UNIVERSIDAD NACIONAL DE COLOMBIA

Modelo interpretable de Aprendizaje de máquina para la predicción de calificaciones crediticias corporativas

Miguel Ángel Rodríguez Fonseca

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de ingeniería de Sistemas e Industrial
Bogotá, Colombia
2022

Modelo interpretable de Aprendizaje de máquina para la predicción de calificaciones crediticias corporativas

Miguel Ángel Rodríguez Fonseca

Trabajo de grado presentado como requisito parcial para optar al título de:
Magíster en Ingeniería - Ingeniería de Sistemas y Computación

Director:
Ph.D. Germán Jairo Hernández Pérez

Línea de Investigación:
Sistemas Inteligentes

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de ingeniería de Sistemas e Industrial
Bogotá, Colombia
2022

A mis padres, hermano, familia y amigos.

A todos mis amigos y compañeros que me han apoyado durante el desarrollo de mi trabajo de grado.

Este Trabajo Final de maestría fue calificado en mayo de 2022 por el siguiente evaluador:

Germán Guerrero Chaparro Msc.
Facultad de Ciencias Económicas
Universidad Nacional de Colombia

Resumen

Las calificaciones crediticias corporativas son uno de los indicadores financieros de mayor relevancia en el análisis de riesgo crediticio. Estas son generadas por diversas agencias calificadoras las cuales basan sus metodologías en el uso de diversas variables financieras de cada compañía y su impacto en el mercado es de tal importancia que su mala gestión puede desencadenar grandes crisis financieras como la ocurrida en 2008. Junto a esto la gran importancia que tienen los modelos internos de calculo de riesgo de crédito y las regulaciones y acuerdos internacionales buscan que se tenga un nivel de explicabilidad de los métodos empleados al realizar esta gestión de riesgo. En este trabajo se propone el uso de Aprendizaje de Máquina automático (AutoML) como herramienta para la generación de un modelo de aprendizaje de máquina que realice una predicción de las calificaciones crediticias corporativas haciendo uso de datos provenientes de hojas de balance, estados financieros e información descriptiva de la compañía. Adicionalmente, como aporte principal de este trabajo se realizó la inclusión, dentro del AutoML, del nivel de interpretabilidad de cada modelo como un segundo objetivo a optimizar, permitiendo la generación de modelos que puedan explicar de mejor manera sus resultados.

Palabras clave: Aprendizaje de máquina automático, interpretabilidad, riesgo de crédito, inteligencia artificial, calificaciones crediticias corporativas, LIME.

Abstract

Interpretable Machine Learning Model for Corporate Credit Ratings forecasting.

Corporate credit ratings are one of the most relevant financial indicators in credit risk analysis. These are generated by different rating agencies which base their methodologies on the use of various financial variables of each company and the experience of their analysts; their impact on the market is of such importance that their mismanagement can trigger major financial crises such as the one that occurred in 2008. Along with this, the great importance of internal models for calculating credit risk and international regulations and agreements seek to have a level of explainability of the methods used to perform this risk management. This paper proposes the use of Automatic Machine Learning (AutoML) as a tool for the generation of a machine learning model that performs a prediction of corporate credit ratings using data from balance sheets, financial statements and descriptive information of the company. Additionally, the main contribution of this work was the inclusion within AutoML of the level of interpretability of each model as a second objective to be optimized, allowing the generation of models that can better explain their results.

Keywords: AutoML, Interpretability, credit risk, Machine learning, artificial intelli-

gence, Corporate credit rating, LIME.

Índice general

Resumen viii

Lista de figuras 1

1. Introducción 2

2. Justificación y objetivos 5
2.1. Justificación 5
2.2. Objetivo General 6
2.3. Objetivos Específicos 6

3. Antecedentes 7
3.1. Calificaciones crediticias 7
3.2. Agencias calificadoras 8
3.2.1. Metodologías de las agencias calificadoras 8
3.3. Aprendizaje de máquina automático 10
3.3.1. Ingeniería de datos y preprocesamiento de datos 11
3.3.2. Selección del modelo, optimización de hiperparámetros 12
3.4. Interpretabilidad 14
3.4.1. Local vs Global 14
3.4.2. Tipo de datos 16
3.4.3. Modelos específicos vs modelos agnósticos 16
3.4.4. LIME 17
3.5. Trabajos previos de ML para riesgo de crédito 17

4. Análisis e implementación del modelo de aprendizaje de máquina automático 22
4.1. Metodología 22
4.2. Conjunto de datos 24
4.2.1. Fuentes de datos 24
4.2.2. Construcción del conjunto de datos 24
4.2.3. Preparación de los datos 26
4.3. Implementación de AutoML 28
4.4. Evaluación de modelos 30
4.5. Interpretabilidad 31

5. Resultados **34**
5.1. Experimento inicial 34
5.2. Experimento serie de tiempo 38

6. Contribución - AutoML con interpretabiliad como segundo objetivo **39**
6.1. Interpretabilidad como segundo objetivo en AutoML 39
6.2. Resultados AutoML con interpretabilidad 41

7. Conclusiones y recomendaciones **44**
7.1. Conclusiones 44
7.2. Recomendaciones 45

A. Resultados revisión de literatura **47**

B. Anexo: Detalle de modelos generados por TPOT + LIME **51**

Bibliografía **52**

Lista de Figuras

3-1. Escalas de calificaciones de lasa tres principales agencias [Santos, 2007]. . . . 9
3-2. Flujo común en aprendizaje de máquina supervisado [Elshawi et al., 2019]. . 11
3-3. Taxonomía técnicas de interpretabilidad de aprendizaje de máquina. Adaptado de [Linardatos et al., 2020]. 15

4-1. Ciclo de vida descrito por CRISP-DM [Studer et al., 2021]. 23
4-2. Escala de calificaciones dada por Financial modeling prep. 26
4-3. Distribución las compañías por sector 26
4-4. Distribución de calificaciones en el conjunto de datos 28
4-5. Explicación generada por LIME 31

5-1. Matriz de confusión TPOT conjunto de prueba 36
5-2. Matriz de confusión H2O conjunto de prueba 37

6-1. Proceso de explicación local para una instancia 40
6-2. Evolución de modelos por generación y frente de pareto generado al entrenar los modelos de TPOT con LIME 42

A-1. Comparación estudios basados en árboles de decisión. 48
A-2. Comparación estudios basados en máquinas de soporte vectorial. 49
A-3. Comparación estudios basados en redes neuronales artificiales. 50

1. Introducción

Luego de la crisis financiera mundial presentada en 2008, conocida como crisis Subprime, ocasionada en gran parte por la incorrecta asignación de calificaciones crediticias a grupos de hipotecas, en las cuales se indicaba que estos grupos se encontraban en grado de inversión, por lo que generaba confianza en los inversionistas, pero en realidad gran parte de estas hipotecas tenían un riesgo mayor, entrando en default y dando inicio a la crisis financiera. Todo esto debido a la falta de un análisis riguroso por parte de las agencias calificadoras junto con falta de conocimiento e historia de este tipo de instrumentos financieros. [Rom, 2009].

Adicionalmente, regulaciones financieras como la European GDPR, han iniciado a exigir a corporaciones financieras que asignen una calificación crediticia a personas por medio de inteligencia artificial que sus modelos estén en la capacidad de dar explicaciones respecto a los resultados finales y que criterios fueron tenidos en cuenta para esto. Lo cual a futuro puede verse expandido a las asignación de calificaciones corporativas, y en general, a todos aquellos modelos que se construyan con nuevas tecnologías.

El uso e innovación en técnicas de aprendizaje de máquina en el sector financiero se ha visto incrementado en los últimos 12 años, esto debido al gran incremento que se ha presentado en la generación de información desde diversas fuentes y los avances de diferentes tecnologías de la información para la implementación de modelos de inteligencia artificial y aprendizaje de máquina. Entre los casos de uso más comunes se centran modelos de detección de fraude, riesgo crediticio, servicio al cliente, inversión, entre otras.

Esto en simultaneo con el exponencial crecimiento que se ha venido dando en la implementación y facilidades actuales que se tienen para el desarrollo de diferentes métodos de inteligencia artificial y específicamente de Aprendizaje de máquina, nuevas herramientas que permiten un despliegue en la nube o diferentes frameworks que facilitan el entendimiento y proceso general de construcción de estos modelos.

Dentro del alcance de este trabajo se tienen diferentes aspectos, en primer lugar el conjunto de compañías esta limitado a las pertenecientes al mercado de Estados Unidos que cuentan con información pública disponible entre el periodo de 2000 a 2020, para predecir sus calificaciones de largo plazo emitidas por las agencias calificadoras. Un segundo aspecto es el modelo desarrollado, para su construcción se emplea el uso de AutoML para la búsqueda de

un modelo óptimo obtenido automáticamente, con la inclusión de la interpretabilidad como parte del proceso de búsqueda, con el fin de tener un modelo final que permita una mayor explicabilidad de sus predicciones.

Una de las principales limitaciones encontradas durante el desarrollo de este trabajo fue la adquisición y licenciamiento de datos, especialmente de las calificaciones crediticias otorgadas por las principales agencias calificadoras, que son nuestra variable objetivo y son usadas como etiquetas en el entrenamiento supervisado usado para ajustar el modelo. Para superar esta limitación se accedieron a calificaciones otorgadas por una fuente independiente que hace públicos sus ratings dentro de una metodología iterativa la APIs (application programming interface) utilizadas. Otra de las limitantes, al usar búsquedas de soluciones por medio de AutoML con una cantidad amplia de características es el consumo de recursos de computo requeridos, por esto se hizo uso de tecnologías en la nube, como Colab de google para ejecutar estos experimentos de forma adecuada.

Como tema central de este trabajo se tomara el análisis de riesgo de crédito corporativo como caso de uso, con el objetivo de generar un modelo de aprendizaje de máquina que genere como salida la calificación crediticia de una compañía en particular basándose en diferentes variables tomadas de sus propios estados financieros, balances y demás características de la compañía. Con el fin de poder llegar a conocer con anticipación las calificaciones o ratings crediticios asignados por las diferentes agencias calificadoras o implementar como un sistema interno de análisis de crédito en el cual, se tenga un nivel adecuado en la interpretación y explicabilidad de la influencia de cada una de las variables.

Para cumplir con este objetivo se realizara la búsqueda y construcción de este modelo de aprendizaje de máquina por medio de una metodología iterativa que permita ir avanzando en la creación y evaluación de modelos para encontrar la configuración óptima y los resultados de interpretabilidad deseados. Para esto se hará uso de técnicas de aprendizaje de máquina automático, las cuales permitirán realizar el proceso de búsqueda de características y modelos de una forma automática, optimizando así los tiempos y resultados al poder tener la combinación y exploración sobre los diversos modelos ya existentes y reduciendo también la necesidad de tener un conocimiento experto sobre los datos y tipos de problemas a solucionar, esto teniendo en cuenta la alta demanda de este tipo de modelos haciendo que se busque una solución al problema de tener poca expertiz en diversos temas de aplicación por parte de los científicos de datos o personas que se dedican al desarrollo de estos modelos.

Finalmente, para entender con mayor claridad el comportamiento del modelos obtenidos con las diferentes herramientas se realiza un análisis de interpretabilidad, entendiendo la relación de las predicciones con las variables de entrada. También incluyendo un nivel de interpretabilidad dentro de la iteración del AutoML con el objetivo de llegar a modelos que permitan

tener una mayor interpretabilidad a nivel local.

2. Justificación y objetivos

2.1. Justificación

La evaluación adecuada del riesgo crediticio de instrumentos financieros, y compañías del mercado tiene importancia en el proceso de inversión, influyendo en la estructura de los portafolios de inversión y en las decisiones finales de los inversores. Por el contrario la inadecuada asignación de calificaciones por parte de reconocidas agencias, puede ser de tal importancia que debido a ello se dio inició a la gran crisis financiera de 2008, conocida como la crisis Subprime.

Esta crisis fue generada debido a que las agencias calificadoras no fueron lo suficientemente estrictas en la revisión de la información recibidas, pasando por alto señales de advertencia, y a pesar que su misión es realizar evaluaciones crediticias a largo plazo, no se tuvieron en cuenta de manera acertada los factores a largo plazo, en particular las irregularidades contables y las estructuras de financiación de instrumentos demasiado complejos. Esto provoco que se tuviera cierta ignorancia en los prestamos hipotecarios, y respecto a las hipotecas históricas quienes adquirieron los créditos no fueron bien documentados, y al empaquetar estos instrumentos para las agencias se hacía mas complicado evaluar de forma correcta.[Rom, 2009]

Por otra parte, los procesos de evaluación de riesgo de crédito, siempre buscan ser bastante solidos en resultados a la vez que son monitoreados internamente durante la gestión de riesgos. Estos procesos de control interno y regulación permiten una estructura flexible, en la que los bancos pueden adoptar sus propios sistemas para medir sus riesgos y gestionarlos de forma eficiente, pero al mismo tiempo, promoviendo una divulgación de sus procesos internos que estén sujetos a revisión y evaluación, haciendo necesario que los modelos empleados puedan ser explicables, permitiendo entender como la información dada como entrada es evaluada y transformada en un criterio o calificación final de riesgo.

En cuanto a la implementación de un modelo de aprendizaje de máquina, en la literatura se ha encontrado en general un aumento en el uso de estas herramientas en la industria financiera, en una gran variedad de enfoques, desde modelos estadísticos hasta implementaciones de redes neuronales profundas, todas con diferentes resultados y siempre basando su ingeniería de características y selección de modelos en la experiencia previa de los analistas, algunas de ellas presentando pequeños análisis sobre la explicabilidad de sus predicciones, así mismo

durante el análisis de literatura no se evidencia que existan trabajos previos que realicen una exploración sobre el aprendizaje de máquina automático para afrontar este problema y su combinación con técnicas de interpretabilidad que permitan realizar una búsqueda de modelos que tengan un mayor nivel de explicación para cada una de las entradas.

Es por esto que se busca realizar un aporte significativo al buscar una exploración tanto de forma automática como incluyendo la interpretabilidad en el entrenamiento, permitiendo así llegar a resultados prometedores y análisis de interpretabilidad más completos, sin poseer un conocimiento demasiado profundo en el proceso de asignación de calificaciones crediticias y también teniendo un menor preprocesamiento de la información a utilizar.

2.2. Objetivo General

Proponer un modelo de aprendizaje de máquina capaz de hacer predicciones de las calificaciones crediticias de entidades corporativas del mercado norteamericano, el cual mantenga un adecuado nivel de precisión e interpretabilidad.

2.3. Objetivos Específicos

- Identificar las dificultades, variables y distintos modelos de aprendizaje de máquina que han sido empleados para solucionar este problema por medio de un análisis de literatura.
- Definir las variables de entrada y salida del modelo que generan un mejor rendimiento, precisión e interpretabilidad en la predicción basado en los resultados obtenidos en los distintos casos del estudio del arte.
- Implementar el modelo que realice la predicción de la calificación crediticia de una corporación según las variables de entrada ingresadas.
- Evaluar los resultados obtenidos por el modelo de aprendizaje de máquina por medio de métricas de precisión y su nivel de interpretabilidad.

3. Antecedentes

3.1. Calificaciones crediticias

Una calificación crediticia es una opinión formal e independiente en la cual se expresa la habilidad que tiene una compañía o una persona natural de cumplir con sus obligaciones crediticias. Estas calificaciones pueden ser emitidas públicamente o no según la agencia encargada de realizar esta evaluación. Desde el punto de vista del prestatario una calificación crediticia generalmente es un requisito para la emisión de un bono público y acceder a ciertas estructuras de prestamos[Santos, 2007].

Las calificaciones de incumplimiento a nivel de emisor (IDR - Issuer default ratings) son asignadas a entes corporativos, entidades soberana, instituciones financieras, bancos, aseguradoras, compañías de leasing, y entidades financieras publicas como gobiernos locales y regionales. Estas calificaciones son indicaciones de la probabilidad de reembolso de acuerdo con los términos establecidos en la emisión de la deuda y expresan el riesgo en un orden de clasificación relativo, es decir, son medidas ordinales del riesgo crediticio u no predicen una frecuencia especifica de incumplimiento o perdida[Fitch, 2021].

Estas calificaciones juegan un papel importante en el mundo de los mercados financieros ya que con estas los inversionistas realizan parte del análisis para conocer el perfil de una compañía, medir el nivel de riesgo que esto llevaría en sus portafolios y tomar una decisión. Una compañía con una calificación mayor puede obtener un préstamo a un menor costo[Du, 2003].

Las calificaciones crediticias también fomentan el desarrollo y el buen funcionamiento de los mercados de capitales, al proporcionar información transparente y conocimientos a los participantes del mercado. También permiten que las empresas y gobiernos puedan recaudar dinero en los mercados capitales, en lugar de solicitar un préstamo en un banco, piden prestado directamente a los inversores mediante la emisión de bonos o notas. Los inversores compran estos títulos de deuda esperando recibir intereses más la devolución de su principal, ya sea en el momento el vencimiento del bono o en pagos periódicos.

Este proceso de emisión y compra de bonos se ve ampliamente facilitado por las calificaciones crediticias al proporcionar una gestión eficiente y ampliamente reconocida y duradera de la medida del riesgo crediticio relativo.[SP, 2021]

Las calificaciones crediticias pueden verse diferenciadas entre calificaciones de corto plazo (menos de 365 días) y calificaciones de largo plazo (1 año o más), esto debido a las diferentes dinámicas que pueden manejar los inversores pero en general los diferentes inversores, acreedores y otras partes interesadas buscan realizar su análisis sobre las calificaciones de largo plazo de una entidad emisora ya que esta puede ser tomada como una medida general de solvencia.[Santos, 2007]

Estas calificaciones también están clasificadas de acuerdo al nivel al que se esta evaluando, pueden ser asignadas a una emisión en particular, o ser una calificación para el otorgamiento de crédito a una compañía en general. En el caso de este trabajo se realizará la predicción sobre estas últimas.

3.2. Agencias calificadoras

Las agencias calificadoras, como S&P Global Ratings, Moodys ratings y Fitch Rating principales y más reconocidas agencias, son empresas editoriales e informativas que se especializan en analizar el riesgo crediticio de emisores y emisiones de deuda individualmente, estas agencias formulan y difunden sus opiniones de calificaciones para que sean usadas por inversores, ayudándolos a considerar sus estrategias y decisiones comerciales. Estas agencias son vistas como proveedores imparciales e independientes de calificaciones de riesgo de crédito debido a que no están involucradas directamente en las transacciones de mercados de capital.[SP, 2021]

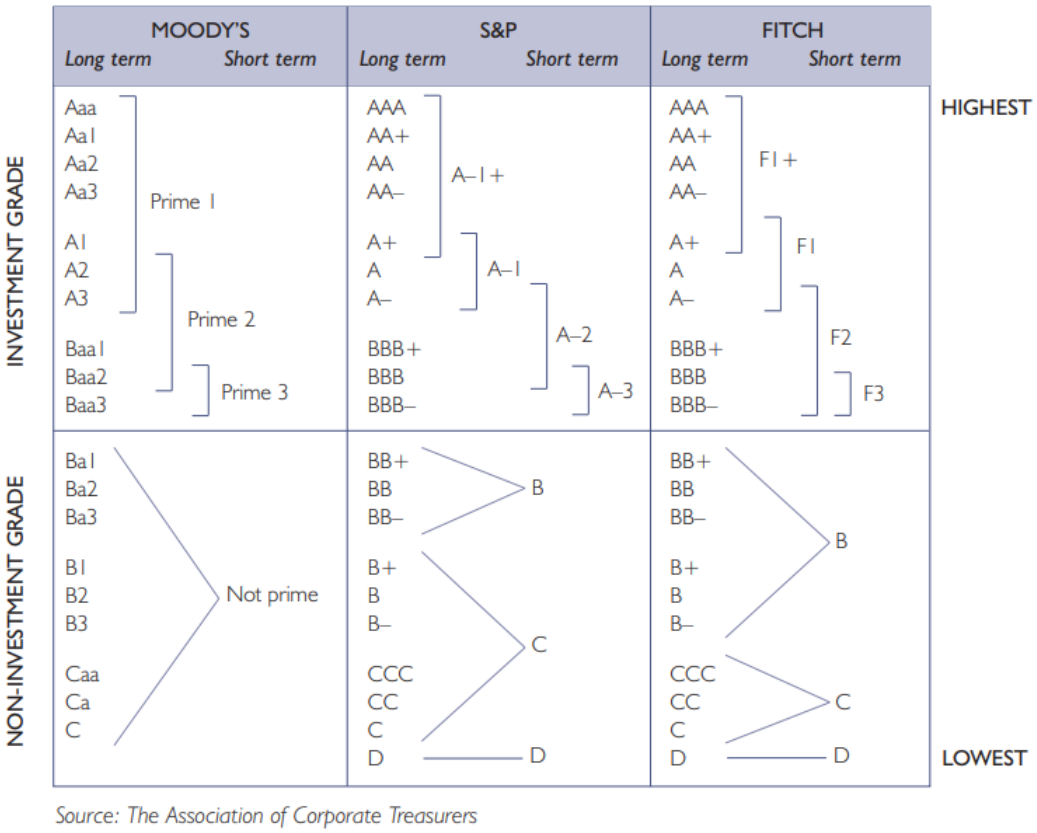
3.2.1. Metodologías de las agencias calificadoras

En general las agencias calificadoras emplean metodologías muy similares para llegar a la determinación de su calificación crediticia, pero al ser cada una independiente en su metodología se pueden ver diferencias en los enfoques tomados y en las calificaciones finales emitidas dependiendo de un sector o producto en especifico, obteniendo algunas discrepancias en las calificaciones emitidas entre ellas.

Cada agencia brinda una descripción detallada de su metodología de calificación en su sitio web para dar un mayor entendimiento en como obtienen su opinión final.

Dentro de las principales áreas en las que se enfocan estas agencias podemos encontrar:

- Riesgo de negocio: En este se evalúan las fortalezas y debilidades de las operaciones de la entidad en análisis, en lo que se puede incluir su posición en el mercado, diversificación



Source: The Association of Corporate Treasurers

Figura 3-1.: Escalas de calificaciones de lasa tres principales agencias [Santos, 2007].

geográfica, fortaleza de su sector económico, fluctuaciones del mercado y dinámicas de compañías competidoras. Con este enfoque se llega a una comparación entre empresas, identificando sus fortalezas y debilidades relativas.

- Riesgo financiero: Evaluación de la flexibilidad de la entidad, con lo cual se analizan medidas de ventas totales y rentabilidad, margenes, expectativas de crecimiento, liquidez, diversidad de financiamiento y predicciones financieras. En la base principal de este enfoque esta el análisis del índice de crédito, el cual es usado para posicionar cuantitativamente a empresas con riesgo de negocio similar.

Otra forma en la que estas metodologías pueden ser divididas y que usualmente son usadas por las agencias son el uso de modelos matemáticos, criterio de analistas o una combinación de estas dos.

- Calificaciones basadas en modelos: Este tipo de metodología enfoca su análisis exclusivamente en datos cuantitativos, los cuales se incorporan en diferentes modelos matemáticos. Como ejemplo, tenemos que para la evaluación de la calidad de los activos, el financiamiento y la rentabilidad de esta entidad la agencia se basaría principalmente en los datos de estados financieros públicos y los registros regulatorios de la compañía.

- Calificaciones impulsadas por analistas: En este tipo de metodología, comúnmente es asignado un analista, junto con un equipo de especialistas que lideraran la evaluación de la solvencia crediticia de la entidad. En este caso los analistas toman información de informes publicados, así como de entrevistas y discusiones con la gerencia de la compañía a evaluar. Con estos datos aplican un juicio analítico para evaluar la situación financiera, el desempeño operativo, las políticas y las estrategias de gestión de riesgos de la entidad.

En general, las agencias calificadoras usan una metodología que se basan en una ponderación entre el análisis de modelos y los resultados por analistas.

3.3. Aprendizaje de máquina automático

Dado el alto incremento en la generación de grandes volúmenes de datos, el uso de técnicas de aprendizaje de máquina ha llamado la atención en diferentes campos de la industria con el fin de dar solución a problemas de alta complejidad y que requieren de cierto nivel de conocimiento experto, el cual en gran parte de los casos no son dominados en su totalidad por los científicos de datos que implementan estas soluciones, generando un mayor gasto de tiempo en el ciclo de desarrollo, en donde se debe realizar las ingeniería de características, selección del algoritmo a utilizar, optimización de parámetros, medición de errores, entre otros. [Elshawi et al., 2019]

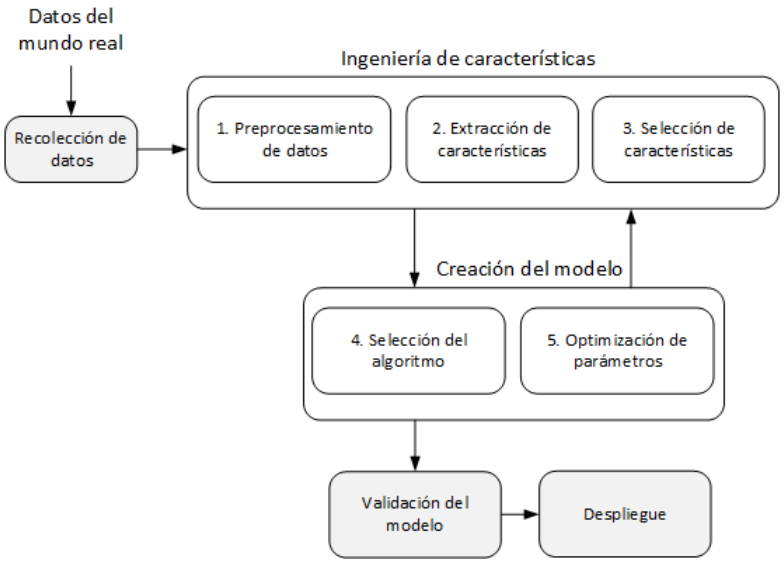


Figura 3-2.: Flujo común en aprendizaje de máquina supervisado [Elshawi et al., 2019].

Con el fin de mejorar este proceso y lograr un mejor desempeño sin tener el conocimiento experto durante el proceso, surge el aprendizaje de máquina automatizado o AutoML, un proceso con el que se automatiza de inicio a fin el proceso de aplicar aprendizaje de máquina a problemas del mundo real, desde automatizar el proceso de ingeniería de variables hasta la optimización de parámetros del modelo. [Yao et al., 2018]

En general, el enfoque de AutoML suele comprender el flujo completo del aprendizaje de máquina, pero este puede ser también utilizado para cubrir únicamente una parte del problema, donde pueda que se necesite más enfoque para llegar a un resultado óptimo, donde no se cuente con el recurso humano que tenga el expertiz.

Dentro de este proceso podemos encontrar las siguientes secciones, las cuales se describen en [Truong et al., 2019]:

3.3.1. Ingeniería de datos y preprocesamiento de datos

La ingeniería de datos, en general, es la primer tarea en el proceso de aprendizaje de máquina y de gran importancia para el rendimiento general del modelo y permitiendo tener un mejor desempeño de las siguientes secciones del proceso.

El objetivo principal de esta fase es lograr construir automáticamente características desde los datos de entrada y que estas puedan ser lo más diferenciables posible. Este tipo de detección de características esta ampliamente relacionado con el contexto y aplicación del problema, así como de la experiencia humana en el campo de aplicación del problema; por

lo cual esta es una de las fases que representa un reto importante a la hora de implementar el AutoML.

Algunos de los métodos más usados en la mejora de características son:

- Reducción de la dimensionalidad: En este proceso se reduce el número de variables a considerar en el modelo. Es de gran utilidad cuando se tienen grandes cantidades de variables en las que hay información redundante. Se puede llevar a cabo seleccionando un subconjunto de las variables originales o transformando las ya existentes para llegar a un conjunto reducido de ellas.

- Generación de características: En este proceso se busca explorar relaciones entre las variables originales, con el fin de generar nuevas características que puedan llegar a mejorar significativamente el rendimiento en el aprendizaje.

- Codificación de características: Este método reinterpreta las características originales basado en diccionarios aprendidos desde los datos, buscando representar la información de las características en una forma que pueda llegar a ser interpretado de mejor manera por el modelo.

3.3.2. Selección del modelo, optimización de hiperparámetros

Para la selección del modelo, una vez se tienen las características de entrada definidas, estas son usadas para entrenar una variedad de modelos, cada uno con varios conjuntos de parámetros, los cuales se buscan optimizar, obteniendo el modelo o conjunto de modelos como el modelo final. Este proceso se realiza de maneras diversas por medio de herramientas y librerías disponibles que generan modelos por medio de AutoML, cada una contiene un conjunto diferente de algoritmos de aprendizaje de máquina entre los que se pueden encontrar regresiones logísticas, algoritmos basados en arboles de decisión, máquinas de soporte vectorial y modelos de redes neuronales.

En algunas herramientas es común el uso de algoritmos genéticos para ir evaluando iterativamente los mejores modelos de cada época y utilizarlos en la siguiente generación; otros método comúnmente utilizado para obtener los mejores modelos es la búsqueda de arquitectura neuronal (NAS) por sus siglas en ingles.

En cuanto a la optimización de hiperparámetros, podemos dividir las técnicas utilizadas en cuatro categorías [Yao et al., 2018]:

- *Búsquedas simples:* En este enfoque las técnicas utilizadas no hacen suposiciones acerca del espacio de búsqueda, los dos enfoques más comunes son:

- *Búsqueda de rejilla*: Esta es la forma más común de realizar la optimización de hiperparámetros. Aquí para obtener el parámetro óptimo se enumeran un conjunto de posibles configuraciones, de las cuales se prueban todas y se selecciona la combinación que tuvo un mejor desempeño. En los espacios de búsqueda continuos esto significa tener que discretizar para tener el nuevo conjunto de búsqueda.
- *Búsqueda aleatoria*: En este enfoque se toman diferentes configuraciones aleatoriamente desde el espacio de búsqueda para seleccionar la mejor configuración posible. Suele tener un mejor resultado que la búsqueda por rejilla ya que puede explorar una mayor parte del espacio de búsqueda.

- *Derivative-Free optimization:* Dentro de esta categoría se encuentran los métodos que permiten optimizar la búsqueda basándose en muestras, con lo cual generan diferentes configuraciones iterativamente tomando como base las configuraciones generadas y evaluadas previamente.
  - *Búsqueda heurística*: Las búsquedas heurísticas están inspiradas en comportamientos y fenómenos biológicos. La mayoría de estas técnicas están basadas en métodos de optimización de poblaciones, donde cada población es un conjunto de configuraciones de un modelo especifico y durante cada iteración surge una nueva población basada en los mejores individuos evaluados. Un ejemplo son los algoritmos genéticos, en los cuales entre cada nueva generación se realizan operaciones de cruce y mutación sobre los mejores individuos generando uno nuevo, el cual será un candidato promisorio en la siguiente generación.
  - *Aprendizaje por refuerzo*: Este método contiene una política la cual actúa como optimizador, y su rendimiento real en el entorno es medido por el evaluador. Sin embargo, a diferencia de los métodos anteriores, los feedbacks (es decir, la recompensa y el estado) no necesitan ser devueltos inmediatamente una vez que se realiza una acción, sino que pueden ser devueltos después de realizar una secuencia de acciones.

- *Gradiente descendiente*: Los problemas de optimización de AutoML son muy complejos, y el objetivo no suele ser diferenciable o incluso no es continuo. Por ello, el gradiente descendiente no es tan popular como los métodos anteriores. Sin embargo, centrándose en algunas funciones de pérdida diferenciables, por ejemplo, la pérdida cuadrada y la pérdida logística, los hiperparámetros continuos pueden ser optimizados por el descenso de gradiente. En comparación con los métodos anteriores, los gradientes ofrecen la información más precisa donde se localizan las mejores configuraciones. A diferencia de los problemas de optimización tradicionales cuyos gradientes se pueden derivar explícitamente del objetivo, en los problemas de AutoML, los gradientes necesitan ser calculados numéricamente.

- *Búsqueda voraz (Greedy search)*: La búsqueda voraz es una estrategia natural para resolver problemas de toma de decisiones en varios pasos. Sigue una heurística que toma una decisión localmente óptima en cada paso con la intención de encontrar un óptimo global. La búsqueda voraz no puede encontrar el óptimo global, pero normalmente puede encontrar un óptimo local que se aproxima al óptimo global con un coste de tiempo razonable. Además, este buen rendimiento empírico también se justifica teóricamente en muchas aplicaciones, como la selección de características y la optimización submodular.

3.4. Interpretabilidad

El aumento en el uso de modelos de aprendizaje de máquina para la búsqueda de soluciones complejos dentro diversas industrias y campos de desarrollo ha traído consigo un nuevo reto, principalmente en aquellos sectores en los cuales se hace necesario un entendimiento mayor de las decisiones o salidas generadas por estos modelos por el gran impacto que estos tienen en la sociedad [Carvalho et al., 2019], y es que los modelos que solucionan estos problemas resultan ser algoritmos complejos que funcionan como cajas negras, lo cual no permite al humano entender la lógica y funcionamiento interno con los que se llega al resultado final, además de tener ciertas regulaciones en algunos de los dominios de aplicación de estas técnicas, tales como el sector de la salud y el sector financiero.

Las técnicas de interpretabiliad para algoritmos de aprendizaje de máquina pueden ser clasificadas de distintas maneras según el punto de vista desde el que se quieran estudiar y el objetivo planteado al aplicarlos sobre un algoritmo. En la figura **3-3** se presentan algunos de los aspectos con los cuales los métodos de interpretabilidad pueden ser agrupados [Linardatos et al., 2020].

3.4.1. Local vs Global

Dentro de esta clasificación se identifican dos tipos de modelos de interpretabilidad:

Local

En este caso se busca explicar una única predicción del modelo, es decir, poder ver cuales características o variables fueron las que tuvieron mayor relevancia para la decisión final de una predicción especifica,la idea principal es enfocarse en esa única instancia y buscar entender como el modelo llegó a la predicción dada. Esto se realiza aproximando una pequeña región de interés en un modelo en el que no se tenga la explicación de sus predicciones utilizando un modelo interpretable de menor complejidad. El razonamiento con el que se

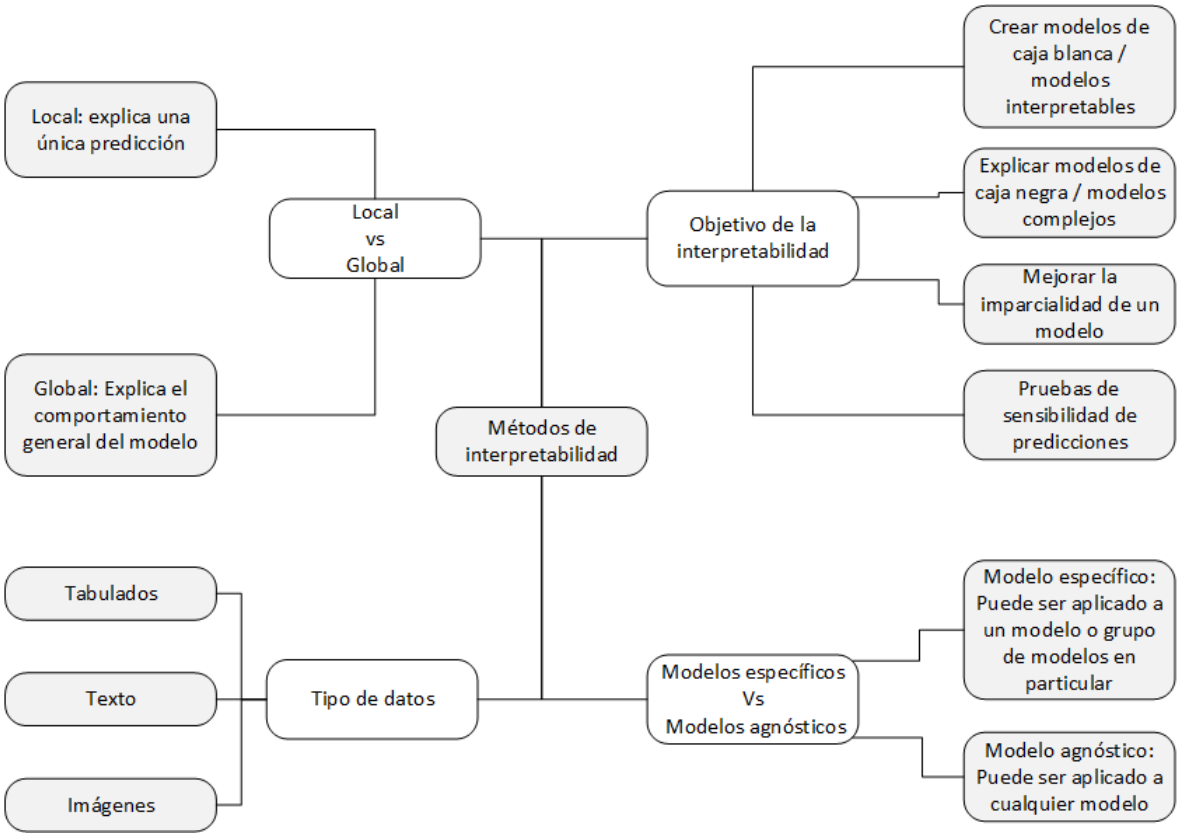


Figura 3-3.: Taxonomía técnicas de interpretabilidad de aprendizaje de máquina. Adaptado de [Linardatos et al., 2020].

llega a este método es que, a nivel local, la predicción podría depender solo de manera lineal o monótona de algunas características en lugar de tener una dependencia compleja de ellas.[Carvalho et al., 2019]

Global

Los métodos globales describen el comportamiento medio de un modelo de aprendizaje automático. Los métodos globales suelen expresarse como valores esperados basados en la distribución de los datos. Por ejemplo, el gráfico de dependencia parcial, un gráfico de efecto de características, es la predicción esperada cuando todas las demás características se marginan. Dado que los métodos de interpretación global describen el comportamiento medio, resultan especialmente útiles cuando el modelizador quiere comprender los mecanismos generales de los datos o depurar un modelo. [Molnar, 2022]

3.4.2. Tipo de datos

Uno de los aspectos importantes a tener en cuenta al utilizar un método de interpretabilidad es el tipo de datos con el que pueden trabajar, los tipos de datos más comunes suelen ser datos tabulares o imágenes, pero también se pueden encontrar métodos que acepten textos como entrada para el análisis de interpretabilidad.

3.4.3. Modelos específicos vs modelos agnósticos

Modelo-específico

Los métodos de interpretación que son modelo-específicos son aquellos que únicamente pueden ser aplicados a cierto tipo de algoritmos de aprendizaje de máquina por lo que pueden llegar a presentar ventajas y ser más útiles para la evaluación de ciertos algoritmos de los que ya se tiene un interpretabilidad especifica respecto a un análisis por un método modelo-agnóstico.

Modelo-agnóstico

Los métodos de interpretabilidad modelos-agnósticos son aquellos que no son dependientes del modelo para su aplicabilidad, esto permite usar por ejemplo el mismo método de interpretabilidad para comparar el funcionamiento de dos algoritmos de aprendizaje de máquina que busquen dar solución al mismo problema. Estos métodos siempre son utilizados una vez el algoritmo ha sido entrenado (post hoc) y por definición, estos métodos no pueden tener acceso a los modelos internos, como los pesos o la información estructural.[Molnar, 2022]

3.4.4. LIME

(Local Interpretable Model-Agnostic Explanations) por sus siglas en inglés es un método de análisis de interpretabilidad en el cual se usan modelos sustitutos locales para explicar las predicciones individuales de los modelos de aprendizaje de máquina que se consideran como çaja negra".

El objetivo principal es entender por qué el modelo realiza cierta predicción, para esto se realizan variaciones a los datos cercanos al punto del cual queremos obtener la explicación y sobre estas variaciones se realizan las predicciones en el modelos de caja negra. con este nuevo conjunto de datos, las muestras permutadas y sus correspondientes predicciones LIME realiza el entrenamiento de un modelo interpretable, teniendo en cuenta el nivel de proximidad de cada permutación con la muestra original asignando un mayor peso a las más cercanas. Este nuevo modelo debería ser una buena aproximación de las predicciones del modelo de caja negra de manera local y no necesariamente de manera global, esta precisión es conocida como fidelidad local.

La explicación producida por LIME se obtiene por:

$$explicacion(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{3-1}$$

Donde g es el modelo interpretable seleccionado dentro de la familia de los posibles modelos interpretables G. Como no todo $g \in G$ podría ser lo suficientemente simple para ser interpretable se adiciona $\Omega(x)$ como una medida de complejidad de la explicación $g \in G$. En problemas de clasificación, $f(x)$ es la probabilidad de que x pertenezca a cierta clase. Para definir la medida de proximidad entre una instancia z a x usamos π_x. Finalmente $L(f, g\pi_x)$ como la medida de que tan poco fiel es g en aproximación a f en la vecindad definida por ${}_p i_x$. Para alcanzar un alto nivel de interpretabilidad y fidelidad local se debe minimizar $L(f, g\pi_x)$.

3.5. Trabajos previos de ML para riesgo de crédito

Podemos encontrar que anteriormente se han llevado diferentes estudios y trabajos en los que se aborda en el reto de la predicción de calificaciones crediticias, muchos de ellos basados en calificaciones a nivel de bonos y también podemos encontrar un amplio campo de estudio en la predicción de calificaciones crediticias para personas naturales, que comúnmente utilizan las entidades financieras comerciales para realizar estudios en la asignación o aprobación de créditos a sus usuarios. Dentro de este conjunto de estudios encontrados se ha evidenciado que hay gran variedad en el tipo de modelos utilizados para abordar este problema, donde se puede encontrar desde métodos estadísticos hasta la implementación de recientes modelos

de aprendizaje de máquina que cuentan con un mayor nivel de complejidad.

A continuación se hace un breve repaso por estos trabajos, donde se identifica su principal aporte, los métodos o modelos usados y los conjuntos de datos usados, posteriormente se mostraran algunos de los resultados obtenidos en estos trabajos.

En los últimos años se ha encontrado que la cantidad de trabajos relacionados ha estado en crecimiento y se han realizado implementaciones de diferentes modelos, los cuales según los conjuntos de datos usados, las métricas de evaluación, la locación geográfica de los datos a evaluar y los mismos modelos han presentado diferentes resultados. Dentro de los trabajos realizados por medio de métodos estadísticos, el primero a destacar es [Du, 2003], en el cual, tomando parte de los trabajos realizados por Horrigan en 1966, West en 1970-1973, Pogue and Soldofsky en 1969, Kaplan and Urwitz en 1979 y Blume, Lim and Mackianly en 1998 buscando extenderlos a situaciones de interés donde se presentaban algunas limitaciones, especialmente para la identificación de cambios en la calificaciones de compañías especificas, para el desarrollo de este trabajo se usaron modelos logísticos y modelos de tasa de riesgo, y usando datos de calificaciones previas, duración, información contable e información de mercado, en este caso se evaluaron las calificaciones entre 1995 y 1998, obteniendo una precisión entre 49.17 % y 51.19 %.

También haciendo uso de modelos estadísticos, como regresiones y modelos z-score encontramos a [Bandyopadhyay, 2006] con una investigación para mercados emergentes en India, donde se realiza una clasificación de compañías entre aquellas que pueden entrar en default y aquellas que son solventes, fueron usadas 542 compañías de la India con datos entre 1998 y 2004 donde posteriormente se seleccionaron 52 compañías con default y 52 solventes. En las variables usadas, este estudio usa propiedades tomadas de los flujos de efectivo, junto con información de las obligaciones, movimientos de activos y algunas características no financieras de cada compañía. El modelo Z-score tiene una capacidad predictiva del 91 % para datos dentro de la muestra y puede detectar firmas malas con un año de anterioridad entre un 92 y 88 % de exactitud.

En cuanto a los modelos más complejos, se han realizado diferentes estudios que abarcan maquinas de soporte vectorial, árboles de decisión y redes neuronales. Dentro de estos se han encontrado diferentes enfoques que realizan diferentes estructuras y arquitecturas de redes, comparando el resultado entre ellas, como lo podemos ver en [Golbayani et al., 2020b], en donde con un conjunto de compañías de tres sectores económicos diferentes: energía, finanzas y salud; con datos entre 2000-2016. En este estudio, se realiza la predicción de la calificación emitida por la agencia Standard and Poors, con datos obtenidos desde las plataformas Bloomberg y compustat, y usando tres diferentes arquitecturas de redes: Perceptrones multicapa (MLP), redes neuronales convolucionales (CNN) y Redes recurrentes de memoria de

corto y largo plazo (LSTM). En estas se obtuvieron diferentes para cada uno de los sectores, destacando la red convolucional con una mejor media de precisión, de 87 % mientras que el perceptron multicapa con una media de 80 %. Por su parte los experimentos realizados en los que se tuvo en cuenta los tiempos anteriores, realizados con redes LSTM, obtuvieron niveles de precisión entre 74 % y 91 % los cuales varían entre cada sector y época analizada, se evidenció mejor desempeño en los años más recientes del conjunto de datos.

Un aspecto de gran importancia que se ha encontrado en la revisión de la literatura, es el reciente interés que se está presentando por partes de bancos centrales, como lo son el de España e Italia, los cuales han realizado publicaciones en las que se hace uso de Machine Learning para afrontar el problema o el estudio del riesgo de crédito. Por su lado, el Banco de España, en [Alonso and Carbo, 2020], realiza un estudio en cuanto a los beneficios del uso de modelos de aprendizaje de máquina en este ámbito, y buscando encontrar una forma de realizar una comparación entre el dilema de la predicción vs el costo de un modelo supervisado. En [Moscatelli et al., 2020], el Banco de Italia muestra como se esta haciendo una transición en el mundo financiero, pasando de modelos estadísticos a nuevas técnicas de ML, para esto realiza la comparación de estos modelos estadísticos con modelos basados en árboles, en estos experimentos han obtenido scores de AuROC (Area under the Receiver Operating Characteristic) de entre 81.1 % y 83.4 %, también mostrando importancia de las variables a nivel global para cada uno de los modelos usados.

Una revisión de literatura más detallada acerca de la predicción de calificaciones corporativas, es realizada en [Golbayani et al., 2020a], donde se realiza una comparación especifica entre los diferentes estudios realizados anteriormente que usaron modelos de aprendizaje de máquina para el problema de calificaciones crediticias corporativas, aquí se hace una segmentación entre los enfoques utilizados, dividiendo en tres grandes grupos los trabajos previos: árboles de decisión, máquinas de soporte vectorial y redes neuronales artificiales. Se resalta en el análisis de literatura la variedad de resultados obtenidos, teniendo modelos con desempeños por debajo de 50 % hasta modelos que alcanzan una precisión cercana al 90 %, enfatizando en el hecho de que cada uno tiene un conjunto de pruebas diferente y pueden estar enfocados en sectores o geografías muy especificas que podrían significar una complejidad mayor o menor, lo cual se complementa con la forma en la que en algunos trabajos se dividen los datos, ya que muchas de las calificaciones y variables no tienen bastantes cambios en el tiempo, haciendo que los conjuntos de pruebas y entrenamiento sean muy similares.

A continuación se presenta un recuento de la cantidad de modelos incluidos por [Golbayani et al., 2020a] en su revisión de literatura:

Tipo de modelo	Cantidad de estudios
Árboles de decisión	8
Máquinas de soporte vectorial	15
redes neuronales artificiales	21

En el anexo A, se detallan los resultados de cada modelo obtenidos en la revisión de literatura de [Golbayani et al., 2020a].

En cuanto a estudios que involucran interpretabilidad, dentro de este campo de investigación se han encontrado recientes publicaciones que hacen énfasis en el uso de estas herramientas para suplir el uso de modelos de tipo caja negra, como lo descrito por [Rudin, 2019], donde se propone el uso de más modelos interpretables en comparación a intentar explicar complejos modelos de Aprendizaje de máquina considerados como caja negra, esto soportado en demostraciones de que el hacer uso de interpretabilidad no implica una disminución de la precisión en los resultados finales de los modelos.

Otro caso reciente, es el expuesto en [Hervé and Erwan, 2020], donde también justificado en la post crisis de 2008 y a las regulaciones se ve un incremento del uso de modelos de aprendizaje de máquina, se da un recorrido por los diferentes métodos de interpretabilidad y explicabilidad que pueden ser aplicados a los modelos e aprendizaje de máquina, todo esto es realizado sobe la base de la evaluación a riesgo de crédito a nivel personas.

En [Bussmann et al., 2021] se propone el uso de un modelo XGBoost, que consta de un conjunto de arboles de decisión, para afrontar la predicción de un default de crédito dentro del mercado europeo, en este caso únicamente se tiene una salida binaria de Default o no default, obteniendo una precisión de hasta 0.93 y realizando un análisis de explicabilidad promedio del modelo por medio de el método TreeSHAP.

Finalmente, como el antecedente más cercano en el que se hace uso de interpretabilidad y aprendizaje de máquina automático para análisis riesgo de crédito, se tiene el descrito por [Patron et al., 2019], donde usando el dataset "Give me some credit", disponible en una competencia de la plataforma Kaggle, se aborda el problema de dar o no dar un crédito a una persona. En este trabajo se buscó realizar la construcción de un modelo de aprendizaje de máquina, generado automáticamente que lograra obtener un desempeño igual o superior al obtenido por en los primeros lugares de la competencia, los cuales se realizaron con modelos tradicionales, realizando un tratamiento rigurosos sobre los datos y hacen una selección manual del modelo a partir de su experiencia y experimentación. Los resultados del modelo final encontrado (0.861) alcanzaron un puntaje muy cercano al de la competencia (0.869). Finalmente ellos realizaron un análisis de interpretabilidad local por medio de LIME identificando las características que influyen en esta decisión, por ejemplo las personas con mayor

edad son consideradas más seguras de pagar en el tiempo.

4. Análisis e implementación del modelo de aprendizaje de máquina automático

4.1. Metodología

La metodología utilizada en la implementación de este proyecto tiene dos bases principales, primero el flujo común de aprendizaje de máquina, enmarcado por diferentes autores y descrito previamente en este trabajo en la figura 3-1, y segundo siguiendo las fases propuestas por el modelo CRISP-DM [Studer et al., 2021] por sus siglas en ingles (*Cross-Industry Standard Process for Data Mining*) y el cual puede ser totalmente adaptado al proceso de desarrollo de modelos de aprendizaje automático, como se puede ver en [Studer et al., 2021] con el modelo propuesto CRISP-ML(Q) - *CRoss-Industry Standard Process model for the development of Machine Learning applications with Quality assurance*. methodology.

Esta es una metodología iterativa que sigue los siguientes pasos:

1. *Entendimiento del negocio*
   - Conocer los conceptos básicos y funcionamiento de las calificaciones crediticias corporativas, agencias calificadoras.
   - Definir los objetivos y su traducción al problema de aprendizaje de máquina
2. *Entendimiento de los datos*
   - Definición de las variables y sus fuentes, que pueden aportar en el proceso de clasificación y predicción de las calificaciones crediticias.
   - Definición del conjunto de datos de salida que será usado como etiqueta en el proceso de entrenamiento.
3. *Preparación de los datos*
   - Validación y combinación de los datos obtenidos desde diferentes fuentes.
   - Limpieza de los datos obtenidos.

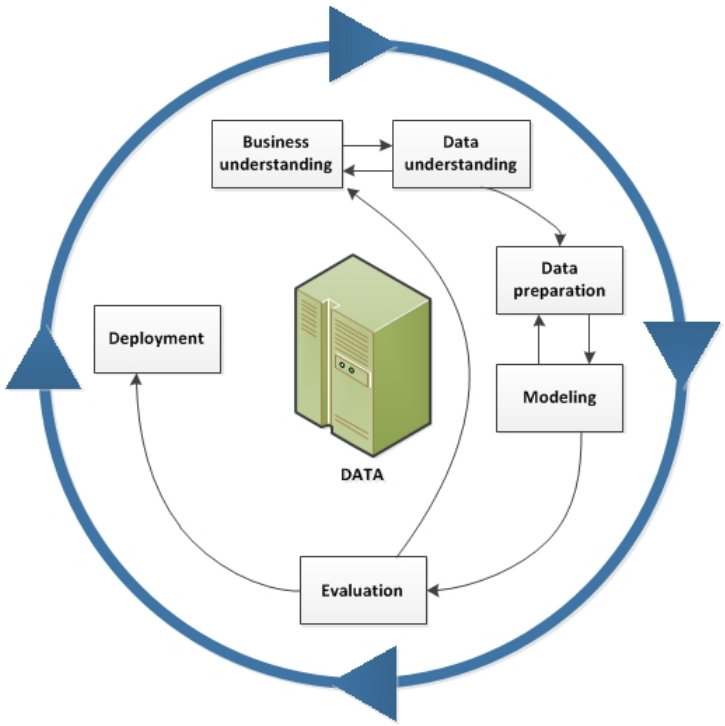


Figura 4-1.: Ciclo de vida descrito por CRISP-DM [Studer et al., 2021].

- Separación de los conjuntos de datos entre los datos de entrenamiento, prueba y validación para los diferentes experimentos a realizar.

4. *Modelación*

   - Configuración de las ejecuciones de entrenamiento a realizar desde las herramientas de aprendizaje de máquina.
   - Entrenamiento automático de los modelos por medio de las herramientas de aprendizaje de máquina.

5. *Evaluación*

   - Validación de los mejores modelos obtenidos durante las ejecuciones del aprendizaje de máquina automático.
   - Comparación entre los modelos obtenidos por las diferentes herramientas utilizadas.
   - Análisis de interpretabilidad de los modelos obtenidos.

4.2. Conjunto de datos

Como se ha descrito en la sección 3.1 y 3.2, el proceso de asignación de una calificación crediticia a una compañía en particular esta principalmente basado en su información descriptiva, información de estados financieros, hojas de balances, flujos de efectivo, entre otros, y dependiendo de la agencia calificadora estos aspectos pueden llegar a tener mayor o menor relevancia en el modelo.

Para el alcance de este trabajo se tomaran en cuenta un conjunto de compañías pertenecientes al mercado de Estados Unidos de América.

4.2.1. Fuentes de datos

Para la construcción del conjunto de datos utilizado en la implementación de este proyecto se usaron dos diferentes plataformas de información financiera que se encuentran disponibles de manera abierta para descargar por medio de API la información financiera publica de cada compañía:

- Financial Modeling Prep[1]
- SimFin[2]

Estas dos plataformas permiten la descarga masiva de datos en formato JSON y contienen la información pública de la compañía desde su primera aparición en el mercado hasta el año inmediatamente anterior debido al tipo de licencia usada para este proyecto, lo cual no implica ningún problema para realizar las predicciones de las calificaciones.

4.2.2. Construcción del conjunto de datos

Dentro de la información disponible dentro de la plataforma de SimFin se encuentra un listado general de las compañías pertenecientes al mercado de los Estados Unidos con su correspondiente Ticker, el cual puede ser usado como llave para la descarga de la información financiera desde cualquiera de las dos plataformas.

Del análisis de literatura realizado, donde se estudiaron los trabajos previos mencionados en 3.5, se ha evidenciado que las variables en común que generan mejores resultados se encuentran registradas en los estados financieros, balances, flujos de efectivo y descripciones de las compañías. Ya que para el desarrollo del modelo se implementará AutoML, el cual nos

[1] https://simfin.com/data/api
[2] https://financialmodelingprep.com/developer/docs

permite realizar una selección de características, se tomo el conjunto completo proveniente de estos reportes para que sean seleccionados durante el entrenamiento.

Una vez se obtuvo el listado de las compañías se construyeron los scripts que permitieron la descarga y almacenamiento de esta información disponible (en cuanto a estados financieros, balances, descripciones, entre otros) en los últimos veinte años (2000-2020) de forma trimestral para cada compañía en archivos delimitados por comas (CSV). Se toma este periodo de tiempo debido a que en este periodo se cuenta con mayor disponibilidad, también permite tener un progreso y visualización de como variaron las calificaciones debido a las crisis entre 2008 y 2009.

En este punto se obtuvo información de 2060 compañías, las cuales dependiendo de su antigüedad en el mercado de los Estados Unidos contaban con reportes históricos de entre 1 y 78 trimestres previos y en cada uno de estos registros con información de 64 diferentes variables, la dimensión total de este conjunto de datos inicial es de 56181 x 64.

En cuanto a la información correspondiente al histórico de las calificaciones crediticias, la cual es la variable objetivo en el modelo a construir, al ser esta propiedad de cada una de las agencias calificadoras no se encontraba disponible para la descarga masiva dentro de las plataformas utilizadas, por esto no fue posible utilizar la información de alguna de las tres principales agencias calificadoras: Moody´s Ratings, S&P Ratings y Fitch. En su lugar se utilizó el rating proporcionado por el proveedor de datos Financial Modeling Prep, el cual es calculado directamente por el proveedor de datos con la metodología descrita en [3] , es una calificación de uso libre y se encuentra definida con la siguiente escala descrita en la figura **4-2**

Para el alcance de este trabajo, se toma como supuesto que estas calificaciones son correctas, ya que provienen directamente de una agencia calificadora y las predicciones finales estarían dadas a coincidir sobre esta misma escala. Otro caso, que no está dentro del alcance actual, es aquel en el que se quisiera predecir una calificación nueva, como puede ser el ponderado entre 2, 3 o más agencias, en este caso se debe analizar con detalle que la variable que se usará como entrada al etiqueta en el entrenamiento del modelo sea calculada correctamente, de lo contrario podría tener afectaciones dentro de las predicciones finales.

Contando con la información de la variable objetivo y combinando esta información con el conjunto de datos inicial se obtuvo una cantidad menor de registros, esto debido a que no todas las compañías en el conjunto inicial tenían una calificación asignada dentro de la escala de rating utilizada en este caso. Las dimensiones del conjunto de datos final son 47777 x 65.

Cómo se puede observar en la figura **4-3**, las compañías que componen el conjunto de datos

[3] https://site.financialmodelingprep.com/developer/docs/recommendations-formula/

Figura 4-2.: Escala de calificaciones dada por Financial modeling prep.

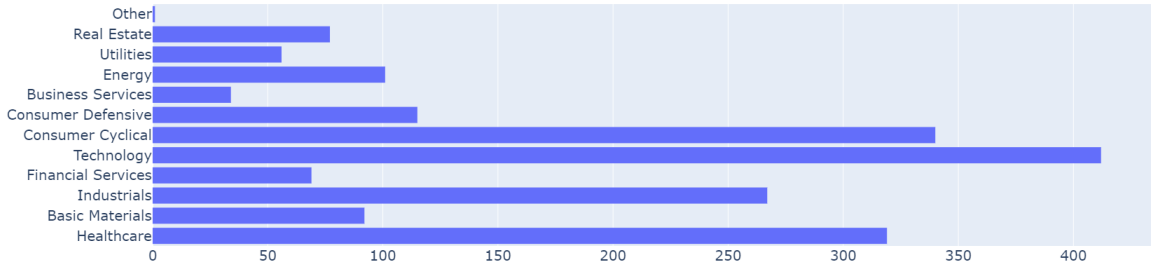


Figura 4-3.: Distribución las compañías por sector

están distribuidas en 11 sectores, entre los cuales predominan las compañías del sector tecnológico, seguidas por las de consumo cíclico, salud e Industrial.

4.2.3. Preparación de los datos

Una vez con el total de los datos descargados, en archivos independientes por tipo de reporte, se realizó la unificación de esta información, (usando como llave principal el Ticker de cada compañía, el cual permite identificar con claridad cada una dentro de el conjunto de archivos y permanece durante el tiempo), en un archivo principal con el contenido de todos los reportes utilizados, en el proceso de esta unificación se dejaron a un lado los registros en los cuales no se contara con la información completa de los reportes (Descripción, Flujo de efectivo, hoja de balance, estado financiero ó rating).

Debido a la naturaleza del problema presentado y como se explica en la sección 4.4, las calificaciones a predecir se caracterizan por estar ordenadas de forma especifica, con lo cual la etiqueta original que se encuentra dada por letras es homologada a una escala numérica en la que un menor grado de inversión es relacionado con un valor menor, mientras que las calificaciones que corresponden a mejores grados de inversiones se homologan a valores más

altos como se puede apreciar en la figura **4-2**.

Algunas de las variables contenidas dentro de los reportes anteriormente mencionados son datos categóricos los cuales se encuentran representados por cadena alfanuméricas, entre estas podemos encontrar el sector económico al cual pertenece la compañía y el tipo de industria, estos dos de gran importancia en algunos de los trabajos previos mencionados en la sección 3.5, para la utilización de estas variables dentro de las herramientas de AutoML utilizadas, TPOT y H2O, los datos contenidos fueron transformados por medio de *One hot encoding*, en el cual se genera una columna para cada una de las opciones posibles dentro de una variable categórica, ya que este método permite transformar estas variables en variables numéricos y evitando que algunos algoritmos de aprendizaje de máquina malinterpreten el orden de los números asignado valorando con un peso diferente cada una de las categorías.

Otro punto importante en el proceso de preparación de los datos, es que al revisar la distribución de cada una de las clases, en este caso calificaciones, se encontró que la gran mayoría se encuentran dentro de las calificaciones A+ y A mientras que la cantidad de registros que se encuentran en las calificaciones más bajas son menos, figura **4-4**. Para esto, durante este trabajo estas categorías inferiores se han agrupado en una categoría llamada *Default*, en la cual ya se tomaría a la compañía en un nivel de riesgo mayor al que se estaría dispuesto a asumir, este umbral podría variar según la entidad financiera que este realizando el análisis, ya que para algunos fondos de inversión pueden resultar interesantes las rentabilidades que pueden llegar a producir invertir en estas calificaciones más bajas.

Dentro de este trabajo se decidió unificar las calificaciones inferiores a B- en la categoría llamada *Default*, asumiendo que las compañías que se encuentren por debajo de esta calificación no serian elegibles para invertir. Para inversores que tengan un nivel de riesgo más alto, la agrupación se realizaría sobre calificaciones de un nivel más bajo, pero se debe tener en cuenta el tener suficientes muestras para que el modelo logre aprender de estas clases durante el entrenamiento.

Finalmente con las transformaciones mencionadas realizadas se dividió el conjunto de datos en dos, Train y Test, para esto y teniendo en cuenta los problemas de evaluación que se pueden presentar al tener registros muy similares en los conjuntos de entrenamientos y pruebas, mencionados en [Golbayani et al., 2020a], se dividió el grupo de nombres de compañías que se tenia inicialmente en un conjunto de 70 % para entrenamiento y un 30 % para pruebas, finalmente al tener en cuenta el total de registros de cada compañía se obtuvo un conjunto de entrenamiento con un total de 35832 (75 %) y un conjunto de pruebas con 11945 (25 %).

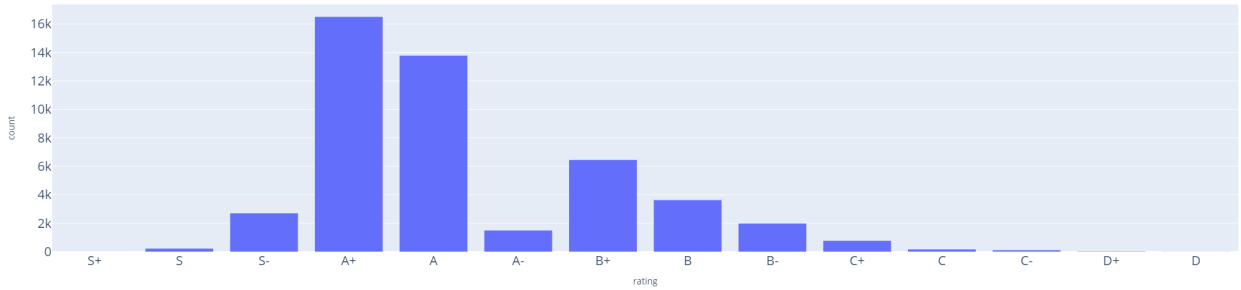


Figura 4-4.: Distribución de calificaciones en el conjunto de datos

4.3. Implementación de AutoML

Para la experimentación de este proyecto se implementaron modelos de aprendizaje de máquina automatizado (AutoML) en python por medio de las librerías TPOT[4] y H2O[5], estas dos librerías de uso libre permitieron realizar la experimentación para evaluar diferentes conjuntos de características, diferentes tipos de modelos y métricas de evaluación con el fin de buscar el modelo con un mejor desempeño en la predicción de las calificaciones crediticias.

Dentro de la implementación de ambas librerías se han dejado abiertos los tipos de modelos disponibles dentro de la búsqueda con el fin de explorar la mayor cantidad de modelos y combinaciones posibles que presenten un mejor resultado. Para el caso de TPOT que funciona por medio de la búsqueda heurística, específicamente con el uso de algoritmos genéticos, se configuró la búsqueda de un modelo de clasificación donde su variable objetivo sea el valor numérico asignado a cada una de las calificaciones, ya que el principal objetivo al implementar un sistema de predicción de las calificaciones crediticias es identificar aquellas compañías a las que se les asignan una probabilidad de default o que se encuentren por debajo de un limite de calificación establecido.

Dentro de la diferencia encontrada en la preparación de los datos para ser usados en ambas librerías, encontramos que la inclusión de variables no numéricas dentro de los conjuntos de pruebas requería un tratamiento distinto en cada caso, para la librería de H2O estas variables fueron usadas en formato alfanumérico dentro de una estructura de dataframe propia de H2O mientras que en el caso de TPOT fue necesario, como se menciono anteriormente, realizar una codificación a valores numéricos para lograr ser usados.

En cuanto a los modelos resultantes en cada una, TPOT al final de la ejecución según la configuración determinada entrega los parámetros y el modelo final que tuvo un mayor des-

[4]http://epistasislab.github.io/tpot/
[5]https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

empeño durante las validaciones cruzadas, por su parte H20 nos permite tener la opción de explorar los modelos que tuvieron un mayor desempeño y los clasifica según la métrica deseada.

En el primer proceso de experimentación, fueron utilizadas la totalidad de las características obtenidas en el conjunto de datos, esto con el fin de que los modelos generados por AutoML estén en la capacidad de identificar cuales son las variables que mayor incidencia tienen en el modelo sin ningún tipo de conocimiento previo.
El modelo final generado desde TPOT, que obtuvo el mejor desempeño después de los experimentos realizados fue un ExtraTreeClassifier, con los siguientes parámetros:

```
ExtraTreesClassifier(bootstrap=False, criterion="entropy",
    max_features=0.8500000000000001, min_samples_leaf=3,
    min_samples_split=7, n_estimators=200)
```

En un segundo proceso de experimentación, se evaluó el desempeño de los modelos generados automáticamente teniendo en cuenta el histórico de un año, por lo cual en este caso el conjunto de datos fue reducido a aquellas compañías que contaran con información de mínimo 4 trimestres, y se usaron únicamente las características que presentaron una mayor importancia en el experimento anterior, para esto fueron analizados los resultados de cada registro del conjunto de datos con LIME y se tomaron aquellas que durante toda la muestra sumaron una mayor importancia. Estas características fueron con las que el modelo encontró mayor relación al evaluar las calificaciones, por lo que pueden diferir de indicadores financiero que comúnmente son usado por las agencias calificadoras.
Estas características fueron:

- Total Assets
- Total Equity
- Net Income
- Net Income (Common)
- Income (Loss) from Continuing Operations
- Net Income/Starting Line
- Pretax Income (Loss)
- Total Liabilities

Como en este caso el conjunto de datos difiere del utilizado en el primer experimento, tanto en su longitud, como en las variables utilizadas, este ejercicio fue replicado con estas características y compañías seleccionadas únicamente con el valor de la fecha a evaluar, sin tener en cuenta históricos.

En este caso el AutoML generó un pipeline que consta de un randomForestClassifier junto con una regresión logística.

```
RandomForestClassifier(LogisticRegression(SelectPercentile(input_matrix,
percentile=84), C=0.0001, dual=False, penalty=l2), bootstrap=True,
criterion=gini, max_features=0.6500000000000001, min_samples_leaf=11,
min_samples_split=12, n_estimators=100)
```

Los resultados de este experimento son presentados en la sección 5.2.

4.4. Evaluación de modelos

Como se ha visto dentro de la literatura y los trabajos previos presentados anteriormente la principal forma en la que son evaluados este tipo de modelos es haciendo uso del acurracy, el cual nos da el porcentaje de predicciones acertadas respecto al número total de elementos, en general en problemas de clasificación este es un buen indicador del rendimiento del modelo, en este caso y teniendo en cuenta lo que se explica en [Golbayani et al., 2020a], la mayoría de las compañías mantienen sus calificaciones a través del tiempo, generando que un modelo que como predicción siempre usa el ultimo valor que tuvo en el entrenamiento pueda llegar a tener altos niveles de acurracy pero no ser muy bueno prediciendo los cambios de los ratings, los cuales son los que dan mayor valor e importancia a este tipo de modelo.

Para evitar esos sesgos dentro de la evaluación de este modelo generado por auto machine learning las compañías fueron divididas de forma aleatoria en conjuntos de entrenamiento (70) y pruebas (30), haciendo que ninguna de las muestras usadas para evaluar el desempeño del modelo hayan sido usadas previamente en el modelo, también se ha tenido en cuenta que este es un problema de clasificación ordinal ya que las clases de salida tienen un orden especifico, lo cual para este problema nos permite determinar también que tan alejada puede estar una predicción de su salida real, con esto se define la siguiente métrica de desempeño basada en lo propuesto en [Golbayani et al., 2020a]

$$F(i) = \frac{\sum_{k \in N} I(\hat{y}_k - y_k <= i)}{N} \tag{4-1}$$

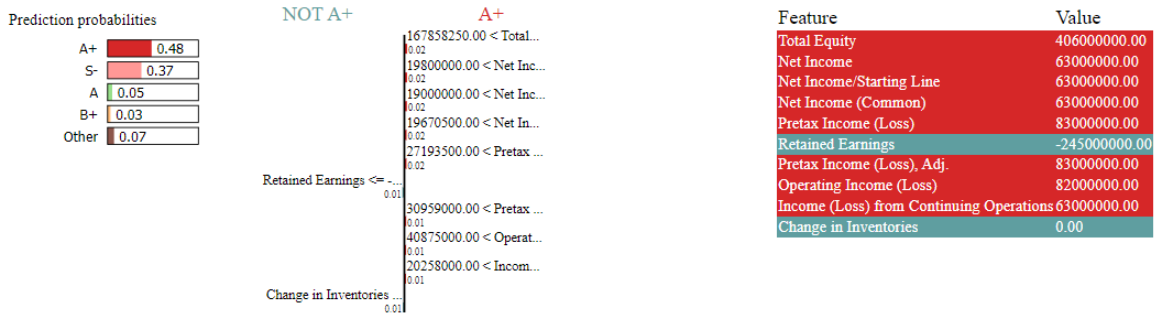


Figura 4-5.: Explicación generada por LIME

Para el uso de está métrica, se hace necesario asignar previamente valores numéricos a cada una de las clases. Una vez con esto es posible aplicar 4-1 donde:

y denota la calificación real de una observación.
$\hat{y}$ denota la predicción dada por un modelo particular.
N es el número total de observaciones en el conjunto de pruebas.
I es la función indicador que toma el valor de 1 siempre que la condición sea verdadera, es decir, cuando la diferencia entre el valor real y el valor generado por el modelo sea inferior a i.

Adicionalmente se presenta la evaluación por medio de la matriz de confusión resultante de las predicciones realizadas con las librerías de AutoML utilizadas en el desarrollo de este trabajo, estas nos permiten ver la cantidad y el porcentaje de calificaciones correctamente asignadas y también permite identificar de una forma visual que tan alejadas están las predicciones de el valor real de la muestra.

4.5. Interpretabilidad

Para la evaluación de interpretabilidad que nos genera el modelo, y entender cuales de las variables son las que tienen una mayor incidencia en la predicción de una calificación en especifico, se realizo una implementación de LIME, explicado más en detalle en la sección 3.4.4, dentro de los modelos generados por las dos herramientas de AutoML.

En este caso LIME toma la probabilidad de predicción de un registro en especifico usando el modelos generado por AutoML y nos muestra las 4 categorías con una mayor probabilidad y un listado de las 10 variables que más peso tuvieron e influyeron en la salida final para ese dato en particular.

En la figura **4-5** podemos ver la interpretación que realiza el modelo sobre uno de los re-

gistros en particular, como se puede apreciar en las probabilidades de predicción, para este caso los datos dados calificaban a la empresa en primer lugar como A+ con un 48 % de probabilidades, en segundo lugar S- con un 37 % de probabilidad y A con un 5 %, mostrando que el modelo esta tomando como segunda y tercera opción las calificaciones adyacentes a la calificación real A+.

Adicionalmente LIME nos indica cuales características influyeron más en la salida, cual es su valor para la muestra en particular y cual es el criterio que hace que pertenezca a esa clase, en el ejemplo presentado, podemos ver que la variable con mayor influencia es Total Equity, y que debido a que tiene un valor superior a 167,858,250.00 esto lo acerca más a la calificación A+, también podemos observar que Retained Earnings es un indicador que aleja este registro de A+, pero que con la ponderación de todas aquellas variables que aportan a A+ es allí donde es clasificado. Este ejemplo es tomado de los experimentos iniciales, en los cuales el modelo evaluaba todas las variables disponibles y por esto pueden presentarse variables similares como las utilidades netas, los cuales solo están representados en el conjunto final con Net Income.

Otro aspecto importante que nos da la explicación de una instancia en LIME, son los limites o thresholds en los cuales están delimitadas las variables para inclinarse más hacia una clase u otra, en este caso podemos observar que para el Total Equity se ha establecido un limite inferior para que la clasificación tenga una mayor probabilidad para ser A+, el valor limite inferior es cercano a 167 millones, mientras el valor de la instancia es superior a los 400 millones. Así mismo, cuando uno de los valores de las características están por fuera de los intervalos de la clase predicha, LIME nos indica sus limites. Esto nos permite tener un mayor entendimiento de las decisiones tomadas y que tan cercanos están los datos de pertenecer a una clase u otra.
Teniendo en cuenta esta interpretabilidad local, se realizó un resumen de las variables de mayor peso al realizar la interpretabilidad en todo el conjunto de pruebas, con esto se obtuvieron las 10 mejores características a nivel global para el experimento de TPOT.

Variables
IndustryID
Total assets
Total equity
Net Income
Net Income (Common)
Income (Loss) from continuing Operations
Net Income (Starting Line)
Pretax Income (Loss)
Total liabilites

En el caso de las implementaciones realizadas en H2O, la interpretabilidad dada para el modelo resultante es generada automáticamente en el proceso de AutoML, en este caso se dan los resultados de la importancia de las variables a nivel global, generando un nivel de importancia relativa, junto con la importancia escalada y el porcentaje en relación al total de variables usadas, en la siguiente tabla se detallan las variables con mayor importancia en el modelo final:

Variable	Importancia relativa	Importancia escalada	Porcentaje
Net Income	38522.75	1.00	0.160
Total Equity	29171.29	0.75	0.121
Cash, Cash Equivalents & Short Term Investments	10276.61	0.26	0.042
Total Liabilities	8437.20	0.21	0.03
Interest Expense, Net	6627.80	0.17	0.02
Revenue	6260.19	0.16	0.02
Share Capital & Additional Paid-In Capital	6175.54	0.16	0.02
Retained Earnings	5660.98	0.14	0.02
Total Current Liabilities	5533.98	0.14	0.02
Net Income/Starting Line	5410.01	0.14	0.02

5. Resultados

5.1. Experimento inicial

Para tener un punto de referencia en el rendimiento y evaluación del modelos resultantes a partir de las iteraciones realizadas en el AutoML, se realizo un modelo de regresión logística multinomial para realizar una clasificación inicial de las calificaciones del conjunto de datos usando las mismas variables de entrada que en el modelo final.

Esta regresión fue entrenada con un máximo de 10000 iteraciones, alcanzando un acurracy de 52 % y 51 % respectivamente para los conjuntos de entrenamiento y pruebas. A continuación se presentan las métricas detalladas generadas por el modelo y la matriz de confusión para el conjunto de prueba.

Calificación	Precision	recall	f1-score	support
Default	0.07	0.02	0.02	798
$B-$	0.28	0.14	0.19	1490
B	0.37	0.33	0.35	2696
B+	0.53	0.45	0.48	4856
$A-$	0.26	0.02	0.03	1106
A	0.51	0.73	0.60	10294
A+	0.56	0.56	0.56	12374
$S-$	0.37	0.21	0.27	2060
S	0.04	0.03	0.03	157
S+	0.00	0.00	0.00	1
acurracy			0.51	35832

Calificación	Precision	recall	f1-score	support
Default	0.02	0.00	0.01	266
$B-$	0.27	0.15	0.20	487
B	0.36	0.30	0.33	925
B+	0.51	0.44	0.47	1587
$A-$	0.19	0.01	0.02	379
A	0.52	0.72	0.60	3482
A+	0.56	0.56	0.56	4127
$S-$	0.35	0.23	0.28	636
S	0.06	0.04	0.04	55
S+	0.00	0.00	0.00	1
acurracy			0.50	11945

Como se aprecia en las tablas, el modelo base no tiene un rendimiento muy bueno ya que únicamente nos da un 50 % de resultados correctos para las predicciones exactas, mientras que teniendo en cuenta las distancias propuestas anteriormente, para las diferencias de un solo notch tenemos un 82 %.

Conjunto	Distancia(0)	Distancia(1)	Distancia(2)
Test	0.502767	0.828715	0.897614

En cuanto a los modelos generados por el AutoML, el correspondiente a la librería de TPOT encontró su mejor rendimiento con una configuración de 20 generaciones y poblaciones de tamaño 10 y realizando 5 cross validations, teniendo un acurracy de 91.53 % en el conjunto de entrenamiento después de 6 horas de entrenamiento y de un 74 % en el conjunto de prueba.

En la figura **5-1**, podemos apreciar la matriz de confusión para el conjunto de pruebas, indicando el número de predicciones exactas y también evidenciando la forma en la que esta distribuyendo los errores, para apoyar esta visualización y haciendo uso de la notch distance, las predicciones nos generan los siguientes resultados:

Conjunto	Distancia(0)	Distancia(1)	Distancia(2)
Test	0.742654	0.913855	0.943826

Con estos resultados, podemos determinar que el modelo nos da un nivel de confianza un poco más alto cuando se da un nivel de tolerancia de 1 notch sobre la clasificación real,

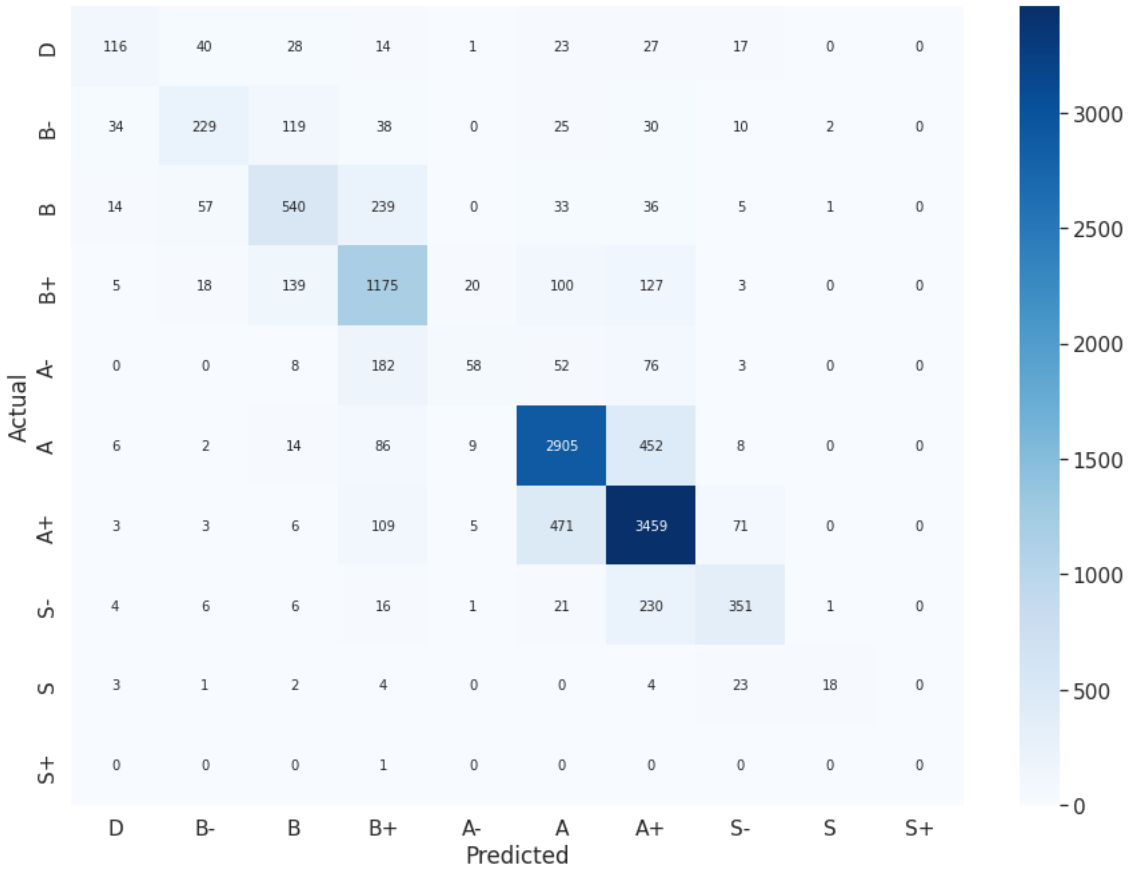


Figura 5-1.: Matriz de confusión TPOT conjunto de prueba

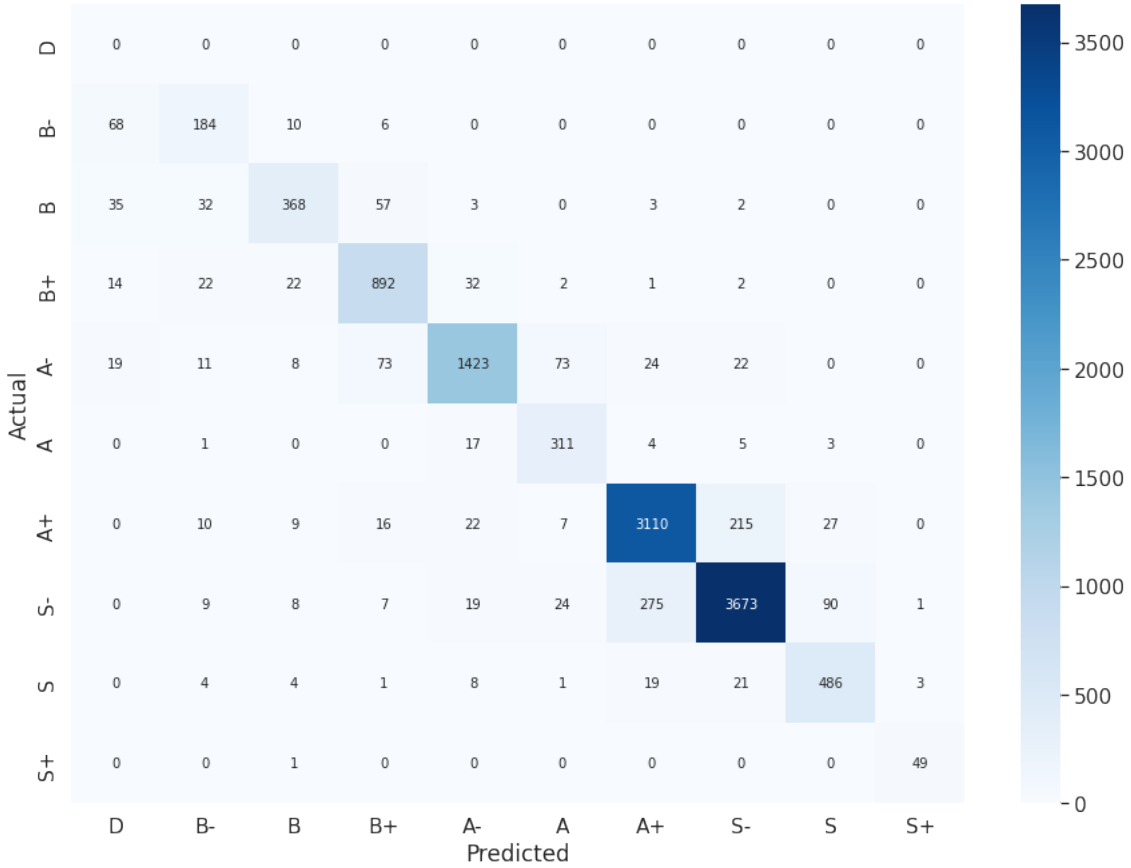


Figura 5-2.: Matriz de confusión H2O conjunto de prueba

alcanzando un 91 % lo cual nos da calificaciones bastantes coherentes con los valores reales y que si se tienen en cuenta el enfoque de modelo de calificaciones crediticias propio para una compañía financiera es un modelo con resultados que serían de gran ayuda para el análisis de crédito.

En cuanto a los resultados obtenidos en la generación automática de modelos de aprendizaje de máquina, haciendo uso de la librería H2O, la iteración que obtuvo el resultado con un mayor rendimiento tanto en el conjunto de entrenamiento como en el conjunto de pruebas fue haciendo por medio del modelo XGBoost descrito en la sección 4.3.

En este caso encontramos que el modelo generado tiene mejores resultados que los presentados por TPOT, llegando en este caso a un 88 % de accurracy en el conjunto de test y haciendo uso de la distancia notch se llega a un 96 % en la tolerancia de 1 notch y de 97 % para la tolerancia de 2 notchs.
A continuación, se presenta la matriz de confusión para este modelo en el conjunto de pruebas:

Conjunto	Distancia(0)	Distancia(1)	Distancia(2)
Test	0.8843	0.96503	0.97640

5.2. Experimento serie de tiempo

Al realizar la experimentación sobre los datos agrupados en forma temporal, como se menciono anteriormente, en donde se buscaba realizar la predicción basado en registros que contenían información histórica de al menos un año (cuatro trimestres de datos), tomando la totalidad de esta serie de tiempo como entrada del modelo y predecir de igual manera la predicción actual, se obtuvieron los siguientes resultados:

Conjunto	Accuray
Test	0.5826

Usando este mismo conjunto de datos, sin realizar la transformación temporal, es decir, teniendo únicamente la información del periodo de tiempo que se desea predecir, y con las mismas configuraciones iniciales dadas a TPOT para la series temporales, se obtuvieron los siguientes resultados:

Conjunto	Accuracy
Test	0.5776

Como vemos, la inclusión de datos históricos en conjunto de datos, genera un leve incremento en el desempeño final del modelo, en comparación con el AutoML generado para el mismo conjunto de datos sin registros históricos, esto debido a que permite reconocer la tendencia de la compañía en el periodo cercano, identificando posibles decaimientos o subidas en su información financiera que podrían indicar cambios en su nivel crediticio.

6. Contribución - AutoML con interpretabiliad como segundo objetivo

6.1. Interpretabilidad como segundo objetivo en AutoML

Teniendo en cuenta la experimentación anterior, los resultados obtenidos y metodología para calcular la interpretabilidad local en los modelos generados por TPOT, con este trabajo se busca contribuir en la implementación de una nueva versión de la librería TPOT en la cual durante su proceso de evaluación de individuos, realizada en medio del algoritmo genético principal con el que funciona la librería se añada como segundo objetivo la minimización del error en la interpretabilidad. Esta modificación se realizó en conjunto con integrantes del grupo de investigación.

Cómo primera etapa de esta modificación se identifico en la librería LIME, el proceso en el que se generan los puntos aleatorios cercanos al punto de referencia que se va a interpretar y con los cuales el algoritmo genera un modelo lineal altamente explicable para conocer la variación que tiene la clase de salida al realizar cambios sobre los valores de las características, aquí se debe resaltar que el modelo lineal es entrenado como una regresión Ridge, que toma como etiqueta la probabilidad de la clase que se esta prediciendo en el modelo a evaluar en el punto especifico.

Como se ha mencionado antes, LIME nos da un nivel de interpretabilidad local, para evaluar el nivel de interpretabilidad del modelo, se hace necesario pasar de esta interpretabiliadad local a una evaluación de interpretabilidad global. Para llegar a ello evaluamos la interpretabilidad local de cada una de las instancias del conjunto de datos, obteniendo los puntos vecinos de la evaluación de LIME y evaluándolos con el modelo a optimizar, estas predicciones nos permiten medir que tan alejado se encuentra el modelo lineal de LIME en esta instancia en especifico, al tener el nivel de desviación de cada instancia del conjunto de datos, podemos tener un nivel de interpretabilidad global del modelo, este proceso es realizado para cada uno de los individuos de la población por cada generación. Ya que al entrenar el modelo se tiene un conjunto completo de datos y en cada punto se evalúa su interpretabilidad, el error resultante para el modelo es calculado como la suma de la desviación del modelo local

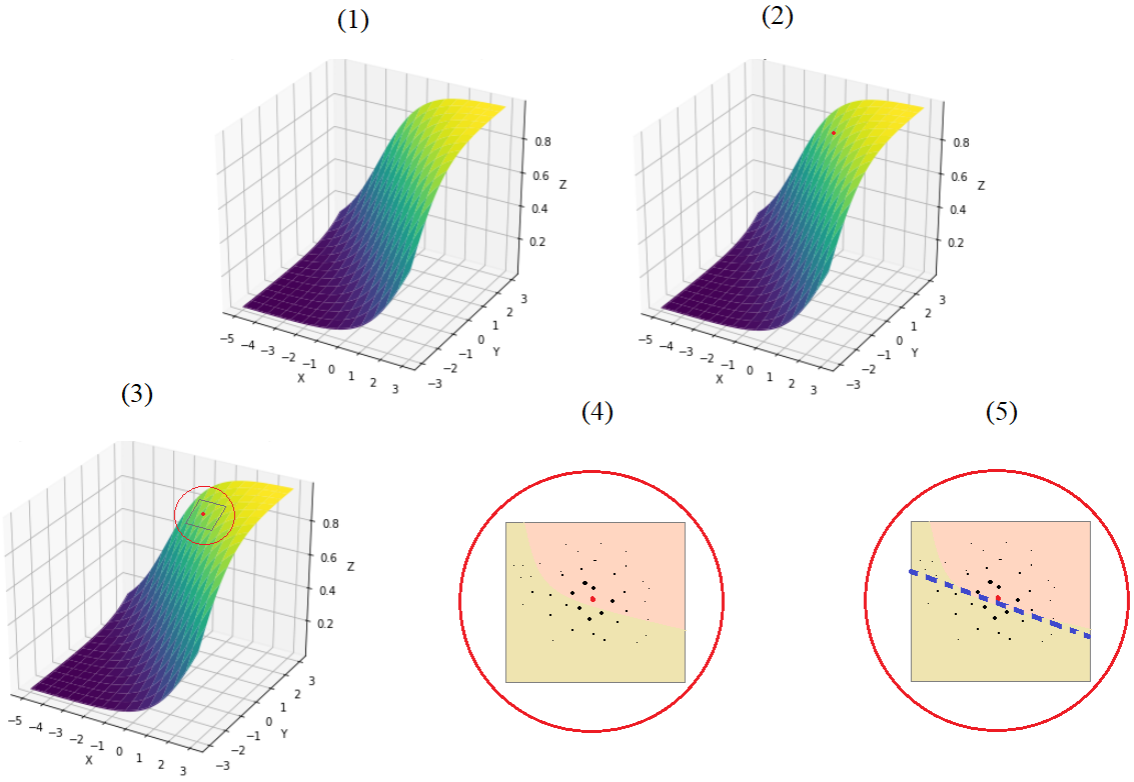


Figura 6-1.: Proceso de explicación local para una instancia

interpretable en cada uno de los registros del conjunto de datos.

En la figura **6-1** podemos detallar el proceso que se realiza en cada instancia de datos durante la optimización, en (1) vemos el conjunto de todos los datos representados por una superficie, posteriormente se itera punto a punto, en este caso en (2) vemos uno de los puntos seleccionados y sobre este se genera un conjunto de puntos vecinos al realizar modificaciones sobre el punto original, en (3) esto esta representado por el plano que contiene el punto seleccionado y sus variaciones. (4) Ya con los datos vecinos generados dentro del plano resultante, podemos observar como se generan las probabilidades de pertenecer o no a la clase predicha por el modelo en cada uno de los puntos vecinos. Sobre esto se genera un modelo interpretable local, que permite identificar como las variaciones de cada una de las características hacen que la clase predicha cambie o se mantenga(5). Finalmente se evalúan las predicciones dadas por el modelo interpretable contra las probabilidades de cada clase dadas por el modelo inicial. Estos errores se van acumulando al hacer este procedimiento en cada uno de los puntos del conjunto de datos, obteniendo el nivel de interpretabilidad propuesto.

Dentro de la configuración original de TPOT, se tenia como doble objetivo el desempeño del

modelo junto con la longitud de los pipelines generados, esto con el fin de buscar modelos menos complejos que solucionaran el problema, aquí se modifico esta configuración para que la evaluación de cada individuo se optimizara con un mayor desempeño como primer objetivo, y una menor desviación acumulada en la interpretabilidad calculada anteriormente como segundo objetivo.

Al completar el proceso de aprendizaje automático en nuestra librería modificada podemos extraer el conjunto total de modelos evaluados junto con sus métricas de desempeño e interpretabilidad calculadas, en donde podemos identificar el frente de pareto, o conjunto de soluciones óptimas que dan solución al problema, estos permiten validar no un único modelo final, si no un conjunto en el que se tienen modelo o muy interpretables o con muy buen rendimiento, permitiendo como punto final seleccionar el que tenga un mejor ajuste a los requerimientos del problema.

Esta modificación fue realizada únicamente para el problemas de clasificación y su utilización implica un alto consumo de recursos y tiempo, debido al proceso que se debe hacer en cada uno de los puntos y en cada uno de los modelos para obtener su interpretabilidad.

En la figura **6-2** podemos ver la evolución de la población en cada una de las generaciones de TPOT, aquí podemos evidenciar como en cada generación las métricas de desempeño (accuracy) y el nivel de interpretabilidad se optimizan, y como en cada generación hay un grupo de soluciones óptimas o frente de pareto. Vemos como en cada generación entrenada el conjunto de modelos esta menos dispersa y tienda a una agrupación alrededor de los mejores valores para la generación final.

6.2. Resultados AutoML con interpretabilidad

En cuanto a la experimentación de la modificación propuesta, al realizar la búsqueda de modelos por medio de TPOT modificado, para usar como segundo objetivo el nivel de interpretabilidad de cada modelos de la población, evaluando su error acumulado durante las interpretaciones locales realizadas a cada punto en el conjunto de datos de entrenamiento, tenemos el siguiente resultado:

Conjunto	Distancia(0)
Test	0.5733

Estos resultados obtenidos del mejor pipeline obtenido por TPOT, con su criterio de mayor acurracy, y que corresponde al siguiente modelo:

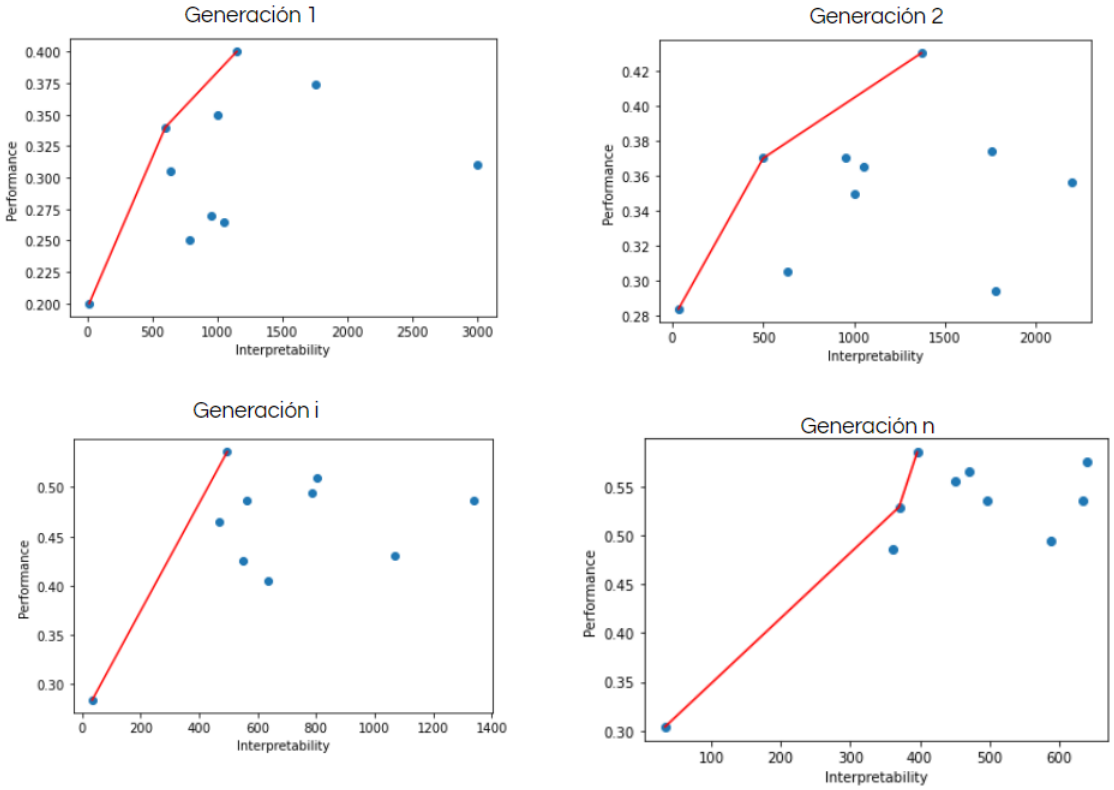


Figura 6-2.: Evolución de modelos por generación y frente de pareto generado al entrenar los modelos de TPOT con LIME

```
GradientBoostingClassifier(PolynomialFeatures(Normalizer(input_matrix,
norm=max), degree=2, include_bias=False, interaction_only=False),
learning_rate=0.01, max_depth=3, max_features=0.7500000000000001,
min_samples_leaf=12, min_samples_split=11, n_estimators=100, subsample=0.05)
```

Pero como se evidencia en la figura **6-2**, presentada anteriormente, podemos seleccionar alguno de los modelos que no sean dominados dentro del frente de pareto, con el fin de buscar el modelo que mayor se ajuste a la necesidad de la implementación, si se necesita un mayor nivel de interpretabiliad o un mayor nivel de acurracy, en este caso podemos ver que los mejores modelos tienen un nivel cercano de interpretabilidad, por lo que el tener un mayor nivel de acurracy permite seleccionar como el modelo ideal al número.

Modelo	Pipeline	Error de interpretabilidad	Acurracy
Modelo 1	MLPClassifier	-6.247100442009989e-29	0.34811129359523973
Modelo 2	GradientBoostingClassifier 1	-317.57535371352964	0.5171357359035399
Modelo 3	GradientBoostingClassifier 2	-360.13468100121446	0.5733106711136805

En el Anexo 1, se da detalle de cada uno de los modelos generados.

Al igual que en la experimentación con series de tiempo, se realizó la comparativa contra los resultados generados con el TPOT original, en este caso se obtuvo:

Conjunto	Accuracy
Test	0.5932

Con esto podemos ver que el nivel de acurracy, tienen una pequeña disminución al tener un modelo de mayor interpretabilidad, pero que representa un mejor entendimiento de como cada una de las características esta influyendo para cada instancia de datos.

7. Conclusiones y recomendaciones

7.1. Conclusiones

Se evidenció que la gran mayoría de trabajos previos mantienen similitudes en las variables que usan como entradas de modelo y que las variaciones que estas pueden llegar a tener se basan en las condiciones externas del contexto sobre el cual se esta desarrollando cada trabajo en especifico, como lo llegan a ser el tipo de agencia calificadora a predecir, el o los sectores específicos sobre los cuales se desea trabajar o incluso la situación particular del mercado que se deseaba evaluar dependiendo de su ubicación geográfica.

En cuanto a las dificultades presentadas que se evidenciaron en la literatura y que igualmente fueron destacadas durante la ejecución de este trabajo, podemos encontrar la disponibilidad de la información, primero en las variables de entrada, que si bien en su mayoría son de carácter público, el tener una única fuente que permita obtener la información de todo el universo puede llegar a tener algunos costos y trabajos posteriores en el procesamiento y limpieza de los datos, ya que para algunos periodos de tiempo la información de algunas compañías no se encontraba disponible en su totalidad. Y segundo en cuanto a la variable dependiente a predecir, las calificaciones históricas de cada compañía, la disponibilidad de estas esta sujeta a licenciamiento por parte de la agencia en especifico sobre la cual se deba trabajar, es por esto que se debió usar la alternativa de una calificación pública diferente a las principales agencias.

Adicionalmente, otro de los retos que se presentan es el imbalance en la cantidad de datos de ejemplo que se tienen de compañías que entran en default en comparación a la gran cantidad de compañías que se encuentran en los niveles más altos, generando que pueda existir un sesgo al momento de evaluar estos casos en los conjuntos de prueba.

Se identificaron en la literatura diferentes enfoques en cuanto a la selección de modelos de aprendizaje de máquina, se evidenció que en general se hace uso de diferentes tipos de arboles de decisión y diferentes configuraciones de redes neuronales llegando a resultados muy similares, por lo que este trabajo aporta la inclusión de esta forma de aprendizaje de máquina automático con la cual se ha podido explorar la combinación de varios de estos tipos de modelos y la exploración automática de los diferentes parámetros, llegando así a resultados similares algunos de los trabajos estudiados utilizando únicamente estas iteraciones

automáticas en las cuales se usaron gran parte de las variables de entrada obtenidas permitiendo que los propios algoritmos de AutoML fueran los que tomaran las características de mayor relevancia a la calificación final.

Al incluirse la interpretabilidad como segundo objetivo, se logra obtener modelos finales con un nivel muy cercano de precisión a los generados por el AutoML sin optimizar la interpretabilidad, esto significando que no hay un deterioro significativo del acurracy, y si una ganancia en cuanto a la interpretabilidad final que se realiza sobre el modelo seleccionado según las necesidades propias del problema. También, en consecuencia de la modificación realizada y su nivel de detalle en cada iteración este enfoque genera un mayor consumo de recursos de computo y tiempo al realizar la generación del modelo.

Se logra realizar una implementación exitosa de los algoritmos de AutoML que permitieron llegar a modelos con buenos nivel de precisión y que se pueden adaptar fácilmente a diferentes configuraciones, como limitar el conjunto de datos a un sector especifico, usar combinaciones de diferentes calificaciones como variable objetivo o cambiar los periodos de tiempo que se usaron para el entrenamiento de este trabajo sin la necesidad de ajustar diferentes parámetros ya que el mismo proceso de AutoML los evalúa mientras busca el modelos que mejor se ajuste a estos cambios.

Finalmente, la evaluación de los resultados permitió establecer que si bien el 74 % de las predicciones son totalmente acertadas, el modelo tiene una precisión de 91 % cuando se tienen en cuenta las predicciones que están alejadas únicamente un nivel dentro de la escala de calificaciones. Lo cual entra en un nivel bueno de acierto aunque se esperaría poder conseguir mejores resultados usando estas mismas técnicas en trabajos futuros.

7.2. Recomendaciones

Se recomienda para futuras implementaciones de modelos que aborden este problema la exploración de conjuntos de datos o compañías que aporten un mayor número de muestras en los casos de Default, esto con el fin de generar una base de entrenamiento que genere más fortaleza en la predicción de las calificaciones inferiores ya que estas son las que pueden agregar mayor valor a los resultados obtenidos.

En cuanto a la selección de las variables se recomienda tener parte de conocimiento básico en la metodología a querer predecir, ya que esto es de gran utilidad al momento de seleccionar gran parte de las características que serán parte del modelo final, junto a esto, también se puede proponer la división del modelo general en la generación de sub modelos de este mismo tipo que se dediquen específicamente a diferentes sectores económicos, esto ya que se

han encontrado varias diferencias en las características de las compañías que pertenecen a diferentes sectores que pueden llegar a generar un poco de ruido cuando son evaluadas todas en conjunto.

Al realizar la medición de la desviación entre el modelo local de LIME frente al modelo a optimizar, se pueden buscar alternativas diferentes a Ridge como modelo interpretable para evaluar el comportamiento de los resultados, esto no fue incluido dentro del alcance del aporte propuesto en este trabajo de grado debido a requiere de un modificación mayor dentro de la librería original de LIME.
Adicionalmente, se recomienda el uso de nuevas herramientas de AutoML de uso licenciado que ofrecen una mayor capacidad de computo y tiempos de ejecución logrando llegar a una mayor profundidad en la búsqueda de características, modelos y hiperparametros y pueden llegar a ofrecer un despliegue de estos modelos en aplicaciones o visualizaciones que llegan a ser mucho más claros y de fácil uso para los analistas y usuarios finales del modelo.[Gilpin et al., 2019]

A. Resultados revisión de literatura

Reference	Year	Num. of Input Var.	Input Data	Credit Rating Provider	Accuracy
(Shin and Han, 2001)	2001	12	Korean Companies	Korean Ratings	69.3%
(Koh et al., 2004)	2004				74.2%
(Ye et al., 2008)	2008	33	Manufacturing/Wholesale/Retail/Services	Moody's	47.1%
(Hájek and Olej, 2014)	2014	20	Investment and non-investment firms	S&P	86.13%
(Veronezi, 2016)	2016	33	Healthcare, Financial, Technology Sectors in USA	S&P and Moody's	$> 70\%$
(Addo et al., 2018)	2018	181 and 10	117019 companies asking for a loan from a bank	European Bank	RMSE: 0.044
(Wallis et al., 2019)	2019	27	308 companies from S&P 500	Moody's	64.6%
(Moscatelli et al., 2019)	2019	26	Italian non-financial firms	Italian Credit Register	$> 84\%$

Figura A-1.: Comparación estudios basados en árboles de decisión.

Reference	Year	Num. of Input Var.	Input Data	Credit Rating Provider	Accuracy
(Huang et al., 2004)	2004	14	36 Commercial Banks	S&P	80.00%
(Huang et al., 2004)	2004	6	Taiwan Dataset	Taiwan Ratings Corporation	79.73%
(Cao et al., 2006)	2006	17	US Manufactoring Companies	S&P	84.61%
(Lee, 2007)	2007	10	3017 Korean Companies	Korea Information Service, Inc	67.22%
(Ye et al., 2008)	2008	33	Four US industries	Moody's	64%
(Yu et al., 2008)	2008	13	Jap. Consumer Credit Card	Application Approvals	79.91%
(Yu et al., 2008)	2008	12	60 UK Corporations	UK Ratings	77.84%
(Kim and Sohn, 2010)	2010	10	Korean SMEs	Korean Default Rates	66.16%
(Hájek and Olej, 2011)	2011	5	852 US firms	S&P	85.96%
(Hájek and Olej, 2011)	2011	12	452 Czech municipals	local experts	89.76%
(Kim and Ahn, 2012)	2012	14	Korea Dataset	Korea Ratings Corporation	67.98%
(Yeh et al., 2012)	2012	18	2470 Taiwanese companies	Taiwanese Rating	74.4%
(Hájek and Olej, 2014)	2014	20	Investment and non-investment grade	S&P	87.28%
(Pai et al., 2015)	2015	18	Taiwan Dataset	Taiwan Ratings Corporation	86%
(Wallis et al., 2019)	2019	27	308 companies from S&P 500	Moody's	60.1%

Figura A-2.: Comparación estudios basados en máquinas de soporte vectorial.

Reference	Year	Num. of Input Var.	Input Data	Credit Rating Provider	Accuracy
(Garavaglia, 1991)	1991	87	797 Companies	S&P	85%
(Moody and Utans, 1994)	1994	10	196 Indudtrial Firms	S&P	36%
(Huang et al., 2004)	2004	5	36 Commercial Banks	S&P	80.00%
(Huang et al., 2004)	2004	6 & 16	Taiwan Dataset	Taiwan Ratings Corporation	75.68%
(Kim, 2005)	2005		1080 US companies	S&P	84%
(Brabazon and O'Neill, 2006)	2006	8	791 Non-financial US Firms	S&P	85%
(Cao et al., 2006)	2006	17	US Manufactoring Companies	S&P	80.28%
(Lee, 2007)	2007	10	3017 Korean Companies	Korea Information Service, Inc	59.93%
(Yu et al., 2008)	2008	13	Jap. Consumer Credit Card	Application Approvals	88.08%
(Yu et al., 2008)	2008	12	60 UK Corporations	UK Ratings	85.87%
(Khashman, 2010)	2010	24	German Dataset	German Ratings	85.9%
(Kim and Sohn, 2010)	2010	10	Korean SMEs	Korean Default Rates	64.23%
(Hájek, 2012)	2012	52	852 US Mining Companies	S&P	80%
(Hájek, 2012)	2012	14	154 US Municipalities	Moody's	71.62%
(West, 2000)	2000	24	German and Australian Data	German and Australian Rating	error:0.2243
(Khemakhem and Boujelbene, 2015)	2015	15	86 Tunisian companies	Tunisian credit rating	82.55%
(Zhao et al., 2015)	2015	20	German Credit Dataset	German Credit Ratings	87%
(Hájek and Olej, 2014)	2014	20	Investment and non-investment grade	S&P	88.44%
(Addo et al., 2018)	2018	181 and 10	117019 companies asking for a loan	European Bank	RMSE: 0.044
(Daniel et al., 2019)	2019	17	Random US companies	S&P and Moody's	83%
(Wallis et al., 2019)	2019	27	308 companies from S&P 500	Moody's	63.6%

Figura A-3.: Comparación estudios basados en redes neuronales artificiales.

B. Anexo: Detalle de modelos generados por TPOT + LIME

```
model 1: MLPClassifier(FastICA(FastICA(FastICA(input_matrix,
FastICA__tol=0.6000000000000001), FastICA__tol=0.6000000000000001),
FastICA__tol=0.05), MLPClassifier__alpha=0.1,
MLPClassifier__learning_rate_init=0.1)

model 2: GradientBoostingClassifier(PolynomialFeatures(PCA(Normalizer
(input_matrix, Normalizer__norm=max), PCA__iterated_power=9,
PCA__svd_solver=randomized), PolynomialFeatures__degree=2,
PolynomialFeatures__include_bias=False,
PolynomialFeatures__interaction_only=False),
GradientBoostingClassifier__learning_rate=0.01,
GradientBoostingClassifier__max_depth=3,
GradientBoostingClassifier__max_features=0.7500000000000001,
GradientBoostingClassifier__min_samples_leaf=12,
GradientBoostingClassifier__min_samples_split=11,
GradientBoostingClassifier__n_estimators=100,
GradientBoostingClassifier__subsample=0.05)

model 3: GradientBoostingClassifier(PolynomialFeatures(Normalizer
(input_matrix, Normalizer__norm=max), PolynomialFeatures__degree=2,
PolynomialFeatures__include_bias=False,
PolynomialFeatures__interaction_only=False),
GradientBoostingClassifier__learning_rate=0.01,
GradientBoostingClassifier__max_depth=3,
GradientBoostingClassifier__max_features=0.7500000000000001,
GradientBoostingClassifier__min_samples_leaf=12,
GradientBoostingClassifier__min_samples_split=11,
GradientBoostingClassifier__n_estimators=100,
GradientBoostingClassifier__subsample=0.05)
```

Bibliografía

[Abdou et al., 2017] Abdou, H. A., Abdallah, W. M., Mulkeen, J., Ntim, C. G., and Wang, Y. (2017). Prediction of financial strength ratings using machine learning and conventional techniques. *Investment Management and Financial Innovations*, 14:194–211.

[Abedin et al., 2018] Abedin, M. Z., Guotai, C., Colombage, S., and Fahmida-E-Moula (2018). Credit default prediction using a support vector machine and a probabilistic neural network. *Journal of Credit Risk*, 14:1–27. Cited By :3 Export Date: 26 September 2019.

[AghaeiRad et al., 2017] AghaeiRad, A., Chen, N., and Ribeiro, B. (2017). Improve credit scoring using transfer of learned knowledge from self-organizing map. *Neural Computing and Applications*, 28:1329–1342. Cited By :4¡br/¿Export Date: 26 September 2019.

[Alexandros, 2017] Alexandros, G. (2017). " forecasting models using machine learning (ml) techniques on banks ' credit rating .".

[Alonso and Carbo, 2020] Alonso, A. and Carbo, J. M. (2020). Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost. *SSRN Electronic Journal*.

[Alvim et al., 2010] Alvim, L. G. M., Santos, C. N. D., and Milidiú, R. L. (2010). Daily volume forecasting using high frequency predictors. pages 248–254. Cited By :5¡br/¿Export Date: 26 September 2019.

[Attigeri et al., 2017] Attigeri, G. V., Pai, M. M. M., and Pai, R. M. (2017). Credit risk assessment using machine learning algorithms. *Advanced Science Letters*, 23:3649–3653. Cited By :3¡br/¿Export Date: 26 September 2019.

[Ayouche et al., 2017] Ayouche, S., Aboulaich, R., and Ellaia, R. (2017). Partnership credit scoring classification probem: A neural network approach. *International Journal of Applied Engineering Research*, 12:693–704. Export Date: 26 September 2019.

[Baltaev and Chavdarov, 2014] Baltaev, A. and Chavdarov, I. (2014). Predicting corporate defaults: Evaluating moody ' s credit rating institute.

[Bandyopadhyay, 2006] Bandyopadhyay, A. (2006). Predicting probability of default of indian corporate bonds: logistic and z-score model approaches. *Journal of Risk Finance*, 7:255–272.

[Bao et al., 2019] Bao, W., Lianju, N., and Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128:301–315. Cited By :2¡br/¿Export Date: 26 September 2019.

[Bhaduri, 2009] Bhaduri, A. (2009). Credit scoring using artificial immune system algorithms: A comparative study. pages 1540–1543. Cited By :1 Export Date: 26 September 2019.

[Bhattacharjee et al., 2017] Bhattacharjee, B., Sridhar, A., and Shafi, M. (2017). An artificial neural network-based ensemble model for credit risk assessment amp; deployment as a graphical user interface. *International Journal of Data Mining, Modelling and Management*, 9:122–141. Cited By :3 Export Date: 26 September 2019.

[Birla et al., 2016] Birla, S., Kohli, K., and Dutta, A. (2016). Machine learning on imbalanced data in credit risk. Cited By :3¡br/¿Export Date: 26 September 2019.

[Bussmann et al., 2021] Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57:203–216.

[Carvalho et al., 2019] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics 2019, Vol. 8, Page 832*, 8:832.

[Chauhan et al., 2020] Chauhan, K., Jani, S., Thakkar, D., Dave, R., Bhatia, J., Tanwar, S., and Obaidat, M. S. (2020). Automated machine learning: The new wave of machine learning. *2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings*, pages 205–212.

[Chen et al., 2011] Chen, Q., Xue, H.-F., and Yan, L. (2011). Credit risk assessment based on potential support vector machine. volume 1, pages 97–101. Cited By :1 Export Date: 26 September 2019.

[Chen et al., 2009] Chen, W., Ma, C., and Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert Systems with Applications*, 36:7611–7616. Cited By :63 Export Date: 26 September 2019.

[Chi et al., 2019] Chi, G., Uddin, M. S., Abedin, M. Z., and Yuan, K. (2019). Hybrid model for credit risk prediction: An application of neural network approaches. *International Journal on Artificial Intelligence Tools*, 28. Export Date: 26 September 2019.

[Chuang and Huang, 2011] Chuang, C.-L. and Huang, S.-T. (2011). A hybrid neural network approach for credit scoring. *Expert Systems*, 28:185–196. Cited By :22¡br/¿Export Date: 26 September 2019.

[Cimpoeru, 2011] Cimpoeru, S. S. (2011). Neural networks and their application in credit risk assessment. evidence from the romanian market. *Technological and Economic Development of Economy*, 17:519–534. Cited By :9 Export Date: 26 September 2019.

[Correa and Gonzalez, 2011] Correa, A. B. and Gonzalez, A. M. (2011). Evolutionary algorithms for selecting the architecture of a mlp neural network: A credit scoring case. pages 725–732. Cited By :9 Export Date: 26 September 2019.

[Danenas and Garsva, 2009] Danenas, P. and Garsva, G. (2009). Support vector machines and their application in credit risk evaluation process. *Transformations in Business and Economics*, 8:46–58. Cited By :18 Export Date: 26 September 2019.

[Danenas and Garsva, 2012] Danenas, P. and Garsva, G. (2012). Pso-based linear svm classifier selection for credit risk evaluation modeling process. volume 1 DISI, pages 338–341. Cited By :1 Export Date: 26 September 2019.

[Danenas et al., 2011] Danenas, P., Garsva, G., and Gudas, S. (2011). Credit risk evaluation model development using support vector based classifiers. volume 4, pages 1699–1707. Cited By :25 Export Date: 26 September 2019.

[Dima and Vasilache, 2009] Dima, A. M. and Vasilache, S. (2009). Ann model for corporate credit risk assessment. pages 94–98. Cited By :4 Export Date: 26 September 2019.

[Du, 2003] Du, Y. (2003). Predicting credit rating and credit rating changes : A new approach. *Working Paper*.

[Elshawi et al., 2019] Elshawi, R., Maher, M., and Sakr, S. (2019). Automated machine learning: State-of-the-art and open challenges.

[Fang and Huang, 2011] Fang, K. and Huang, H. (2011). Variable selection for credit risk model using data mining technique. *Journal of Computers*, 6:1868–1874. Cited By :3¡br/¿Export Date: 26 September 2019.

[Fenerich et al., 2018] Fenerich, A. T., Steiner, M. T. A., Neto, P. J. S., Tochetto, E. A., Tsutsumi, D. P., Assef, F. M., and Santos, B. S. D. (2018). Use of machine learning techniques in bank credit risk analysis. volume 2018-December. Export Date: 26 September 2019.

[Fitch, 2021] Fitch (2021). Rating definitions.

[Gilpin et al., 2019] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pages 80–89.

[Golbayani et al., 2020a] Golbayani, P., Florescu, I., and Chatterjee, R. (2020a). A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *arXiv*.

[Golbayani et al., 2020b] Golbayani, P., Wang, D., and Florescu, I. (2020b). Application of deep neural networks to assess corporate credit rating. pages 1–18.

[Groth and Muntermann, 2011] Groth, S. S. and Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50:680–691. Cited By :71 Export Date: 26 September 2019.

[Guo et al., 2012] Guo, X., Zhu, Z., and Shi, J. (2012). A corporate credit rating model using support vector domain combined with fuzzy clustering algorithm. *Mathematical Problems in Engineering*, 2012.

[Hajek, 2010] Hajek, P. (2010). *Credit Rating Modelling by Neural Networks*. Cited By :3 Export Date: 26 September 2019.

[Hajek et al., 2014] Hajek, P., Olej, V., and Myskova, R. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, 20:721–738.

[Hajek et al., 2017] Hajek, P., Olej, V., and Prochazka, O. (2017). Predicting corporate credit ratings using content analysis of annual reports –a naïve bayesian network approach. *Lecture Notes in Business Information Processing*, 276:47–61.

[Hervé and Erwan, 2020] Hervé, P. and Erwan, R. (2020). Explain artificial intelligence for credit risk management.

[Hilscher et al., 2010] Hilscher, J., Bruche, M., Campbell, J., Cecchetti, S., Jarrow, R., Lebaron, B., Ramadorai, T., Shleifer, A., Singhal, M., Stein, J., Szilagyi, J., and Verdelhan, A. (2010). Credit ratings and credit risk.

[Hsieh and Hung, 2010] Hsieh, N.-C. and Hung, L.-P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*, 37:534–545. Cited By :85 Export Date: 26 September 2019.

[Hu, 2009] Hu, J. (2009). Personal credit rating using artificial intelligence technology for the national student loans. pages 103–106. Cited By :2¡br/¿Export Date: 26 September 2019.

[Huang, 2012] Huang, C.-F. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing Journal*, 12:807–818. Cited By :107 Export Date: 26 September 2019.

[Huang et al., 2020] Huang, J., Chai, J., and Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14.

[Huang, 2009] Huang, S.-C. (2009). Integrating nonlinear graph based dimensionality reduction schemes with svms for credit rating forecasting. *Expert Systems with Applications*, 36:7515–7518. Cited By :19¡br/¿Export Date: 26 September 2019.

[Huang et al., 2010] Huang, W., Zhu, X., and Su, Q. (2010). Research on application of personal credit scoring based on bp-logistic hybrid algorithm. volume 4, pages V4735–V4739. Export Date: 26 September 2019.

[Huang et al., 2004] Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, 37:543–558.

[Hájek and Olej, 2014] Hájek, P. and Olej, V. (2014). Predicting firms' credit ratings using ensembles of artificial immune systems and machine learning – an over-sampling approach. *IFIP Advances in Information and Communication Technology*, 436:29–38.

[Jagric et al., 2011] Jagric, V., Kracun, D., and Jagric, T. (2011). Does non-linearity matter in retail credit risk modeling? *Finance a Uver - Czech Journal of Economics and Finance*, 61:384–402. Cited By :7 Export Date: 26 September 2019.

[Khashman, 2009] Khashman, A. (2009). A neural network model for credit risk evaluation. *International Journal of Neural Systems*, 19:285–294. Cited By :39¡br/¿Export Date: 26 September 2019.

[Khashman, 2010] Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37:6233–6239. Cited By :104¡br/¿Export Date: 26 September 2019.

[Khashman, 2011] Khashman, A. (2011). Credit risk evaluation using neural networks: Emotional versus conventional models. *Applied Soft Computing Journal*, 11:5477–5484. Cited By :49¡br/¿Export Date: 26 September 2019.

[Khemakhem and Boujelbene, 2018] Khemakhem, S. and Boujelbene, Y. (2018). Predicting credit risk on the basis of financial and non-financial variables and data mining. *Review of Accounting and Finance*, 17:316–340. Cited By :1 Export Date: 26 September 2019.

[Kim and Ahn, 2012] Kim, K.-J. and Ahn, H. (2012). A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers and Operations Research*, 39:1800–1811. Cited By :57 Export Date: 26 September 2019.

[Kotsiantis et al., 2009] Kotsiantis, S. B., Kanellopoulos, D., Karioti, V., and Tampakas, V. (2009). An ontology-based portal for credit risk analysis. pages 165–169. Cited By :1 Export Date: 26 September 2019.

[Kulkarni and Dhage, 2019] Kulkarni, S. V. and Dhage, S. N. (2019). Advanced credit score calculation using social media and machine learning. *Journal of Intelligent and Fuzzy Systems*, 36:2373–2380. Cited By :1¡br/¿Export Date: 26 September 2019.

[Kvamme et al., 2018] Kvamme, H., Sellereite, N., Aas, K., and Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207–217. Cited By :12¡br/¿Export Date: 26 September 2019.

[Ledell and Poirier, 2020] Ledell, E. and Poirier, S. (2020). H2o automl: Scalable automatic machine learning.

[Lei et al., 2019] Lei, K., Xie, Y., Zhong, S., Dai, J., Yang, M., and Shen, Y. (2019). Generative adversarial fusion network for class imbalance credit scoring. *Neural Computing and Applications*. Export Date: 26 September 2019.

[Leo et al., 2019] Leo, M., Sharma, S., and Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, 7. Export Date: 26 September 2019.

[Letizia and Lillo, 2019] Letizia, E. and Lillo, F. (2019). Corporate payments networks and credit risk rating. *EPJ Data Science*, 8. Export Date: 26 September 2019.

[Li et al., 2017] Li, Z., Tian, Y., Li, K., Zhou, F., and Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, 74:105–114. Cited By :21 Export Date: 26 September 2019.

[Lin and Zhang, 2012] Lin, Y. and Zhang, Y. (2012). Credit risk assessment based on neural network. pages 402–404. Cited By :2¡br/¿Export Date: 26 September 2019.

[Linardatos et al., 2020] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy 2021, Vol. 23, Page 18*, 23:18.

[Lisboa et al., 2009] Lisboa, P. J. G., Etchells, T. A., Jarman, I. H., Arsene, C. T. C., Aung, M. S. H., Eleuteri, A., Taktak, A. F. G., Ambrogi, F., Boracchi, P., and Biganzoli, E.

(2009). Partial logistic artificial neural network for competing risks regularized with automatic relevance determination. *IEEE Transactions on Neural Networks*, 20:1403–1416. Cited By :30¡br/¿Export Date: 26 September 2019.

[Liu et al., 2010] Liu, X., Fu, H., and Lin, W. (2010). A modified support vector machine model for credit scoring. *International Journal of Computational Intelligence Systems*, 3:797–804. Cited By :3 Export Date: 26 September 2019.

[Liu et al., 2009] Liu, Y., Cai, Q., Luo, Y., Qian, J., and Ye, F. (2009). Artificial neural networks for corporation credit rating analysis. volume 1, pages 81–84. Cited By :3¡br/¿Export Date: 26 September 2019.

[Lu and Cao, 2009] Lu, H.-K. and Cao, J. (2009). Personal credit scoring model based on integration of rough set and ga-neural network. *Nanjing Li Gong Daxue Xuebao/Journal of Nanjing University of Science and Technology*, 33:194–198. Cited By :1¡br/¿Export Date: 26 September 2019.

[Luo et al., 2017] Luo, C., Wu, D., and Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65:465–470. Cited By :19 Export Date: 26 September 2019.

[Malik and Hermawan, 2018] Malik, R. F. and Hermawan (2018). Credit scoring using cart algorithm and binary particle swarm optimization. *International Journal of Electrical and Computer Engineering*, 8:5425–5431. Export Date: 26 September 2019.

[Mall, 2018] Mall, S. (2018). An empirical study on credit risk management: The case of nonbanking financial companies. *Journal of Credit Risk*, 14:49–66. Export Date: 26 September 2019.

[Metz and Cantor, 2006] Metz, A. and Cantor, R. (2006). Moody's credit rating prediction model. *Moody's Investors Service*, pages 1–20.

[Molnar, 2022] Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

[Moonasar, 2007] Moonasar, V. (2007). Credit risk analysis using credit risk analysis using artificial intelligence : Evidence from a leading south african banking institution. *Methodology*.

[Moscatelli et al., 2020] Moscatelli, M., Parlapiano, F., Narizzano, S., and Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161.

[Nanni and Lumini, 2009] Nanni, L. and Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 36:3028–3033. Cited By :118¡br/¿Export Date: 26 September 2019.

[Olson et al., 2016] Olson, R. S., Bartley, N., Urbanowicz, R. J., and Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. *GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference*, pages 485–492.

[Paper et al., 2017] Paper, F. I. E. T., Carlenius, B., Døvik, E. K., Kolberg, J. K., and Aanes, B. (2017). Corporate credit rating using deep learning with genetic algorithms norwegian school of economics. pages 1–68.

[Patron et al., 2019] Patron, G., Leon, D., Lopez, E., and Hernandez, G. (2019). An interpretable automated machine learning credit risk model.

[Pazhoheshfara et al., 2010] Pazhoheshfara, P., Azadeh, A., and Saberia, M. (2010). Improving accuracy of artificial neural networks for credit scoring models using voting algorithm. Cited By :1 Export Date: 26 September 2019.

[Petropoulos et al., 2018] Petropoulos, A., Siakoulis, V., Stavroulakis, E., and Klamargias, A. (2018). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *The use of big data analytics and artificial intelligence in central banking*, 50:30–31.

[Ping, 2009] Ping, Y. (2009). Hybrid classifier using neighborhood rough set and svm for credit scoring. pages 138–142. Cited By :21 Export Date: 26 September 2019.

[Ping and Yongheng, 2011] Ping, Y. and Yongheng, L. (2011). Neighborhood rough set and svm based hybrid credit scoring classifier. *Expert Systems with Applications*, 38:11300–11304. Cited By :40¡br/¿Export Date: 26 September 2019.

[Provenzano et al., 2020] Provenzano, A. R., Trifirò, D., Jean, N., Pera, G. L., Spadaccino, M., Massaron, L., and Nordio, C. (2020). An artificial intelligence approach to shadow rating. *SSRN Electronic Journal*, pages 1–18.

[Rahayu et al., 2010] Rahayu, S. P., Zain, J. M., Embong, A., and Purnami, S. W. (2010). Credit risk classification using kernel logistic regression with optimal parameter. pages 602–605. Export Date: 26 September 2019.

[Ribeiro and Chen, 2017] Ribeiro, B. and Chen, N. (2017). Financial credit risk assessment via learning-based hashing. *Intelligent Decision Technologies*, 11:177–186. Cited By :1¡br/¿Export Date: 26 September 2019.

[Rom, 2009] Rom, M. C. (2009). The credit rating agencies and the subprime mess: Greedy, ignorant, and stressed?

[Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.

[Ryser and Denzler, 2009] Ryser, M. and Denzler, S. (2009). Selecting credit rating models: A cross-validation-based comparison of discriminatory power. *Financial Markets and Portfolio Management*, 23:187–203. Cited By :6 Export Date: 26 September 2019.

[Saitoh, 2016] Saitoh, F. (2016). Predictive modeling of corporate credit ratings using a semi-supervised random forest regression. volume 2016-December, pages 429–433. Cited By :1 Export Date: 26 September 2019.

[Salehi and Mansoury, 2011] Salehi, M. and Mansoury, A. (2011). An evaluation of iranian banking system credit risk: Neural network and logistic regression approach. *International Journal of Physical Sciences*, 6:6082–6090. Cited By :3 Export Date: 26 September 2019.

[Santos, 2007] Santos, K. (2007). Corporate credit ratings: a quick guide. 44:45–49.

[Selau and Ribeiro, 2011] Selau, L. P. R. and Ribeiro, J. L. D. (2011). A systematic approach to construct credit risk forecast models. *Pesquisa Operacional*, 31:41–56. Cited By :1 Export Date: 26 September 2019.

[Sitt and Wu, 2011] Sitt, M. and Wu, T. (2011). Evaluation of credit risk. pages 1–5.

[SP, 2021] SP (2021). Guide to credit rating essentials what are credit ratings and how do they work?

[Studer et al., 2021] Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., and Müller, K.-R. (2021). Towards crisp-ml(q): A machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction 2021, Vol. 3, Pages 392-413*, 3:392–413.

[Truong et al., 2019] Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, B., and Farivar, R. (2019). Towards automated machine learning: Evaluation and comparison of automl approaches and tools.

[Tsai and Chen, 2010] Tsai, C.-F. and Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. *Applied Soft Computing Journal*, 10:374–380. Cited By :56 Export Date: 26 September 2019.

[Turkson et al., 2016] Turkson, R. E., Baagyere, E. Y., and Wenya, G. E. (2016). A machine learning approach for predicting bank credit worthiness. pages 81–87. Cited By :6¡br/¿Export Date: 26 September 2019.

[Twala, 2009] Twala, B. (2009). Combining classifiers for credit risk prediction. *Journal of Systems Science and Systems Engineering*, 18:292–311. Cited By :6 Export Date: 26 September 2019.

[Vahid and Ahmadi, 2016] Vahid, P. R. and Ahmadi, A. (2016). Modeling corporate customers' credit risk considering the ensemble approaches in multiclass classification: Evidence from iranian corporate credits. *Journal of Credit Risk*, 12:71–95. Cited By :1 Export Date: 26 September 2019.

[Vidovic and Yue, 2020] Vidovic, L. and Yue, L. (2020). Machine learning and credit risk modelling machine learning in finance.

[Wang et al., 2012] Wang, J., Hedar, A.-R., Wang, S., and Ma, J. (2012). Rough set and scatter search metaheuristic based feature selection for credit scoring. *Expert Systems with Applications*, 39:6123–6128. Cited By :38¡br/¿Export Date: 26 September 2019.

[Wu, 2010] Wu, Y. (2010). Research of credit sale risk evaluation based on bp neural network. Export Date: 26 September 2019.

[Yao, 2009] Yao, P. (2009). Credit scoring using ensemble machine learning. volume 3, pages 244–246. Cited By :4 Export Date: 26 September 2019.

[Yao et al., 2018] Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., Yang, Q., and Yu, Y. (2018). Taking human out of learning applications: A survey on automated machine learning.

[Ye et al., 2008] Ye, Y., Liu, S., and Li, J. (2008). A multiclass machine learning approach to credit rating prediction. *Proceedings - International Symposium on Information Processing, ISIP 2008 and International Pacific Workshop on Web Mining and Web-Based Application, WMWA 2008*, pages 57–61.

[Yi, 2009] Yi, J. (2009). Credit scoring model based on the decision tree and the simulated annealing algorithm. volume 4, pages 18–22. Cited By :12 Export Date: 26 September 2019.

[Zhang et al., 2009] Zhang, C., Li, Y., Liu, M., and Liu, Z. (2009). An enhanced approach for investment risk forecasting of electric power projects. volume 1, pages 29–32. Export Date: 26 September 2019.

[Zhao and Chen, 2009] Zhao, S.-F. and Chen, L.-C. (2009). The bp neural networks applications in bank credit risk management system. pages 527–532. Cited By :5¡br/¿Export Date: 26 September 2019.

[Zhu et al., 2010] Zhu, C., Zhan, Y., and Jia, S. (2010). Research on bp neural network evaluation model of credit risk of bank clients. Cited By :3 Export Date: 26 September 2019.