Análisis predictivo sobre insolvencia de empresas en Colombia

Laura Sofía Caita Giraldo

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Tutora
Lina María Sepúlveda Cano, PhD Ingeniería Línea Automática

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2022

TABLA DE CONTENIDO

1. RESUMEN EJECUTIVO

Cuando se crea una empresa, la idea que suele surgir en sus creadores es que pueda conseguir el éxito y expandirse en el mercado. Determinar los elementos que puedan producir insolvencia financiera y, posteriormente, la quiebra, permiten conseguir una oportuna intervención por parte de las organizaciones con el fin de evitar pérdidas. De igual manera, pueden ser un aviso para los bancos y proveedores, en caso de que estas empresas hagan una solicitud de créditos o préstamos.

Los algoritmos de aprendizaje de máquina creados con la idea de predecir si una empresa va a entrar en insolvencia suelen dejar de lado el aspecto legal, muchas veces, se realizan con bases de datos balanceadas, lo cual no refleja el mundo real. Esta monografía tiene los tiene en cuenta a ambos para realizar un análisis con factores diferenciadores y aplicados al mercado colombiano. El texto de referencia utilizado para su desarrollo fue el artículo "*Predict insolvency; a study using boosting algorithms in Colombian firms*" en el cual Correa & Lopera (2020) buscaban, a través de un algoritmo de Boosting, tener un modelo que prediga una posible insolvencia.

Debido a lo anterior, se plantea un algoritmo de clasificación "Solvencia" o "Insolvencia", el cual parte del análisis de tres bases de datos diferentes de la Superintendencia de Sociedades con datos altamente desbalanceados. Se realizó todo el proceso de exploración, estandarización, selección de hiperparámetros, uso de matrices de confusión y curva ROC. El ejercicio iterativo utiliza, además, métricas como *F1*, *Recall*, *Precision* y *Accuracy*, teniendo especial cuidado con los efectos del posible sobre-entrenamiento de los modelos.

Los resultados que se obtuvieron fueron satisfactorios, con un F1 del 98% en el modelo Árbol de decisión, el cual fue el que obtuvo una mejor puntuación. Cabe resaltar la importancia de la implementación de técnicas como el *SMOTE* para evitar falsos resultados con una alta puntuación, como se vi pudo evidenciar con los modelos basados en Regresión Logística y k-

NN. Asimismo, se observó que una Red Neuronal pequeña puede tener un buen desempeño en este tipo de problemas.

2. DESCRIPCIÓN DEL PROBLEMA

2.1 PROBLEMA DE NEGOCIO

El mundo empresarial se mueve en medio de incertidumbre constante, por lo que predecir si una empresa va a llegar a bancarrota ha sido un problema importante tanto para empresarios como para los banqueros, en especial después del crack del 29 como lo explican Carrizo, J er al (2016). Actualmente, el Aprendizaje Automático, combina la información de diversas variables financieras en un análisis interdependiente que generalmente está determinado dos grupos de clasificación que en este caso son “insolvencia” y “solvencia”.

Las empresas insolventes y sus acreedores se ven afectados cuando entran en proceso de insolvencia. La predicción eficaz de la insolvencia en una etapa temprana es relevante para que los acreedores tomen decisiones adecuadas y para reduzcan el riesgo crediticio (Liang, Lu, Tsai y Shih, 2016). Asimismo, les sirve a los bancos para tener claro el panorama en caso de que pidan un préstamo e incluso saber si podrán pagar sus deudas.

2.2 APROXIMACIÓN DESDE LA ANALÍTICA DE DATOS

La monografía busca clasificar de forma satisfactoria cuáles empresas, a partir de su balance general, estado de resultados y algunos indicadores financieros van a entrar en insolvencia entendida por la RAE como *“el estado en que una empresa es incapaz de pagar sus deudas acumuladas”*, y así conseguir que tanto sus acreedores como propietarios tomen decisiones

anticipadas para evitarlo. Asimismo, se tendrá en cuenta la ley 1116 de 2006[1], la cual explica dos procesos en el Art. 1 a los que pueden entrar las empresas una vez entran a la insolvencia.

El primero es la reorganización, en el cual se reestructura la parte administrativa, las operaciones y se analizan los pasivos y activos. Mientras que, el segundo es la liquidación, que busca la eliminación pronta y ordenada de la empresa buscando el aprovechamiento del deudor. Por otra parte, para el desarrollo del proyecto, se utilizaron tres bases de datos, todas públicas y disponibles en el Sistema Integrado de Información Societaria (SIIS) las cuales buscan un equilibrio entre la normatividad y lo financiero.

2.3 ORIGEN DE LOS DATOS

El grueso de los datos usados en esta monografía está disponible en la página de la Superintendencia de Sociedades[2]. En total se tomaron tres bases de datos, todas del periodo 2020. Las dos primeras bases de datos se relacionan con el consolidado trimestral de empresas "plenas consolidadas" que reportan información a la superintendencia con corte a diciembre (balance general y estado de resultados). Por otro lado, la tercera base de datos se conformó por el informe de empresas insolventes y liquidadas del mismo año.

Finalmente, luego de que las tres bases de datos se combinaran, se buscaron 10 indicadores financieros, los cuales fueron calculados con las fórmulas correspondientes y se dejaron en las últimas columnas de la tabla, quedando así 1122 filas (empresas) y 50 columnas (variables) que representan los datos iniciales del modelo de Aprendizaje de Máquina.

---

[1]Leyes desde 1992 - Vigencia expresa y control de constitucionalidad [LEY_1116_2006]. (s/f). Senado de la República de Colombia.
Empresas que reportan información: https://siis.ia.supersociedades.gov.co/#/massivereports

[2] Empresas insolventes y liquidadas:
https://supersociedades.gov.co/delegatura_insolvencia/Paginas/publicaciones.aspx

2.4 MÉTRICAS DE DESEMPEÑO

Como es un ejercicio de clasifican se recurre a las métricas derivadas propias de este tipo de modelo para evaluar los desempeños. Entre estas se encuentran:

- Matriz de confusión o error
- Precisión
- *Recall*
- Precisión
- *F1-Score*
- Área bajo la curva de funcionamiento del receptor (ROC)

Dado que el problema presenta datos desbalanceados, se opta por tomar la curva ROC y el F1 como métricas principales, y con base en ellas, se determina cual es el mejor modelo. Debido a la naturaleza de los datos se estima que un porcentaje de F1 de 75% o más es adecuado. Igualmente, se le dará prioridad a la técnica SMOTE, para evitar falsas predicciones con altos porcentajes.

3. DATOS

3.1 DATOS ORIGINALES

El Sistema Integrado Información Societaria (SIIS), provee distintas bases de datos públicas que permiten revisar los comportamientos de las empresas en el ámbito financiero. Específicamente, pueden consultarse de forma trimestral los estados financieros comparados de todas las empresas que declaran. Para el proyecto se utilizaron dos de estas bases de datos específicamente de empresas plenas consolidadas; el estado de situación financiera y el de resultados.

Ambas bases de datos se fusionaron, quedando en total con 1146 filas y 89 columnas, de las cuales fueron removidas 51 por cualquiera de las siguientes razones: 1) Más del 80% de los

datos no tenían información, o 2) No eran relevantes para este tipo de problema ni servían para calcular indicadores financieros. Posteriormente, fueron calculados 10 indicadores financieros mediante sus respectivas formulas y, finalmente, se contrastó con una última tabla, también del SIIS, de la lista de empresas que, durante el mismo periodo, se encontraron en algunos de los procesos de insolvencia.

Del total de empresas de la base de datos, aproximadamente el 10% entró en insolvencia, lo cual puede explicarse con factores externos del año 2020 como la pandemia. Aun así, los datos siguen presentando un desbalance alto, factor que será tenido en cuenta durante todo el desarrollo de la monografía. Finalmente, la base de datos quedó constituida por 1122 filas (empresas u observaciones) y 50 columnas (variables).

A continuación, en la **Tabla 1**, se listan las variables utilizadas.

Tabla 1.

Descripción de las variables

Nombre columna	Descripción	Tipo	Tabla original
NIT	Identificación empresa	Número	Situación financiera
Razón social de la sociedad	Nombre	Texto	Situación financiera
CIIU	Sector	Número	Situación financiera
Tipo societario	Naturaleza de la empresa	Texto	Situación financiera
Departamento de la dirección del domicilio	Departamento	Texto	Situación financiera
Ingresos de actividades ordinarias	Ingresos de las actividades normales	Número	Estado resultados
Costo de ventas	Precio de una venta	Número	Estado resultados
Ganancia bruta	Ganancia	Número	Estado resultados
Otros ingresos	Ingresos de otras actividades	Número	Estado resultados
Gastos de ventas	Gastos de una venta	Número	Estado resultados
Gastos de administración	Gastos personal y papelería	Número	Estado resultados
Otros gastos	Gastos de otras actividades	Número	Estado resultados

Ganancia por actividades de operación	Ganancia	Número	Estado resultados
Costos financieros	Costos operaciones	Número	Estado resultados
Ganancia antes de impuestos	Ganancia sin contar impuestos	Número	Estado resultados
Ingreso (gasto) por impuestos	Gasto por impuesto	Número	Estado resultados
Ganancia (pérdida) atribuible a los propietarios de la controladora	Pérdida propietarios de la controladora	Número	Estado resultados
Efectivo y equivalentes al efectivo	Efectivo	Número	Estado resultados
Cuentas comerciales por cobrar y otras cuentas por cobrar corrientes	Cuentas por cobrar	Número	Estado resultados
Inventarios corrientes	Inventarios	Número	Estado resultados
Total activos corrientes distintos de los activos	Activos corrientes diferentes	Número	Situación financiera
Activos corrientes totales	Activos	Número	Situación financiera
Propiedades, planta y equipo	Propiedades	Número	Estado resultados
Total de activos no corrientes	Activos no corrientes	Número	Situación financiera
Total de activos	Activos	Número	Situación financiera
Cuentas por pagar comerciales y otras cuentas por pagar	Cuentas por pagar	Número	Situación financiera
Otros pasivos no financieros corrientes	Pasivos no financieros	Número	Situación financiera
Total de pasivos corrientes distintos de los pasivos	Pasivos corrientes diferentes	Número	Situación financiera
Pasivos corrientes totales	Pasivos corrientes	Número	Situación financiera
Total de pasivos no corrientes	Pasivos no corrientes	Número	Situación financiera
Total pasivos	Pasivos	Número	Situación financiera
Capital emitido	Capital que sale de la empresa	Número	Estado resultados
Otras reservas	Reservas	Número	Estado resultados

Ganancias acumuladas	Ganancias	Número	Estado resultados
Total patrimonio atribuible a propietarios de la controladora	Total patrimonio	Número	Estado resultados
Patrimonio total	Patrimonio de la empresa	Número	Situación financiera
Total de patrimonio y pasivos	Total patrimonio mas pasivos		Situación financiera
Capital de trabajo	Capacidad para llevar a cabo sus actividades con normalidad en el corto plazo.	Número	Indicador financiero
Índice de insolvencia	Probabilidad de que una empresa no tenga la capacidad de cubrir sus obligaciones en el tiempo estimado.	Número	Indicador financiero
Prueba ácida	Capacidad de pago de la empresa sin la necesidad de realizar sus inventarios o sus activos fijos.	Número	Indicador financiero
Liquidez general	Es la fortaleza de la tesorería en relación a todo el pasivo circulante.	Número	Indicador financiero
Leverage	Nivel de endeudamiento de una empresa u organización en relación a sus activos o patrimonio	Número	Indicador financiero
Endeudamiento sobre activos	Cuanta deuda se usa para financiar activos	Número	Indicador financiero
Concentración endeudamiento corto plazo	Porcentaje del total de los pasivos presenta vencimiento en el corto plazo	Número	Indicador financiero
Concentración endeudamiento largo plazo	Porcentaje del total de los pasivos presenta vencimiento en el largo plazo	Número	Indicador financiero

Rotación de cartera	Promedio de días que una empresa tarda en rotar la cartera	Número	Indicador financiero
Rotación proveedores	Promedio de días que una empresa tarda en cursar un pago a un **proveedor**	Número	Indicador financiero
Proceso	Según la ley 1116 hay dos tipos: liquidación y reorganización	Texto	Empresas con insolvencia
Insolvencia	Si hay o no	Texto	Empresas con insolvencia

Fuente: Elaboración propia.

3.2 DATASETS

A partir del ejercicio previo, el Dataset o conjunto de datos se cargó al respectivo notebook, en el cual se realiza en primer momento la exploración de los datos, su imputación, detección de datos atípicos u *outliers* y escalamiento. En un segundo notebook se realiza la partición 80%-20% con el fin de mejorar la escalabilidad y reducir la contención al disminuir la contribución de las operaciones de base de datos mediante la función "train_spit_test", la cual se usó en todos los modelos que se entrenaron.

3.3 DESCRIPTIVA

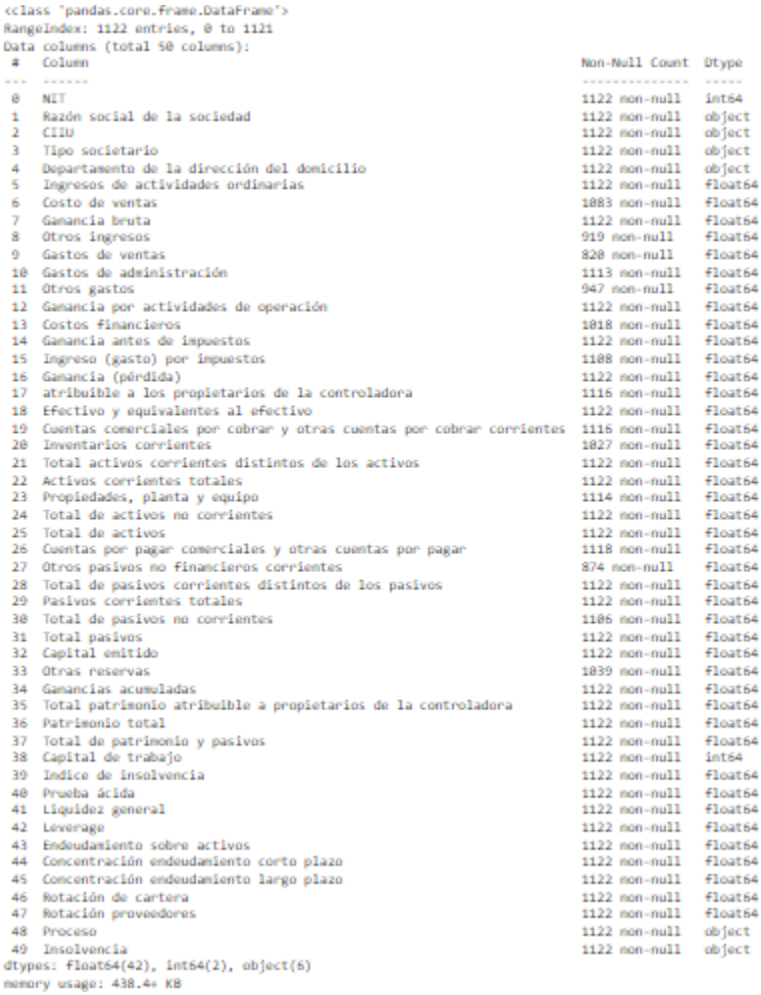
Como ya se mencionó anteriormente, la base de datos cuenta con 50 columnas y 1122 filas. La etiqueta contiene dos posibles valores: "Si" representa que la empresa entró a proceso de insolvencia y "No" que no lo ha hecho. Una de las primeras cosas a notar del ejercicio es la distribución de clases, pues los datos presentan un desbalance distribuido en un 89,48% de empresas que no han entrado en insolvencia contra el 10,52% que si lo han hecho.

Un paso importante que realizar antes de hacer el análisis descriptivo de los datos fue la identificación de los datos nulos. Como se observa en la **Figura 1.,** todas las variables con datos faltantes son numéricas. Debido a que no se quería eliminar ningún registro y que para este tipo

de caso tampoco funciona reemplazar los datos nulos por el promedio de los datos vecinos, ya que son los estados financieros de cada empresa, se optó por colocar “0”.

Figura 1.

Columnas con datos nulos



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1122 entries, 0 to 1121
Data columns (total 50 columns):
 #   Column                                                                Non-Null Count  Dtype
---  ------                                                                --------------  -----
 0   NIT                                                                   1122 non-null   int64
 1   Razón social de la sociedad                                           1122 non-null   object
 2   CIIU                                                                  1122 non-null   object
 3   Tipo societario                                                       1122 non-null   object
 4   Departamento de la dirección del domicilio                            1122 non-null   object
 5   Ingresos de actividades ordinarias                                    1122 non-null   float64
 6   Costo de ventas                                                       1083 non-null   float64
 7   Ganancia bruta                                                        1122 non-null   float64
 8   Otros ingresos                                                        919 non-null    float64
 9   Gastos de ventas                                                      828 non-null    float64
 10  Gastos de administración                                              1113 non-null   float64
 11  Otros gastos                                                          947 non-null    float64
 12  Ganancia por actividades de operación                                 1122 non-null   float64
 13  Costos financieros                                                    1018 non-null   float64
 14  Ganancia antes de impuestos                                           1122 non-null   float64
 15  Ingreso (gasto) por impuestos                                         1108 non-null   float64
 16  Ganancia (pérdida)                                                    1122 non-null   float64
 17  atribuible a los propietarios de la controladora                      1116 non-null   float64
 18  Efectivo y equivalentes al efectivo                                   1122 non-null   float64
 19  Cuentas comerciales por cobrar y otras cuentas por cobrar corrientes  1116 non-null   float64
 20  Inventarios corrientes                                                1027 non-null   float64
 21  Total activos corrientes distintos de los activos                     1122 non-null   float64
 22  Activos corrientes totales                                            1122 non-null   float64
 23  Propiedades, planta y equipo                                          1114 non-null   float64
 24  Total de activos no corrientes                                        1122 non-null   float64
 25  Total de activos                                                      1122 non-null   float64
 26  Cuentas por pagar comerciales y otras cuentas por pagar               1118 non-null   float64
 27  Otros pasivos no financieros corrientes                               874 non-null    float64
 28  Total de pasivos corrientes distintos de los pasivos                  1122 non-null   float64
 29  Pasivos corrientes totales                                            1122 non-null   float64
 30  Total de pasivos no corrientes                                        1105 non-null   float64
 31  Total pasivos                                                         1122 non-null   float64
 32  Capital emitido                                                       1122 non-null   float64
 33  Otras reservas                                                        1039 non-null   float64
 34  Ganancias acumuladas                                                  1122 non-null   float64
 35  Total patrimonio atribuible a propietarios de la controladora         1122 non-null   float64
 36  Patrimonio total                                                      1122 non-null   float64
 37  Total de patrimonio y pasivos                                         1122 non-null   float64
 38  Capital de trabajo                                                    1122 non-null   int64
 39  Indice de Insolvencia                                                 1122 non-null   float64
 40  Prueba ácida                                                          1122 non-null   float64
 41  Liquidez general                                                      1122 non-null   float64
 42  Leverage                                                              1122 non-null   float64
 43  Endeudamiento sobre activos                                           1122 non-null   float64
 44  Concentración endeudamiento corto plazo                               1122 non-null   float64
 45  Concentración endeudamiento largo plazo                               1122 non-null   float64
 46  Rotación de cartera                                                   1122 non-null   float64
 47  Rotación proveedores                                                  1122 non-null   float64
 48  Proceso                                                               1122 non-null   object
 49  Insolvencia                                                           1122 non-null   object
dtypes: float64(42), int64(2), object(6)
memory usage: 438.4+ KB
```

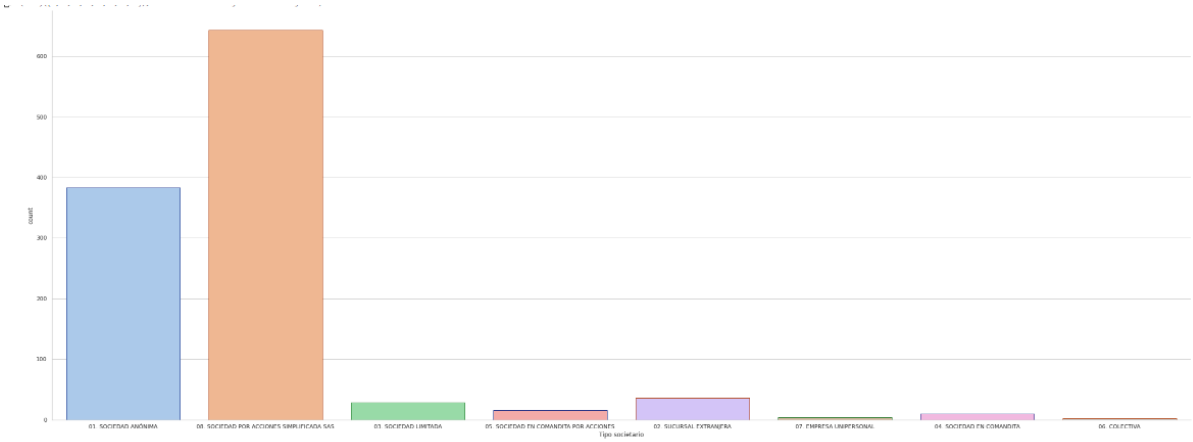
Fuente: 01_insolvencia_analisis (2022)

Cuando se tuvo la base de datos sin registros nulos, se procedió a dividir las variables en dos grupos: categóricas y numéricas. Esto con el fin de hacer más sencilla la visualización de los datos.

Con respecto a los datos categóricos, se observa lo siguiente, desde la **Figura 2** a la **4**.

Figura 2.

Distribución de tipos societarios

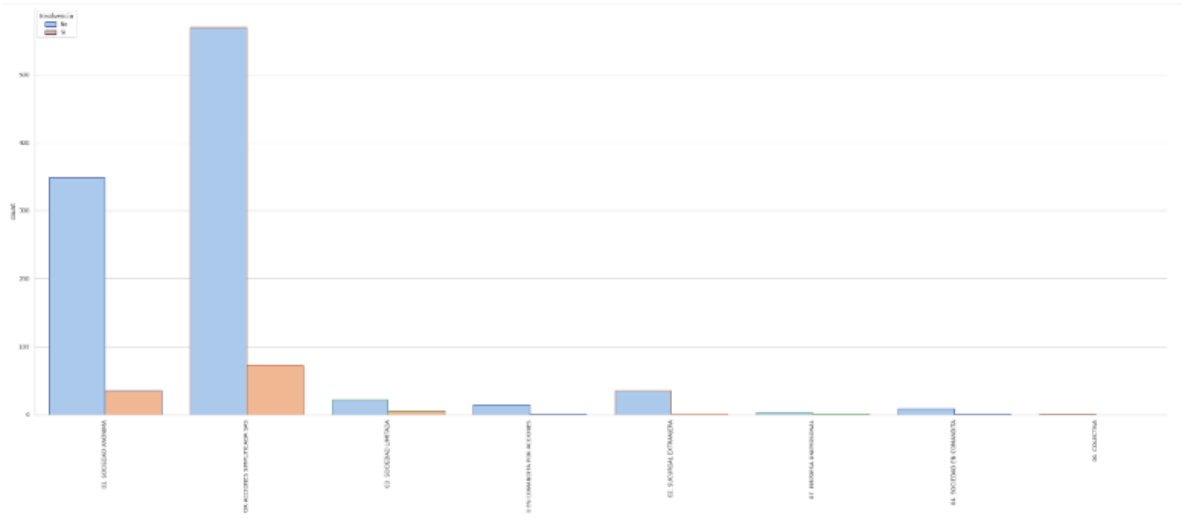


Fuente: 01_insolvencia_analisis (2022)

El tipo societario, más común es la sociedad por acciones simplificadas, seguido de la sociedad anónima. Y la que menos común es la sociedad colectiva.

Figura 3.

Tipo societario e insolvencia

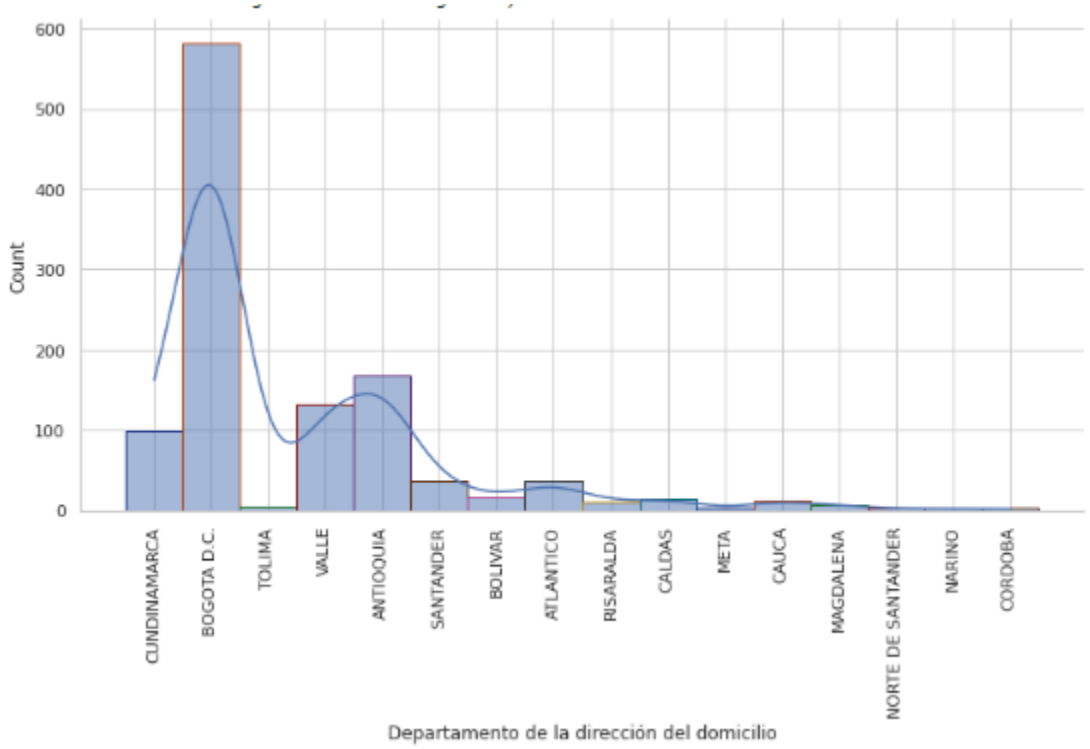


Fuente: 01_insolvencia_analisis (2022)

Al contrastar con las **Figuras 2 y 3**, es lógico que la mayoría de las empresas que están en proceso de insolvencia sean de los grupos con mayor número de empresas. Llama la atención, el caso de las sociedades limitadas, pues pese a no tener muchas empresas un porcentaje importante de estas se encuentra en proceso de insolvencia.

Figura 4.

Departamentos con mayor cantidad de empresas



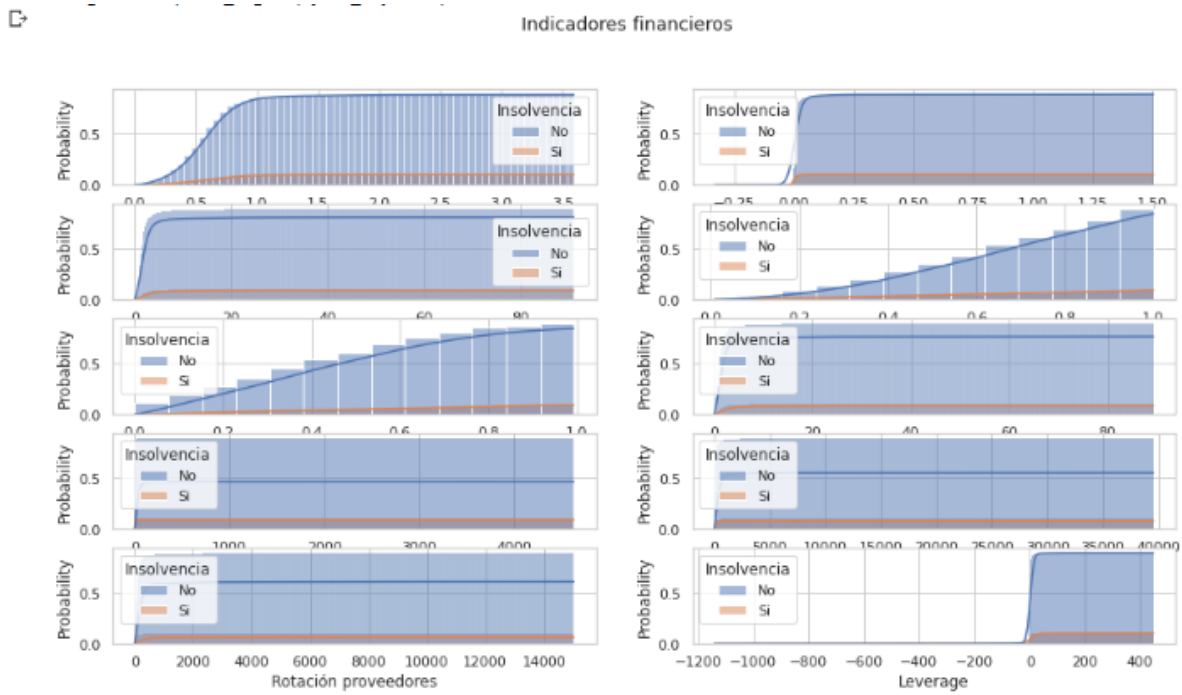
Fuente: 01_insolvencia_analisis (2022)

Como puede verse en la **Figura 4**, la mayor cantidad de empresas se concentran en los departamentos que cuentan con las ciudades más grandes del país como Bogotá, Medellín y Cali.

Por otro lado, con respecto a los valores numéricos puede observarse la **Figura 5.**

Figura 5.

Indicadores financieros



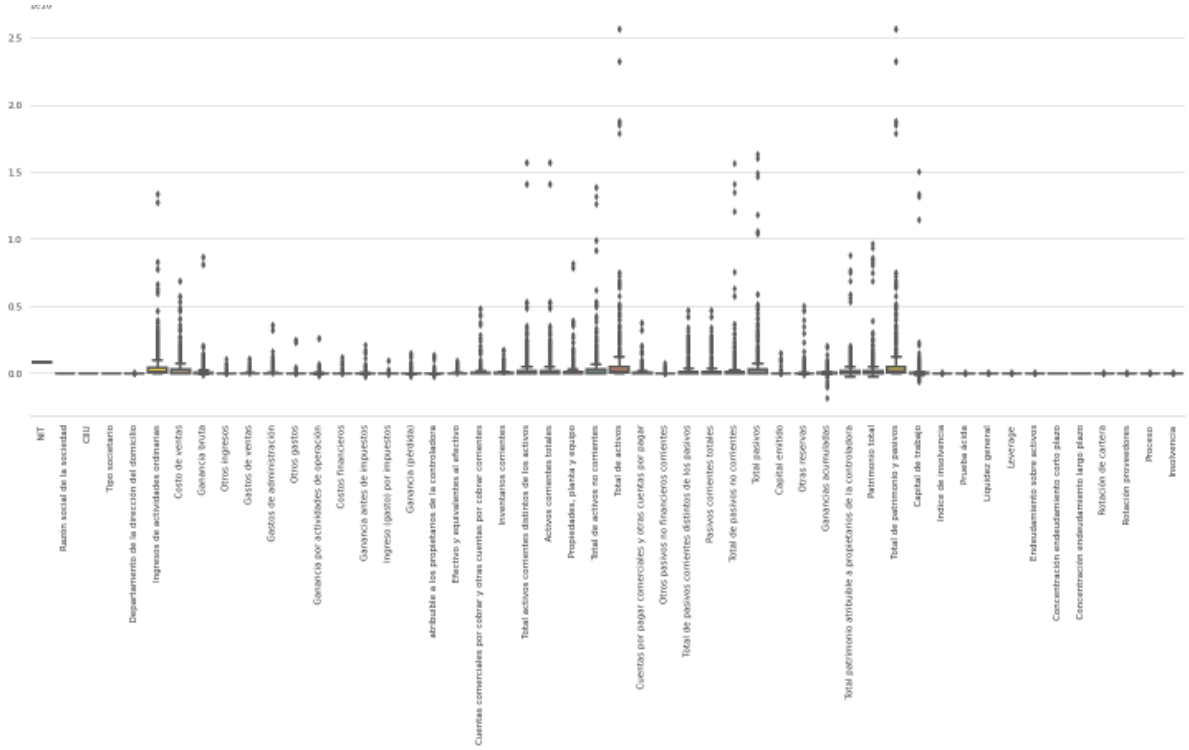
Fuente: 01_insolvencia_analisis (2022)

Llama la atención que la curva para las empresas que presentan insolvencia se mantenga constante en casi todos los indicadores financieros y que su probabilidad siempre se mantenga en valores muy bajos.

Con respecto al pre-procesamiento, el dataset cuenta con una gran cantidad de datos atípicos. En esta base de datos los registros atípicos que son necesarios de tratar para que corran los modelos, como puede verse en la **Figura 6.**

Figura 6.

Datos atípicos

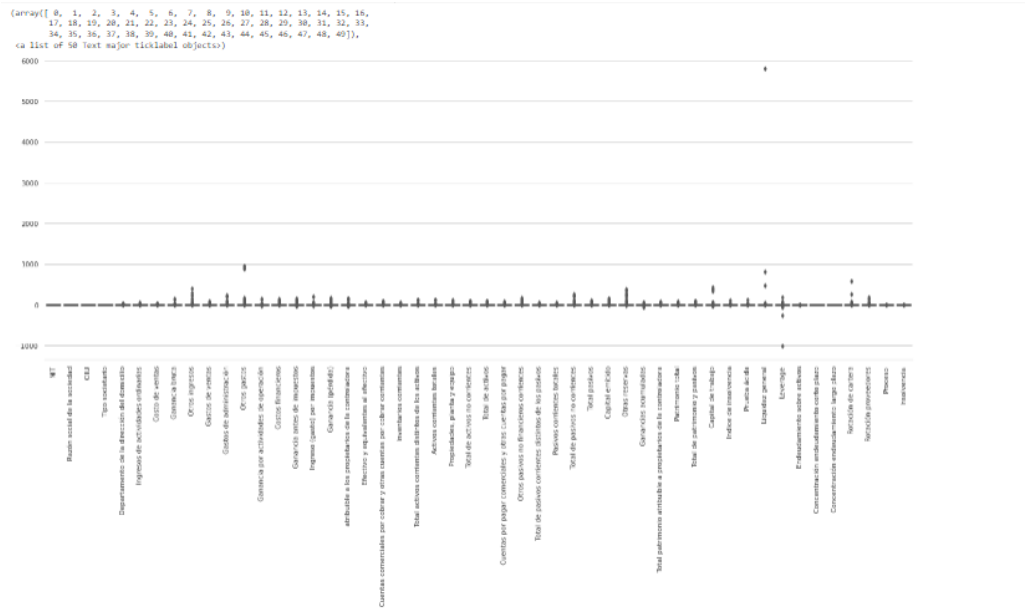


Fuente: 01_insolvencia_analisis (2022)

Si bien, para este tipo de problemas no es lo más recomendado quitar los datos atípicos, ni tampoco es posible reducirlos a los valores más altos porque muchos son números que no tienen relación directa entre sí, se concluyó que los valores más altos, los cuales eran 82 se iban a quitar, para no generar muchas complicaciones a la hora de correr el modelo.

Figura 7.

Datos con los atípicos disminuidos



Fuente: 01_insolvencia_analisis (2022)

Si bien, se quitaron muchos de los datos atípicos más altos, se dejaron algunos debido a que era mejor dejarlos en esa variable.

Figura 8.

Matriz de correlación

Fuente: 01_insolvencia_analisis (2022)

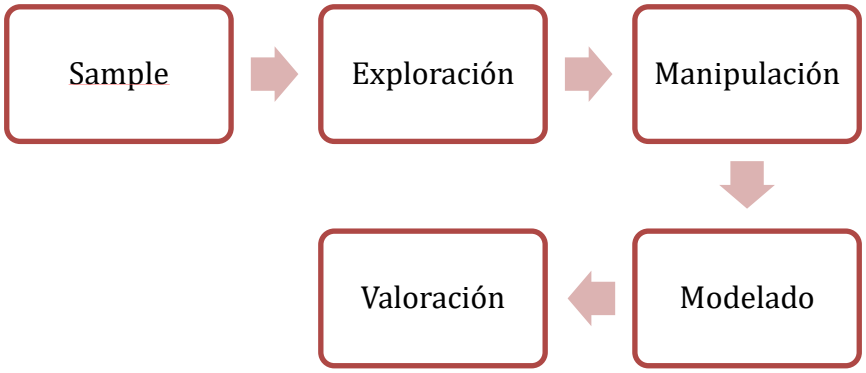
La **Figura 8.** Muestra una diagonal interesante y que algunas de las variables, en especial las de los estados financieros tienen una gran relación entre ellas.

4. PROCESO DE ANALÍTICA

4.1 PIPELINE PRINCIPAL

La metodología que se utilizó para el proyecto se conoce como SEMMA, la cual cuenta con 5 fases según Camargo & Silva (2010) *Sample* (muestra), *Explore* (explorar), *Modify* (modificar), *Model* (modelar) y *Assess* (evaluar) como se muestra en la **Figura 9.**

Figura 9.
Metodología SEMMA



Fuente: Elaboración propia

1. **Fase muestreo:** *"busca extraer una porción de datos lo suficientemente grande para contener información significativa, pero reducida para manipularla rápidamente"*. (Camargo & Silva, 2010). Dentro del proyecto, esta etapa correspondió a la primera selección de las bases de datos en el SIIS, la selección de los diez indicadores financieros y su combinación.
2. **Fase exploración:** "*propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables"*. (Camargo & Silva, 2010). Durante esta etapa, la base de datos construida en Excel pasó por un proceso de imputación debido a que tenía varios datos nulos, y se realizaron gráficos para entender mejor los datos.
3. **Fase manipulación:** "*es en la que se manipulan o modifican los datos por medio de la creación, selección y transformación de variables, para centrar el proceso de selección del modelo"*. (Camargo & Silva, 2010). En este punto, fue cuando se detectaron y

eliminaron algunos datos atípicos de la base para que no afectaran el desempeño del modelo y se escalaron los datos.

4. **Fase modelado:** "*consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado"* (Camargo & Silva 2010). En esta fase se realizó la partición de los datos y se seleccionaron los modelos para la monografía: Regresión Lineal, Árbol de Decisión, k-NN y Redes Neuronales.
5. **Fase valoración:** "*es la valoración de los resultados mediante el análisis de modelos contrastados con otros métodos estadísticos"* (Camargo & Silva 2010). En este punto, los cuatro modelos son comparados con base en las métricas establecidas previamente con el fin de concluir cual es la mejor opción.

Por otra parte, hay que resaltar que el proyecto se divide en dos notebooks y el proceso anterior de limpieza de datos que se realizó en Excel. Como se puede ver, se partió de la unión de tres bases de datos públicas y el cálculo de unos indicadores financieros. Luego, se realizó la imputación de los datos faltante y su respectiva exploración. Se continuó, la detección de datos atípicos y el escalamiento.

Luego de todo este proceso de limpieza se procedió a calcular los modelos. Primero sin tener en cuenta el sobre-entrenamiento y segundo con la técnica SMOTE. Finalmente se revisaron los resultados y se tomó el modelo con mejor *score.*

4.2 PREPROCESAMIENTO

4.2.1 Imputación de datos y revisión variables duplicadas

En primer lugar, con la función datos.info() se confirmó que existían columnas con datos faltantes, los cuales representaban el 7,79% del dataset. Para no afectar el modelo se optó por

rellenarlo con 0, ya que por la naturaleza de los datos no se podía cambiar por la media. Como se ve en la **Figura 10.**

Figura 10.

Imputación de datos con SimpleImputer

```
from sklearn.impute import SimpleImputer
```

```
[ ] imp = SimpleImputer(missing_values = np.nan, strategy= 'constant', fill_value=0)
    imp.fit(datos)
    Datos_Imputacion = datos.fillna(0)
    print(Datos_Imputacion)
```

Fuente: 01_insolvencia_analisis (2022)

Luego de esto, se quitaron los datos duplicados a través de .drop_duplicates(), el cual solo encontró un registro repetido.

4.2.2 Transformación variables categóricas

A partir de los anterior, se procedió a separar las variables en dos bloques para hacer la exploración de datos. Una vez se realizaron las gráficas y los análisis pertinentes, se utilizó la función Label Encoder para transformar todas las variables categóricas en numéricas como se ve a en la **Figura 11.**

Figura 11.

LabelEncoder

```
[ ] from sklearn.preprocessing import LabelEncoder
    encoder = LabelEncoder()

    Datos_Imputacion[categoricas] = Datos_Imputacion[categoricas].apply(encoder.fit_transform)
    Datos_Imputacion[categoricas]
```

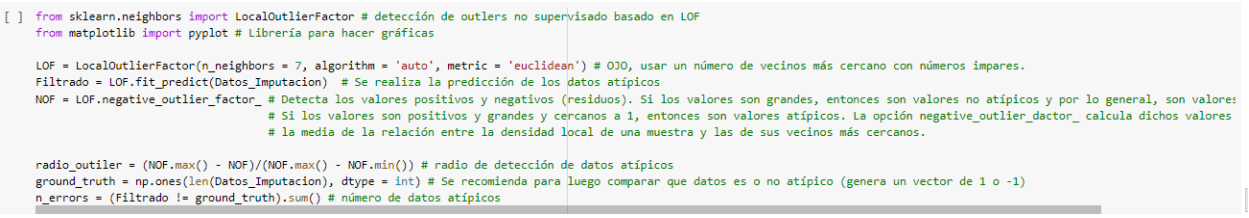
Fuente: 01_insolvencia_analisis (2022)

4.2.3 Eliminación datos atípicos

Como se observó en las **Figuras 6 y 7** se realizó un proceso de eliminación de los datos atípicos a través de la función LOF

Figura 12.

Utilización de la función LOF para detectar datos atípicos



```
[ ] from sklearn.neighbors import LocalOutlierFactor # detección de outlers no supervisado basado en LOF
    from matplotlib import pyplot # Librería para hacer gráficas

    LOF = LocalOutlierFactor(n_neighbors = 7, algorithm = 'auto', metric = 'euclidean') # OJO, usar un número de vecinos más cercano con números impares.
    Filtrado = LOF.fit_predict(Datos_Imputacion)  # Se realiza la predicción de los datos atípicos
    NOF = LOF.negative_outlier_factor_ # Detecta los valores positivos y negativos (residuos). Si los valores son grandes, entonces son valores no atípicos y por lo general, son valores
                                       # Si los valores son positivos y grandes y cercanos a 1, entonces son valores atípicos. La opción negative_outlier_dactor_ calcula dichos valores
                                       # la media de la relación entre la densidad local de una muestra y las de sus vecinos más cercanos.

    radio_outiler = (NOF.max() - NOF)/(NOF.max() - NOF.min()) # radio de detección de datos atípicos
    ground_truth = np.ones(len(Datos_Imputacion), dtype = int) # Se recomienda para luego comparar que datos es o no atípico (genera un vector de 1 o -1)
    n_errors = (Filtrado != ground_truth).sum() # número de datos atípicos
```

Fuente: 01_insolvencia_analisis (2022)

Se descubrió que 90 de los registros presentaban una gran cantidad de datos atípicos luego de dividirlos por percentiles, y se tomó la decisión de borrarlos para que se pudiera trabajar con datos más cercanos. Se tuvieron en cuenta tanto los valores más altos como los más bajos. Al final de este, quedaron 1032 registros para continuar con el pre-procesamiento.

4.2.3 Escalamiento de datos

Una vez se realizó la eliminación de los datos atípicos se escalaron para poder dejarlo listo para la implementación del modelo. Se decidió por el escalamiento estándar mediante la siguiente función.

Figura 13.

Escalamiento

```
from sklearn.preprocessing import StandardScaler
SS = StandardScaler()
Datos_Escalados_Estandar = SS.fit_transform(Datos1)
print(Datos_Escalados_Estandar)
```

Fuente: 01_insolvencia_analisis (2022)

Estos datos escalados se guardaron en un DataFrame y a partir de estos, se trabajo en el segundo Notebook.

4.3 MODELOS

Dentro del ejercicio se desarrollaron en total 4 modelos, explicados a continuación:

1. Regresión logística: en el cual se mide la relación entre la variable dependiente y la afirmación que se desea predecir, con una variable independiente, el conjunto de características disponibles para el modelo. Para ello utiliza una función logística que

determina la probabilidad de la variable dependiente. (Rodríguez, 2018). Su fórmula matemática es:

$$f(x)=1+e-x1$$

2. k-nearest neighbors (KNN): Es un método que busca las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de datos que le rodean. Su calculo matemático se puede hacer por medio de la fórmula de distancia euclidiana. (Italo, 2018).

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

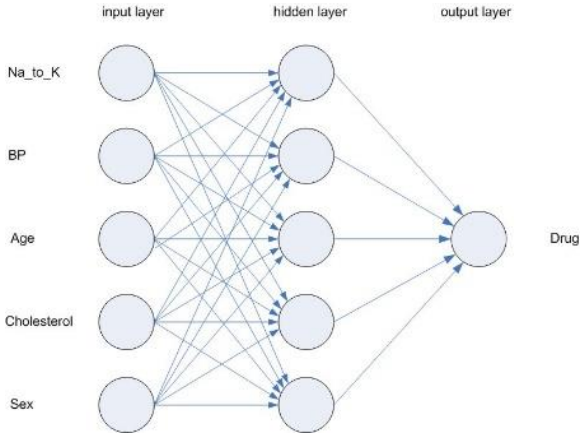
3. DecisionTreeClassifier: Los árboles de decisión son una técnica de aprendizaje automático, su nombre indica, esta técnica toma una serie de decisiones en forma de árbol. Los nodos intermedios representan soluciones. Los nodos finales dan la predicción que vamos buscando. (Heras, 2020). Matemáticamente se construyen utilizando el algoritmo voraz:

$$J(a, l_a) = \frac{m_{izquierdo}}{m} Gini_{izquierdo} + \frac{m_{derecho}}{m} Gini_{derecho}$$

4. Red neuronal: es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona relacionando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas. (IBM, 2021). Su estructura puede observarse en la **Figura 14.**

Figura 14.

Estructura redes neuronales.



Fuente: IBM, 2021.

En cada uno de los modelos se utilizó la optimización de parámetros más relevante a través del método GridSearchCV. Adicionalmente, para los tres primeros modelos, se tuvo en cuenta la técnica SMOTE para evitar falsos positivos y disminuir el sobre-entrenamiento.

5. METODOLOGÍA

5.1 BASELINE

Como línea base del proyecto se tuvo en cuenta el artículo "*Predict insolvency; a study using boosting algorithms in Colombian firms*" en el cual Correa & Lopera (2020) buscan, a través de un algoritmo de *Boosting*, tener una idea preventiva acerca de una posible insolvencia. Se tuvieron en cuenta algunos de los indicadores financieros y la ley 1116, pero no se utilizó ningún algoritmo de ensamble para el modelo, sino que se buscaron los mejores parámetros en modelos más sencillos y se probó una red neuronal.

Con respecto a la primera iteración se probó con tres modelos: DecisionTreeClassifier, k-NN y Regresión Logística, los tres se intentaron con los parámetros por defecto de cada uno. Inesperadamente con los tres dieron valores de *Accurancy* altos, como se muestra en la **Figura 15.**

Figura 15.

Resultados primera iteración

```
Classifiers:  LogisticRegression Has a training score of 88.0 % accuracy score
Classifiers:  KNeighborsClassifier Has a training score of 89.0 % accuracy score
Classifiers:  DecisionTreeClassifier Has a training score of 100.0 % accuracy score
```

Fuente: 02_insolvencia_modelos (2022)

Que esto ocurra en la primera iteración, con un Dataset altamente desbalanceado hace pensar en que posiblemente haya sobre-entrenamiento.

5.2 VALIDACIÓN

El proceso de partición inició con la función "train_test_split", con una distribución de 80% de los datos de entrenamiento frente a un 20% para test. Como los datos están desbalanceados, se tuvo en cuenta el parámetro "Stratify = Y", para que hubiese una mejor distribución de las clases. Lo anterior se puede observar en la **Figura 16.**

Figura 16.

Partición de los datos

```
#split 80-20
from sklearn.model_selection import train_test_split
y = Datos3["Insolvencia"]
X = Datos3.drop("Insolvencia", axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

Fuente: 02_insolvencia_modelos (2022)

5.3 ITERACIONES y EVOLUCIÓN

Luego de la primera iteración se decidió incorporar el método GridSearchCV, el cual sugiere los mejores parámetros para los tres modelos que se probaron, y observar si con esto mejoraba aún más el *Accurancy* o si continuaba igual, sabiendo que con esto el sobre-entrenamiento continuaría presente. Como se ve en la ***Figura 17.***

Figura 17.

Búsqueda de mejores parámetros con GridSearchCV

```
#Con overfitting y mejores parámetros

log_reg_score = cross_val_score(log_reg, X_train, y_train, cv=5)
print('Logistic Regression Cross Validation Score: ', round(log_reg_score.mean() * 100, 2).astype(str) + '%')


knears_score = cross_val_score(knears_neighbors, X_train, y_train, cv=5)
print('Knears Neighbors Cross Validation Score', round(knears_score.mean() * 100, 2).astype(str) + '%')


tree_score = cross_val_score(tree_clf, X_train, y_train, cv=5)
print('DecisionTree Classifier Cross Validation Score', round(tree_score.mean() * 100, 2).astype(str) + '%')
```

```
Logistic Regression Cross Validation Score:  88.48%
Knears Neighbors Cross Validation Score 89.58%
DecisionTree Classifier Cross Validation Score 100.0%
```

Fuente: 02_insolvencia_modelos (2022)

Si bien, el score subió un poco no representa un cambio sustancial con respecto a la primera iteración. Con el fin de confirmar que se estaba presentando sobre-entrenamiento se implementa esta vez la *Curva ROC* y se obtuvieron valores distintos.

Figura 18.

Métricas con la curva ROC

```
] from sklearn.metrics import roc_auc_score

  print('Logistic Regression: ', roc_auc_score(y_train, log_reg_pred))
  print('KNears Neighbors: ', roc_auc_score(y_train, knears_pred))
  print('Decision Tree Classifier: ', roc_auc_score(y_train, tree_pred))
```

```
Logistic Regression:  0.628314785373609
KNears Neighbors:  0.5045310015898251
Decision Tree Classifier:  1.0
```

Fuente: 02_insolvencia_modelos (2022)

Con esta iteración se hizo evidente el sobre-entrenamiento del modelo y se decidió aplicar otras técnicas como SMOTE y utilizar otras métricas para revisar el desempeño de cada modelo. En la **Figura 19**. Pueden verse los resultados de la regresión logística.

Figura 19.

SMOTE con regresión logística

```
Length of X (train): 885 | Length of y (train): 885
Length of X (test): 147 | Length of y (test): 147
--------------------------------------------------------------------------------

accuracy: 0.6858757062146892
precision: 0.16310800691030686
recall: 0.4947368421052632
f1: 0.2436941245408671
--------------------------------------------------------------------------------
```

Fuente: 02_insolvencia_modelos (2022)

En la **Figura 20**. Pueden observarse los resultados para el KNN

Figura 20.

Resultados SMOTE KNN

```
Length of X (train): 885 | Length of y (train): 885
Length of X (test): 147 | Length of y (test): 147
-----------------------------------------------------------------

accuracy: 0.7570621468926554
precision: 0.14702756892230576
recall: 0.2976608187134503
f1: 0.18922137578011008
-----------------------------------------------------------------
```

Fuente: 02_insolvencia_modelos (2022)

Y en la **Figura 21.** Los de DecisionTree

Figura 21.

Resultados SMOTE DecisionTree

```
Length of X (train): 885 | Length of y (train): 885
Length of X (test): 147 | Length of y (test): 147
-----------------------------------------------------------------

accuracy: 0.9977401129943504
precision: 0.99
recall: 0.9888888888888889
f1: 0.9891575091575092
-----------------------------------------------------------------
```

Fuente: 02_insolvencia_modelos (2022)

Fue evidente que si había overfitting y que este es un factor que debe tenerse en cuenta a la hora de modelar Datasets desbalanceados. Para este proyecto, el más sorprendente fue el KNN porque a pesar de usar el hiperparámetro “Stratify” se estaba yendo todo para un mismo lado.

La última iteración realizada fue probar una Red Neuronal no muy compleja. En la red se utilizó “Adam” como optimizador, solo se usaron dos capaz, con menos de 40 neuronas y como resultado en un primer momento se obtuvo 89% de *Accurancy*. Al igual que con los otros modelos se aplicó la técnica SMOTE y su score bajó a 66,6%.

5.4 HERRAMIENTAS

Dentro del proyecto se utilizaron distintas herramientas: Python como lenguaje de programación, Google Collab para los notebooks, Github para subir el repositorio, Excel para combinar las bases de datos y Google drive para subir los documentos.

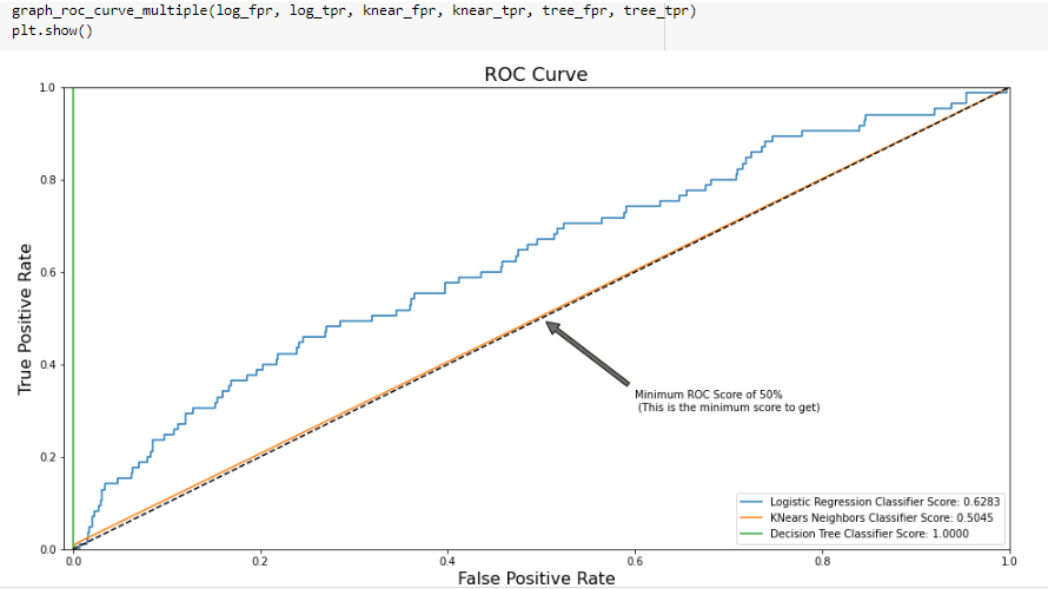
6. RESULTADOS

6.1 MÉTRICAS

Luego de realizar las iteraciones, el mejor resultado obtenido fue el 98% de F1 en el Árbol de Decisiones. Incluso, fue un modelo que, a diferencia de los otros tres no presentó mayor overfitting. Asimismo, con otras métricas como la Curva Roc puede verse también el score para este modelo como se puede ver en la **Figura 22.**

Figura 22.

Curva ROC



Fuente: 02_insolvencia_modelos (2022)

Igualmente, la última iteración con la red neuronal tuvo un *Accurancy* del 66% luego de aplicar la técnica SMOTE. Es posible que, si se añaden más neuronas, quedando un poco más compleja y usando más hiperparámetros pueda subir su score por encima de 75%.

Figura 23.

Resultado red neuronal con SMOTE

```
[172] score_Keras = model.evaluate(Xsm_train, ysm_train, batch_size=200)
      print('Accuracy on validation data with Keras: ' + str(score_Keras[1]))

      6/6 [==============================] - 0s 5ms/step - loss: 0.6400 - accuracy: 0.6667
      Accuracy on validation data with Keras: 0.6666666865348816
```

Fuente: 02_insolvencia_modelos (2022)

6.2 EVALUACIÓN CUALITATIVA

Al comparar las métricas obtenidas con las métricas de negocio requeridas, el único de los modelos que cumple con los porcentajes requeridos (mayores al 75%) fue el Árbol de Decisión. Aún así, el proyecto cuenta con una brecha importante con respecto a otros trabajos de este tipo y es que no se tuvieron en cuenta más años dentro de los datos debido a que la información es complicada de conseguir.

Por otra parte, el orden en que se ejecutaron los modelos permite observar, cómo para este tipo de problemas, los modelos sencillos tienden a caer en sobre-entrenamiento, lo que hace necesario aplicar técnicas que ayuden a equilibrar el desbalanceo de las clases.

7. CONCLUSIONES

En la actualidad hay cientos de modelos de Machine Learning que buscan predecir insolvencia en las empresas. Sin embargo, no hay un estándar sobre el tipo de datos que deben incluirse para estos modelos lo que puede dificultar la obtención de los datos. Con este proyecto fue interesante el alto score que tuvo la primera iteración porque si no se hubiera aplicado alguna técnica para ver si existía overfitting fácilmente se habría podido asumir que era un excelente modelo y altamente confiable.

Lo anterior, deja como aprendizaje que con los Datasets desbalanceados hay que prestar especial importancia a la hora de la partición y revisar los resultados con diversas métricas para que no haya falsos positivos. Igualmente, se entendió que no en todos los casos elegir los mejores parámetros puede causar una gran diferencia en el resultado.

Con los cuatro modelos, hubo una tendencia a hacer overfitting por su desbalanceo, y por eso fue necesario probar con otras técnicas como SMOTE y revisar métricas distintas al *Accurancy*. A pesar de eso, se encontró un modelo con métricas altas, el cual fue el Árbol de decisión con un *F1* de 98%.

Finalmente, el proyecto, a pesar de solo tomar un año de referencia, permite tener una idea sobre cuáles de los indicadores financieros y así mismo, de sus estados son datos relevantes que deben ser monitoreados constantemente para saber si hay posibilidades de entrar en Insolvencia.

8. BIBLIOGRAFÍA

Liang, D., Lu, C. C., Tsai, C. F., & Shih, G. A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research, 252*(2), 561–572. https://doi.org/10.1016/j.ejor.2016.01.012

Asale, R. (2022). *insolvencia | Diccionario de la lengua española*. «Diccionario de la lengua española» - Edición del Tricentenario. https://dle.rae.es/insolvencia

Leyes desde 1992 - Vigencia expresa y control de constitucionalidad [LEY_1116_2006]. (s/f). Senado de la República de Colombia. Recuperado el 7 de junio de 2022, de http://www.secretariasenado.gov.co/senado/basedoc/ley_1116_2006.html

Camargo H. y Silva M. (2010). Dos Caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP. Artículo de tecnología. Vol. 9 Nro1. Obtenido el 25 de junio desde http://www.uelbosque.edu.co/sites/default/files/publicaciones/revistas/revista_tecnologia/volumen9_numero1/dos_caminos9-1.pdf

Correa-Mejía, D. A., & Lopera-Castaño, M. (2020). Financial ratios as a powerful instrument to predict insolvency; a study using boosting algorithms in Colombian firms. *Estudios Gerenciales*, *36*(155), 229–238. https://doi.org/10.18046/j.estger.2020.155.3588

Heras, J. M. (2019, mayo 7). Árboles de Decisión con ejemplos en Python. *IArtificial.net*. https://www.iartificial.net/arboles-de-decision-con-ejemplos-en-python/

IBM Docs. (2021, agosto 17). Ibm.com. https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model

José, I. (2018, noviembre 8). *KNN (K-Nearest Neighbors) #1*. Towards Data Science. https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d

Riveros, J., Carrizo, F., Ferraro, A. D., Toledo, S., & Gardenal, L. (s/f). Edu.ar. Recuperado el 7 de junio de 2022, de https://www.fcecon.unr.edu.ar/web-nueva/sites/default/files/u16/Decimocuartas/marchese_y_otros_modelos_de_predictibilidad_de_quiebras_e_insolvencia_basados_en_analisis_de_estados_financieros.pdf

(S/f). Edu.pe. Recuperado el 7 de junio de 2022, de https://repositorio.usmp.edu.pe/bitstream/handle/20.500.12727/1266/flores_cjd.pdf?sequence=1&isAllowed=y

9. ANEXOS

Repositorio de GitHub: https://github.com/lauracaita1/CaitaLaura_2022_Insolvencia