Proyecto Primer Parcial Entrenamiento de Datos.

C. D. Arcentales, y C. N. Mejia.

Universidad Internacional del Ecuador.

***Resumen*— El entrenamiento de modelos no lineales suelen ser complejos, pero gracias a los avances en la inteligencia artificial se han logrado encontrar algoritmos para lograr predicciones en base a datos. En este caso se han analizado 3 algoritmos diferentes, el de árbol de decisión, el de métodos en base a reglas, y el de máquina de soporte vectorial, se ha entrenado a estos modelos y se les ha dado variables dependientes e independientes, luego se ha calculado la exactitud de cada modelo, dando como resultado que el de árbol de decisión es el modelo más exacto.**

***Abstract*-- Training nonlinear models is usually complex, but thanks to advances in artificial intelligence, algorithms have been found to achieve predictions based on data. In this case, 3 different algorithms have been analyzed, the decision tree, the rule-based methods, and the support vector machine. These models have been trained and given dependent and independent variables, then has calculated the accuracy of each model, resulting in the decision tree being the most accurate model.**

INTRODUCCIÓN

Los modelos de aprendizaje ya sean lineales o no lineales, sirven para el reconocimiento de diferentes tipos de patrones. Las relaciones de estos patrones suelen ser lineales o no lineales, por lo tanto, para hacer predicciones acerca de estos, se pueden aplicar ciertos algoritmos en base a variables dependientes e independientes.

Algunas aplicaciones de modelos lineales son: Predicción de precios de bienes inmuebles, evaluación de riesgo crediticio, predicción de ventas, etc. Por otro lado, las aplicaciones de modelos no lineales pueden ser: reconocimiento de voz e imágenes, diagnóstico médico, sistema de recomendación en línea, predicción del clima, análisis de sentimientos, etc. Preparando su trabajo digital.

METODOLOGÍA

A. Librerías

El primer paso para el desarrollo del proyecto es la importación de las librerías necesarias para los procesos a desarrollar a continuación.

B. Preprocesamiento de Datos

Para empezar se carga el DataFrame mediante la librería pandas y se procede a observar los tipos de datos y si existen datos faltantes o ‘nan’, para lo cual se utilizan comandos como ‘.count()’ y ‘.describe()’ para analizar si los datos existentes son congruentes con el problema propuesto, tras lo cual se remplazan y completan los valores ‘nan’ y ‘0’ y se procede a analizar mediante el comando ‘.unique()’ si los elementos que conforman las columnas son parte del sistema a analizar. Como en el caso de la columna ‘Estado civil’ donde se observan elementos ‘nan’ y ‘0’, o en los casos de las columnas de la ‘M1’ hasta ‘M6’ donde existen elementos incongruentes (-2,0) según el problema planteado, tras lo cual se remplazan por los elementos mas repetidos dentro del sistema.

C. Entrenamiento del sistema.

Una vez preprocesado el conjunto de datos se separa en un conjunto de entrada (x) y en un conjunto de salida (y), donde para el conjunto de entrada se excluyen las columnas ‘ID’ y ‘SP’ y para el conjunto de salida solo se utiliza la columna ‘SP’; se entrena el conjunto de datos mediante los algoritmos ‘RandomForestClassifier’ y ‘XGBClassifier’, los cuales nos permiten observar estadísticas como la exactitud, precisión, sensibilidad, especificidad, y f1-score.

C. El mejor modelo entrenado.

Mediante las librerías ‘pickle’ y ‘joblib’ los cuales permiten que se guarde el mejor modelo entrenado y se pueda cargar de manera optima en otros sistemas.

CONCLUSIÓN

En nuestro modelo se calcularon las siguientes métricas junto con sus valores: exactitud (0.81), precisión (0.84), sensibilidad (0.93), especificidad (0.39) y f1-score (0.89). Las métricas para tomar en cuenta son la sensibilidad y la especificidad, la más alta y baja, respectivamente. Lo que se puede concluir con esto es que nuestro modelo es capaz de identificar de mejor manera los casos positivos, por otro lado, no predice bien los casos negativos. El caso consiste en determinar si el cliente va a pagar su deuda o no dependiendo de los datos del cliente, en este caso el modelo es eficaz para determinar quienes sí van a pagar la deuda debido a su alta sensibilidad. Por otro lado, para determinar quienes no van a pagar sus deudas, resulta ineficaz el modelo, resulta

RECOMENDACIONES

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” *Phil. Trans. Roy. Soc. London,* vol. A247, pp. 529-551, April 1955.

[2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism,* 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, “Title of paper if known,” no puplicado.

[5] R. Nicole, “Title of paper with only first word capitalized,” *J. Name Stand. Abbrev.,* en impresión.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Transl. J. Magn. Japan,* vol. 2, pp. 740-741, August 1987 [*Digests 9th Annual Conf. Magnetics Japan,* p. 301, 1982].

[7] M. Young, *The Technical Writer’s Handbook.* Mill Valley, CA: University Science, 1989.