Aplicación de aprendizaje automático supervisado en la detección de depresión en zonas rurales

Luis David Huante
LTIC
ENES-Morelia
Email: luisdhuantes@gmail.com

***Resumen*—Este proyecto consiste en la comparación de 2 métodos de aprendizaje automático supervisado, Regresión Lineal y KNN (K Nearest Neighbors) en la predicción de la presencia de depresión en habitantes de zonas rurales.**

I. Introducción

Hoy en día, la depresión es una problemática muy presente en la sociedad actual. Diversos factores de índole social, cultural, económica, entre otros, influyen fuertemente en la posibilidad de diferentes sectores poblacionales de desarrollar depresión. En este proyecto se analizará un conjunto de datos que contiene diversas métricas concernientes a las condiciones de vida de los habitantes de zonas rurales y su relación con la depresión. A través de las diferentes características (sexo, ingresos, gastos, número de hijos, etc), podemos inferir cuáles de ellas tendrán depresión a partir de la aplicación de 2 de los distintos métodos de aprendizaje automático (KNN y Regresión Logística), para posteriormente comparar resultados entre sí y elegir al más apto dados los datos proporcionados.

II. Materiales y métodos

Se utilizó un conjunto de datos de 1427 filas y 23 columnas que constituye un estudio sobre las condiciones de vida de las personas que viven en zonas rurales. Contiene los atributos de ID de encuesta, ID de ciudad, Sexo, Edad, Estado marital, Nivel de eduación, Número de niños, Total de miembros de familia, Ganado activo, Activos duraderos, Activos seguros, Gastos de manutención, Otros gastos, Salario entrante, Salario propio de granja, Salario propio de negocio, Salario externo, Salario de agricultura, Gastos de granja, Mercado laboral primario, Inversiones duraderas, Inversiones no duraderas y Estatus de depresión.

II-A. Data Wrangling y Data Cleaning

Las columnas relativas a cualquiera de las ID de los encuestados así como las columnas con datos faltantes fueron eliminadas de la tabla durante el análisis. Tras verificar el conjunto de datos, sólo se encontraron 2 columnas y 2 filas sin utilidad de las 1427 filas y 23 columnas presentes.

III. Procedimientos y resultados

III-1. K Nearest Neighbors: Para comenzar el análisis, tomamos el 0,8 % de los datos para entrenamiento y el resto para prueba. Ahora se debe encontrar el valor óptimo de k, por lo que se generaron 2 gráficas que muestran el comportamiento del error conforme K aumenta.

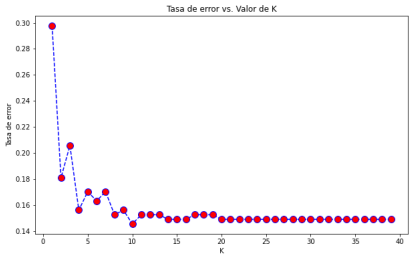

Figura 1. La gráfica muestra la función del error conforme K aumenta

Tras analizar en la gráfica los posibles valores para K, el número con el cual se alcanzó el mayor balance fue 9. A partir de este valor de K, disminuía drásticamente el error. Aunque este se mantiene en un margen de 18 % y 16 %, debemos optar por encontrar el equilibrio y minimizarlo sin afectar otras variables. Se alcanzó un porcentaje de Accuracy de 84.39 % teniendo con una desviación estándar del 2 % por cada experimento realizado.

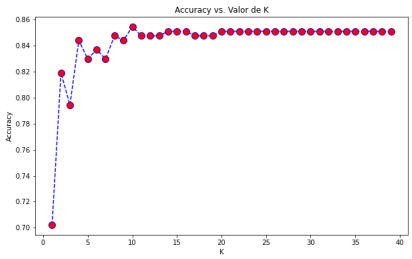

Figura 2. Se muestra la función del Accuracy Score conforme K aumenta

Con los resultados obtenidos, es decir, con la accuracy de 85 % en K=9 podemos definir ese valor como K. Siempre es mejor una K un poco más grande para evitar problemas como el sobreajuste o subajuste, por lo que en este caso se elegirá a K como 9, obteniendo una mejor eficiencia para ese valor de K.

III-2. Regresión Logística: Para poder realizar la comparación de los resultados debemos ahora aplicar Regresión Kogística. Se cambiaron algunas características en el conjunto

de datos. Por ejemplo, pesar de que todas las columnas son numéricas, los datos aún no están listos para construir el modelo. Se necesita convertir la variable categórica a binarias. (get-dummies-Variables ficticias). Se obtuvo un Accuracy Score de 85.26 %. Mayor que el conseguido en K Nearest Neighbors.

En este caso, el modelo no identifica a nadie como propenso a la depresión mientras que marca a 52 personas como sanas que no lo son. La última circunstancia tampoco es muy buena. Se tenía un total de 52 casos de depresión, de los cuales el modelo no encontró ninguno.

Sin embargo, ya está claro que el modelo está sub-ajustado, ya que presentó un resultado débil en las métricas. Sin embargo, podemos atribuir este subajuste a otro conjunto de problemas: la cantidad insuficiente de datos o incluso que el número de características es mayor que el número de objetos.

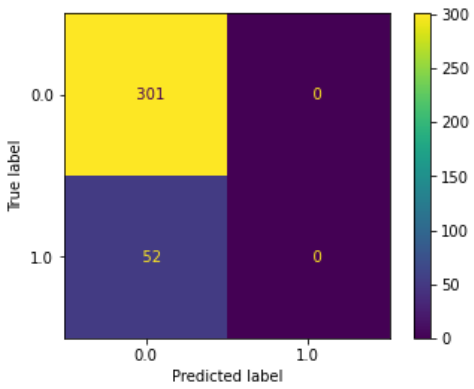

Figura 3. La matriz de confusión denota un subajuste en el modelo

IV. Conclusiones

Tras haber analizado y obtenido los resultados de los dos métodos, es notable la diferencia entre ambos modelos. Primeramente, el análisis realizado con K Nearest Neighbors dió un Accuracy Score más alto que el conseguido con Regresión Logística. Estos resultados se consiguieron a través de un proceso de exploración, limpieza y formateado de datos, a través de los cuales se consiguió el conjunto de dtos final que iba a utilizarse con los métodos de clasificación. Los datos venían casi listos para analizar, por lo que sólo fue necesario eliminar una instancia con datos faltantes. Al empezar el análisis con K Nearest Neighbors, primero se realizó una gráfica para visualizar los valores de K con los que se podía conseguir mejores resultados. La gráfica resultó ampliamente útil para observar el comportamiento del error conforme aumentaba K, y al final el valor más óptimo fue K = 9. Esto llevó a un valor de Accuracy del 84.39 En contraste, al realizar el análisis con Regresión Logística, se obtuvieron resultados ligeramente mejores. Se llegó a una Accuracy de 85.26 %. Analizando la matriz de confusión, se puede deducir que el modelo está sub-ajustado, ya que presentó un resultado débil en las métricas. Sin embargo, podemos atribuir este subajuste a problemas como la cantidad insuficiente de datos o incluso que el número de características es mayor que el número de objetos, lo cual al final se traduce en un desajuste en la función sigmoide, es decir la función logística.

Las variables que más influían en la predicción eran la de los ingresos, gastos y patrimonio, por lo que podemos deducir que los habitantes de zonas rurales tienden a desarrollar depresión cuando existe un patrimonio económico más bajo con gastos elevados.

A pesar de que ambos modelos no variaron mucho, sí hubo diferencias entre ellos, siendo el de Regresión Logística el más óptimo para la tarea dados los datos proporcionados, siendo bastante preciso. Fuee la calidad de los datos el mayor obstáculo para su desempeño, ya que el modelo es tan bueno como los datos se lo permitan. Finalmente, se definió a Regresión Logísitca como el modelo ganador.

Este proyecto resultó muy útil para reforzar los conocimientos matemáticos y de programación implicados en el aprendizaje automático, así como para apreciar los conjuntos de datos en los que estos métodos pueden ser aplicados, ya que las posibilidades y el potencial del aprendizaje es muy vasto.

Bibliografía

1. N/A. (2021). pandas documentation mayo 28, 2022, de Pandas Sitio web: https://pandas.pydata.org/docs/
2. N/A. (2021). scikit-learn documentation mayo 28, 2022, de SciKit-Learn Sitio web: https://scikit-learn.org/stable/index.html
3. Babativa, D. (2019) Depression mayo 28, 2022, de Kaggle Sitio Web: https://www.kaggle.com/datasets/diegobabativa/depression