferri, C. (2004). Introducción a la minería de datos: Pearson.

Hossain, M. Gopiseti, L. & Miah, M. (2017). Prediction of Internacional Roughness Index of flexible pavemenst from climate and traffic data using artificial neuronal network modeling.

Hossain, M. Gopiseti, L. & Miah, M. (2020). Artificial neural network modelling to predict international roughness index of rigid pavements.

Jain, S. Aggarwald, S. & Parida, M. (2005). HDM-4 Pavement Deterioration Models for Indian National Higway Network.



Kargah, N. Tabatabaee, N. Stoffels, S. (2010) Network-Level Pavement roughness prediction model for rehabilitation recommendation.

Marcelino, P., Antunes, M., Fortunato, E. & Castilho, M. (2019). Machine learning approach for pavement performance prediction.

Pineda, C. (2021). Aprendizaje automático y profundo en Python.

Rico, A., Orozco, J., Téllez, R. & Pérez, A. Primera fase Sistema Mexicano para la Administración de los Pavimentos: Instituto Mexicano del Transporte.

Robinson, R. Danielson, U. & Snaith, M. (1998). Road Maintenance Management Concepts and Systems.: Macmillan.

Sigdel, T. & Pradhananga, R. (2021). Development of IRI Prediction Model for National Highways of Nepal.

Solminihac, H. (2005). Gestión de Infraestructura Vial.: Alfaomega.

Tapia, M. (2004). Pavimentos. Universidad Nacional Autónoma de México.

Ziari, H., Sobhani, J., Ayoubinejad, J. & Harmann, T. (2015). Prediction of IRI in short and long terms for flexible pavements: ANN and GMDH methods.