Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencia de Datos y Matemáticas

Proceso de afinación a un modelo de predicción

Situación Problema 4

Análisis de métodos de razonamiento e incertidumbre

Camila Navarro Llaven A00517244

César Guillermo Vázquez Álvarez A01197857

Eder Gersahí Martínez León A00831442

Renata Vargas Caballero A01025281

Verónica Victoria García De la Fuente A00830383

Supervisado por
Marco Otilio Peña Díaz

Monterrey, Nuevo León. Fecha, 28 de noviembre de 2022

1. Problematización

En la actualidad, los modelos de machine learning son una poderosa herramienta al momento de dar explicación al comportamiento de fenómenos, así como soluciones a problemas que involucran cierto nivel de complejidad y, por si fuera poco, nos ayudan a tener noción acerca de la conducta que va tener a futuro una problemática. Esta tecnología surgió en el siglo pasado, concretamente entre los años de 1952 a 1956 con las invenciones de las primera maquinas consideradas inteligentes, años después por diversas situaciones externas a esta nueva rama, se impulsó la investigación y el desarrollo de esta tecnología, dando como resultado un aumento de potencia del cálculo de estos modelos, además de comenzar a ser adoptados por las compañías en las áreas de procesos y servicios para poder obtener ventajas competitivas sobre la competencia.

Existen una gran cantidad de modelos matemáticos de machine learning, y estos a su vez se encuentran clasificados dependiendo al tipo de enfoque que se le quiera dar al proceso o problemática, igualmente dependen del tipo de datos e información que se esté manejando. El machine learning se centra en tres categorías principales, siendo estas, aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo, cada uno con sus respectivos modelos.

El uso de machine learning presenta diversas ventajas, por ejemplo:

- Mayor eficiencia.
- Optimización de procesos y toma de decisiones.
- Evaluación en tiempo real.
- Automatización de procesos.
- Minimización de errores.
- Toma de acciones preventivas.

Es muy importante mencionar que al aplicar modelos de machine learning, se tiene la obligación de evaluar su desempeño y los resultados que estos arrojen, pues se tiene que corroborar que dichos modelos son capaces de ejecutar su tarea o cumplir su objetivo con el mínimo de errores, ya que, dependiendo del problema, se pudiera estar hablando de pérdidas económicas o la vida de una persona.

En relación con lo anterior, la problemática que se quiere abordar es referente al mejoramiento en la aplicación de los modelos de machine learning, pues como se mencionaba anteriormente, estos tienen que tener el menor error posible al momento de ser ejecutados, dicho esto, al medir el rendimiento de los modelos, y no obtener el valor deseado, se puede optar por añadir elementos al modelo que lo ayuden a mejorarse poco a poco, estas modificaciones pueden ir desde extender la base de datos que se está utilizando, usar más variables, añadir o modificar parámetros y hasta combinar e implementar otras técnicas estadísticas y/o matemáticas.

2. Enfoque

Utilizar métodos probabilísticos, visto desde un enfoque tecnológico para generar un modelo de predicción, específicamente un algoritmo de machine learning de aprendizaje supervisado conocido como máquina de vectores de soporte. Dicho modelo nos permitirá clasificar una variable objetivo en función de demás variables del conjunto de datos a utilizar.

3. Propósito

Aplicar un modelo de machine learning, concretamente máquina de vectores de soporte, a una base de datos del Titanic, la cual contienen información acerca de cada uno de los pasajeros. Dicha información es sobre el número del boleto, su tarifa, la clase del boleto, número de cabina, sexo, entre otros, además de una variable que nos indica si la persona sobrevivió o no al acontecimiento del Titanic. En relación con lo anterior, con el modelo que se generé se pretende predecir si los pasajeros del Titanic sobrevivieron a partir de la información que nos ofrecen las variables que se mencionaron antes, por lo que se busca que el modelo tenga un buen desempeño al momento de predecir dicha etiqueta al momento de compararla con la real etiqueta original de la base de datos, y esto se pretende lograr mediante un proceso de afinación del modelo, el cual consta de 5 versiones de este mismo, donde la primera tendrá el peor rendimiento y el modelo final será el que tenga el mejor desempeño.

4. Información

Máquina de vectores de soporte, SVM por sus siglas en inglés (Support Vector Machines), es un modelo que fue desarrollado en la década de los 90´s, al principio

se pensó para clasificación binaria, pero actualmente su implementación se ha extendido a problemas de clasificación múltiple y regresión.

SVM se basa en el Maximal Margin Classifier, y este se fundamenta en el concepto de hiperplano, dicho esto, recordemos que un hiperplano se define como un subespacio plano y afín de dimensiones *p-1*. Hablando de dos dimensiones, el hiperpalno sería un subespacio de una dimensión, o en otras palabras una recta, si se tuviera un espacio tridimensional, entonces el hiperplano es de dos dimensiones y así sucesivamente.

La definición matemática de un hiperplano está dada por los parámetros β_0, β_1, β_2 hasta β_p, donde se tienen definidos todos los puntos por el vector $(X = x_1, x_2, ..., x_n)$. La ecuación resultante es la siguiente:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = 0$$

En ella vemos que los puntos que satisface la ecuación se encuentran o mejor dicho pertenecen al hiperplano, pero cuando x no satisface la ecuación anterior, entonces se tiene lo siguiente:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p < 0$$

o bien

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p > 0$$

Esto nos indica la posición donde caerá el punto x, si es de un lado u otro.

A continuación en la *Figura* 1 se observa un hiperplano de dos dimensiones y la representación de las 3 ecuaciones antes declaradas.

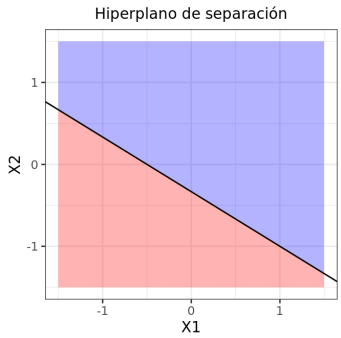

Figura 1: Ejemplo de Hiperplano.[1]

A raíz de este razonamiento, se llega a presentar casos donde hay una gran cantidad de formas de separar los datos, como se observa en la *Figura* 2.

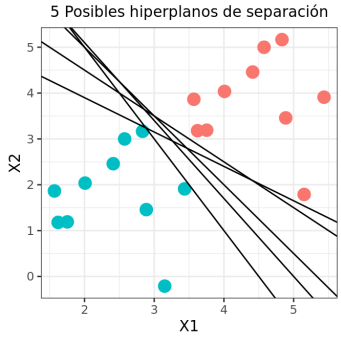

Figura 2: Distintas opciones de Hiperplano para separar datos. [1]

Para poder encontrar la solución óptima de este conjunto, aplicamos el hiperplano óptimo de separación o también conocido como hiperplano de margen máximo, el cual se encarga de buscar aquel hiperplano que se encuentra más alejado de todo nuestro conjunto de datos, y para calcularlo es necesario tomar en cuenta la distancia perpendicular de todas y cada una de las observaciones en referencia a un determinado hiperplano, esto al aplicarlo resulta ser inconveniente, pues existen infinitos hiperplanos contra que medir las distancia, es por eso que se tiene lo que se conoce como máquina de vectores de soporte.

Las máquinas de vectores de soporte engloban los conceptos expuestos anteriormente, pero con ciertas modificaciones para que su implementación sea óptima. En pocas palabras este modelo se encarga de seleccionar la mejor forma de separar los conjuntos de datos, dejando la mayor distancia entre la "línea" y las observaciones. Cabe mencionar que para generar dicha línea, no es necesario tomar en cuenta todas y cada una de las observaciones con las que se cuenta, simplemente se hace uso de los puntos más cercanos a la línea, pues dichos puntos fungen como un apoyo o soporte para la ubicación de nuestra línea, aquellos puntos son conocidos como vectores de soporte, e influyen tanto en la posición y orientación del hiperplano que maximiza el margen de separación. En la *Figura* se observa el funcionamiento del modelo de SVM.

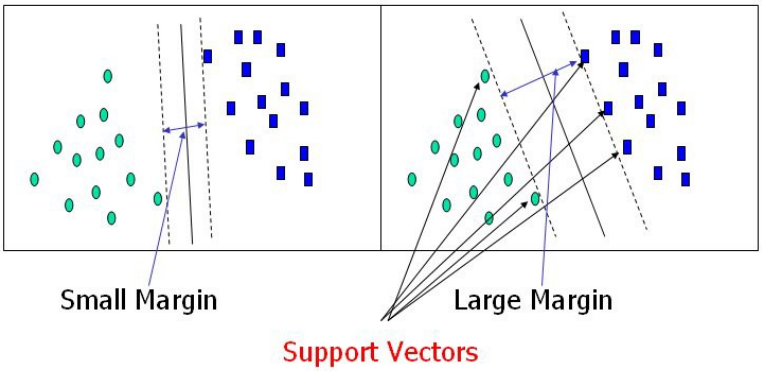

Figura 3: Funcionamiento de Maquina de vectores de soporte (SVM).[2]

5. Razonamiento

Para el entrenamiento del modelo SVM se hicieron con diferentes parámetros de entrenamiento, pero para poder trabajar con ello, tuvimos que normalizar y limpiar nuestros datos. Por ejemplo antes de predecir cuantas personas iban a sobrevivir primero predijimos la edad de las personas que nos faltaban para así poder entrenar con el parámetro de edad el modelo de predicción de sobrevivientes.

Para el entrenamiento del modelo de la edad utilizamos los parámetros de genero, titulo y la clase. Y para el entrenamiento del modelo de sobrevivientes utilizamos los siguientes parámetros y obtuvimos el siguiente performance:

- Genero - Performance: 76.55 %
- Edad - Performance: 62.2 %
- Genero y Título - Performance: 75.3 %
- Genero, Título y Lugar de embarcamiento - Performance: 75.8 %
- Genero, Título, Lugar de embarcamiento y Clase- Performance: 78.9 %

Creemos que el score es menor con la edad ya que también esta predecía y genera un ruido en el entrenamiento para el modelo de predicción de sobrevivientes. Lo que se nos ocurre para mejorar el resultado es en lugar de predecir la edad exacta hacerlo

por rangos de edad y así optener mejores resultados.

Resultados de kaggle:

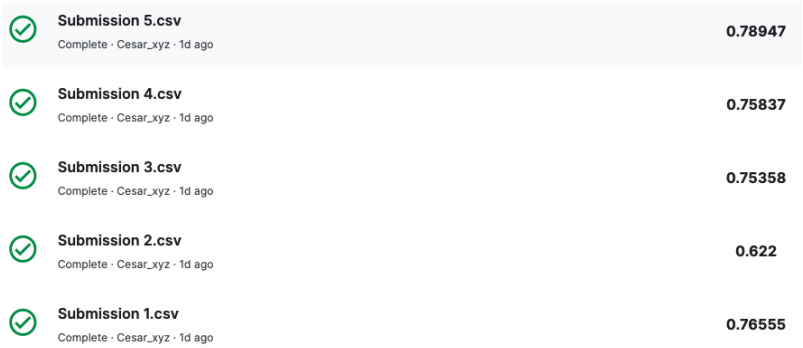

Figura 4: Kaggle performance and submit.

6. Conclusiones

Podemos ver que el desempeño del modelo mejora con distintos ajustes que involucran no solo hiperparámetros del modelo, sino también en los datos que decidimos seleccionar para hacer las predicciones.

Referencias

[1] f. J.A., "Máquinas de Vector Soporte (Support Vector Machines, SVMs)," 4 2017. [Online]. Available: https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

[2] f. R., Rohith, "Support Vector Machine — Introduction to Machine Learning Algorithms," 11 2022. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[3] P. V.G., "Una Breve Historia del Machine Learning - Think Big Empresas," 6 2021. [Online]. Available: https://empresas.blogthinkbig.com/una-breve-historia-del-machine-learning/

[4] "Machine Learning — Qué es, tipos, ejemplos y cómo implementarlo," 9 2022. [Online]. Available: https://www.grapheverywhere.com/machine-learning-que-es-tipos-ejemplos-y-como-implementarlo/

[5] "Beneficios que aporta aplicar Machine learning en la empresa," 4 2019. [Online]. Available: https://zemsaniaglobalgroup.com/machine-learning-en-la-empresa/