INTELIGENCIA ARTIFICIAL "HEART CLEANING"

Josselyn Amarilis Macías Veliz, Mary Elena Muñoz Macias

16 January, 2023

Abstract

En el siguiente artículo se explica e identifica el uso de dos algoritmos de aprendizaje, los cuales tienen conceptos diferentes (tree y el logistic regression) son algoritmos útiles para realizar tareas de clasificación de caracteristicas. Además se tendrá una comparación de estos dos algoritmos mediante un análisis de los mismos y se mostrará la eficiencia de ambos modelos.

1 Introdución

Los algoritmos de clasificación como el tree y el de logistic regression se utilizan para realizar la clasificación de las tareas. El algoritmo logistic regression implica tener grandes cantidades continuas de datos, es uno de los algoritmos Machine Learning más usados para la clasificación de dos clases, es muy sencillo implementarlo se utilizan para realizar cualquier problema de clasificación binaria, una vez que estos datos son entrenados, pasan a predicción y es ahí que se observan los valores que son pronosticados con logistic regression, y luego de eso se realiza una conexión a matriz, proporciona las instancias de una clase real a otra que se predice, muestra los datos clasificados, es decir poder observarlos de una manera más específica, también se utilizó el test and Score en general, este algoritmo se puede utilizar para varios problemas de clasificación.

A diferencia del algoritmo tree es un tipo de aprendizaje supervisado que crea modelos con valor en variables y que tenga un destino, éste algoritmo se lo considera como el más sencillo, pero a la vez uno de los más poderosos del Machine Learning, son muy usados cuando se trabaja con muchos datos y que son relativamente complejos, el objetivo principal de este modelo es que aprenda a calcular una frontera de decisión que permita asignar una u otra categoría de dato de interés.

2 Marco Teorico

(Gozalez, 2021)(1) Indica que el aprendizaje automático es una rama de la ciencia en el campo de la inteligencia artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto significa identificar patrones complejos en millones de datos. Una máquina que realmente aprende es un algoritmo que puede observar los datos y predecir el comportamiento futuro. También en este contexto, automáticamente significa que estos sistemas mejoran de forma autónoma con el tiempo sin intervención humana. En resumen, con el aprendizaje automático puede pasar de reactivo a proactivo. Los datos históricos del grupo de clientes, debidamente organizados y procesados en su conjunto, crean una base de datos que puede ser utilizada para predecir comportamientos futuros, favoreciendo aquellos que mejoran los objetivos comerciales y evitando aquellos que los perjudican. Es imposible que una sola persona analice y saque conclusiones de esta gran cantidad de datos, y hacer predicciones es aún más difícil. Los algoritmos, por otro lado, pueden detectar patrones de comportamiento en función de las variables proporcionadas y detectar qué variables llevaron a la cancelación de la inscripción como cliente. El siguiente diagrama es un ejemplo simplificado de predicciones basadas en datos ficticios de compañías telefónicas, pero utilizando herramientas reales de aprendizaje automático. Una máquina aprende cuando es capaz de acumular experiencia (a través de datos, programas, etc.) y desarrollar nuevos conocimientos para que su desempeño en tareas específicas mejore con el tiempo. Herramienta que buscan mejorar el análisis de datos, en pro de una predicción futura, ya sea por la implementación de nuevos sistemas o simplemente el mejoramiento de los ya existentes, mediante el uso de algoritmos basados en información antigua o reciente que permita el funcionamiento óptimo del sistema a trabajar. En esta rama la integración de la inteligencia artificial mediante el machine learning, busca la eficiencia de los equipos generadores de diagnósticos médicos así se podrán descartar errores humanos en el análisis de datos y se reducirían costos en investigación.

3 Algoritmos de clasificacion TREE

(Gonzalez, 2018)(2) Se refiere a que el árbol de decisión o clasificación de árboles de decisión es un tipo de algoritmo de aprendizaje supervisado que se utiliza principalmente en problemas de clasificación, aunque funciona tanto para variables de entrada y salida categóricas como continuas. En esta técnica, dividimos los datos en dos o más conjuntos homogéneos según el diferenciador más significativo en las variables de entrada. El árbol de decisión identifica la variable más significativa y su valor que produce los mejores conjuntos de población homogéneos. Se evalúan todas las variables de entrada y todos los puntos de división posibles y se elige el de mejor resultado. Los algoritmos de aprendizaje basados en árboles se consideran uno de los mejores y más utilizados métodos de aprendizaje supervisado. Los métodos basados en árboles admiten modelos de predicción con alta precisión, estabilidad y fácil interpretación. A diferencia de los modelos lineales, representan bastante bien las relaciones no lineales.

3.1 Clasificación TREE

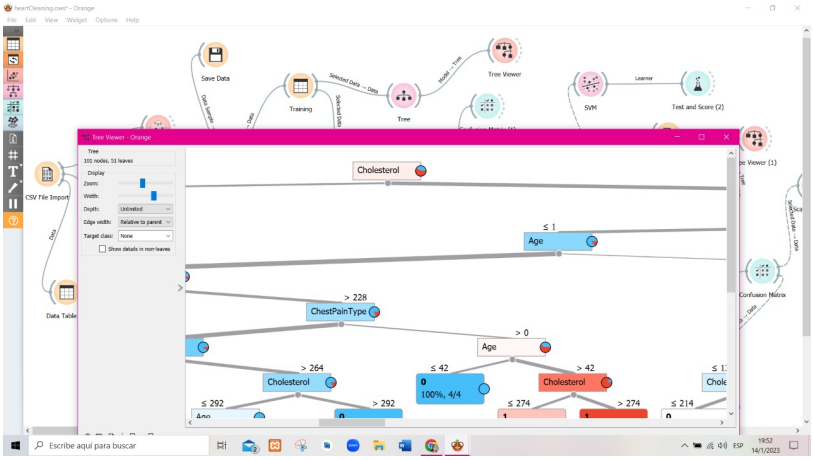

Figure 1: En esta imagen está simplificado, la predicción real tiene muchos más nodos

3.2 Las ventajas de este tipo de algoritmo son:

Fácil de comprender. El resultado del árbol de decisión es muy fácil de entender incluso para personas sin antecedentes analíticos, no se requieren conocimientos estadísticos para leerlo e interpretarlo. Útil en la exploración de datos, el árbol de decisión es una de las formas más rápidas de identificar las variables más importantes y la relación entre dos o más. Usando árboles de decisión, podemos crear nuevas variables o características que tengan un mejor poder predictivo para la variable objetivo. Se requiere menos limpieza de datos. Requiere menos limpieza de datos en comparación con otras técnicas de modelado. Una vez más, no se ve afectado por los valores atípicos y faltantes en los datos. El tipo de datos no es una restricción. Puede manejar variables numéricas y categóricas. Método no paramétrico, se considera un método no paramétrico, lo que significa que los árboles de decisión no hacen suposiciones sobre el espacio de distribución y la estructura del clasificador.

3.3 Algoritmo de logistic regression

Las técnicas de clasificación son una parte importante del aprendizaje automático, porque alrededor del 70 % del problema es la clasificación. Hay muchos algoritmos de clasificación, pero la regresión logística es común y es un método de regresión útil para resolver problemas de clasificación binaria. La regresión logística es un algoritmo de clasificación utilizado para predecir la probabilidad de una variable dependiente categórica. En la regresión logística, la variable dependiente es una variable binaria que contiene datos codificados como 1-0, sí-no, abrir-cerrar, etc. La variable de resultado o objetivo es de naturaleza dicotómica. Dicotómico significa que solo hay dos clases posibles. Por ejemplo, puede usarse para problemas de detección de cáncer o para calcular la probabilidad de que ocurra un evento. La regresión logística es uno de los algoritmos de aprendizaje automático más simples y más utilizados para clasificar dos clases. Es fácil de implementar y puede usarse como base para cualquier problema de clasificación binaria. Describe y estima la relación entre una variable dependiente binaria y las variables independientes. Este modelo

logístico binario se utiliza para estimar la probabilidad de una respuesta binaria en función de uno o más predictores o variables independientes. Esto permite decir que la presencia de un factor de riesgo aumenta la probabilidad de un resultado en un cierto porcentaje. Como todo análisis de regresión, la regresión logística es un análisis predictivo. Se utiliza para describir datos y explicar la relación entre la variable dependiente binaria y una o más variables independientes de nivel nominal, ordinal, de intervalo o de razón. En general, este algoritmo se puede usar para varios problemas de clasificación, como la detección de spam, la predicción de diabetes, si un determinado cliente comprará un determinado producto o participará en una competencia, hay muchos más ejemplos que se pueden usar.

3.4 Logistic Regression

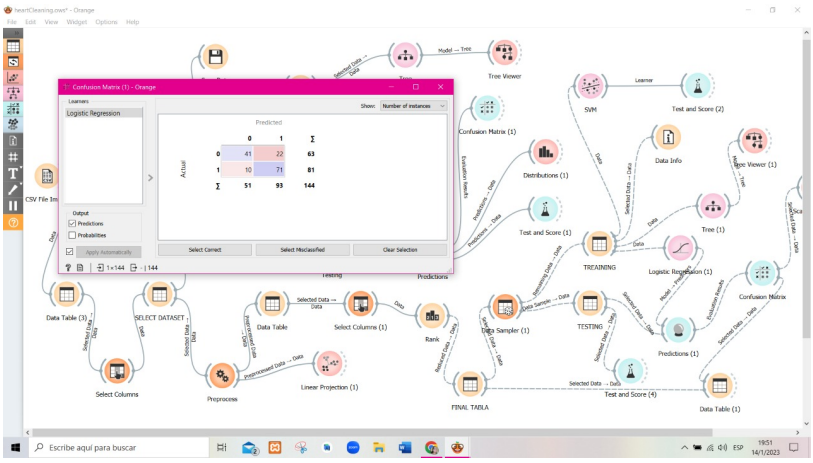

Figure 2: En esta imagen se muestran los datos seleccionados

4 Experimentos

En este apartado, se hace una comparación de los dos algoritmos utilizados y los efectos encontrados, estos algoritmos descritos en la parte de introducción, fueron cuidadosamente seleccionados, ya que se realizó una investigación exhaustiva, para así ampliar nuestros conocimientos a estos temas. utilizamos dos modelos uno: En cuanto al modelo tree nos permite evaluar mediante una representación gráfica los posibles resultados, costos y consecuencias de una decisión compleja. Este método es muy útil para analizar datos cuantitativos y tomar una decisión basada en números, estos resultados son mostrados mediante la gráfica tree viewer la cual muestra la caracteristica tomada en el modelo, obrsevando su clasificaicon. Mientras que el otro modelo Logistic regression al conectarlo con el test and score podemos observar que tiene una precisión del 82.1% en función a la variable que esta seleccionada. la regression lineal viene de la ecuación lineal la cual divide elementos en dos clases. Este modelo funciona perfectamente, siempre y cuando no se utilice data multiclase sólo data binaria. Los dos modelos que sirven para evaluar los datos, debemos conectarlos al test and score el cual es el que realiza las pruebas y determina la precisión. También con la matriz de confusión cual muestra el resultado obtenido por el modelo dependiendo la característica que se haya evaluado en el modelo entrenado.

4.1 Modelo Tree y Logistic regression entrenado y evaluados

Una vez entrenado el modelo y evaluado por el test and score observamos los resultados que son mostrados en las siguientes figuras.

4.2 Modelo Evaluado Tree

El modelo Tree al momento de visualizar los datos se escoge el gráfico Tree Viewer para verificar escogemos un dato, ya sea por edad, tipo de enfermedad, etc. Además muestra, si el dato es mayor se va para el lado derecho, si es menor para el izquierdo. El modelo arbol nos da varias oopciones sobre como queremos mostrar los datos, también especificar el número de instancia a utilizar, estos no se dividen, se escogen dependiendo la cantidad de trabajo ya sea mucho o

poco dependiendo el control que mantenga, es decir si no desea un arbol muy profundo. también la opción de parar el algritmo en un limite de control.

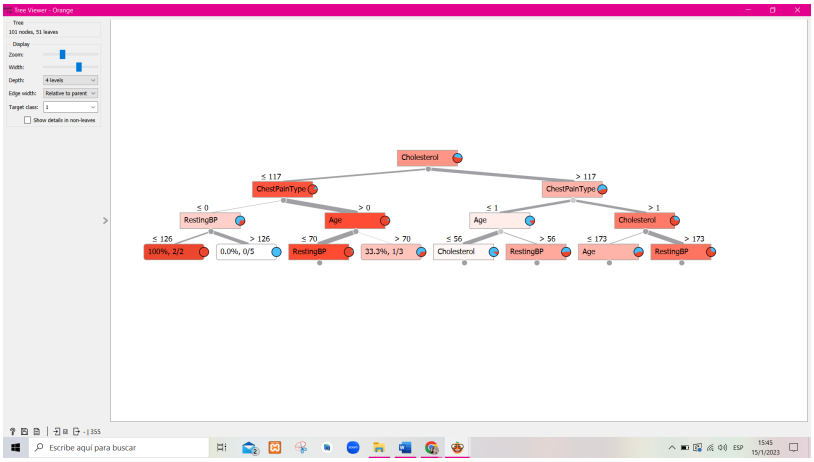

Figure 3: en esta imagen se observa la caracteristica de colesterol

4.3 Modelo Evaluado Logistic

Al entrenar este modelo para que sea evaluado nos ayuda a predecir si la persona es más propensa a sufrir problemas al corazón por su colesterol, edad, y las demás características, también cuánto tiempo va a permanecer con vida, en que tiempo le darán resultados de su enfermedad. Existe un motivo de confusión frecuente es la técnica de regresión logística. Su nombre podría inducirnos a pensar que puede usarse en problemas de regresión. Sin embargo, la regresión logística, sólo funciona para problemas de clasificación.

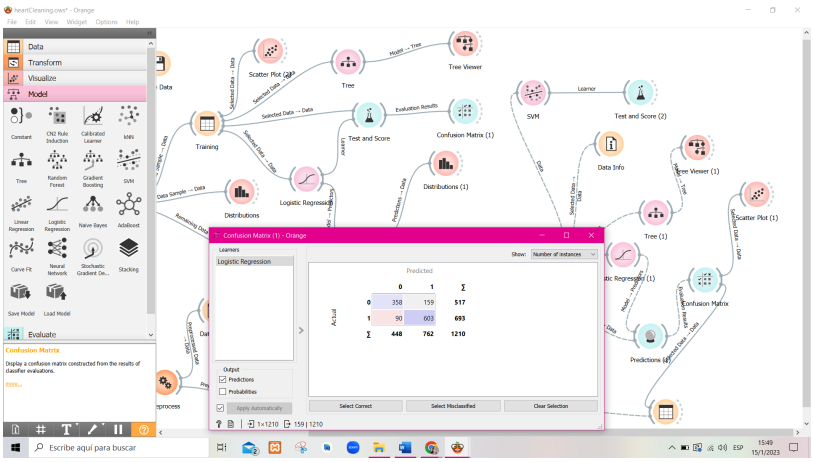

Figure 4: Se muestra la tabla una vez seleccionado y entrenado los datos se evaluan, de esa evaluación se observa en confunsion matrix los datos seleccionado y la precision de este modelo

5 Discusión

Después de los dos experimentos realizamos pruebas en cada uno de los dos modelos. No podemos decir cuál debe seleccionarse, pero hasta ahora está claro que el modelo de Logistic Regression que es el clasificador se desempeñó mejor en comparación al modelo TREE. Para tomar la decisión final graficamos en Tree Viewer muestra un árbol con las características escogidas por otra parte en el Logistic Regression seleccionamos ROC Analysis pero antes realizamos un test and score con el que se obtuvo un mejor resultado.

6 Conclusiones

En este artículo, se proponen dos modelos de aprendizaje automático el modelo Tree y logistic regression, el cual al momento de visualizar los datos se hacen en Tree Viewer y ROC analysis, matrix confusion para así verificar el dato escogido, se muestra también si el dato es mayor o menor y los clasifica según sus características, así mismo el modelo árbol nos dan varias opciones sobre cómo queremos mostrar nuestros datos. El modelo evaluado Logistic nos ayudó a predecir si la persona es más propensa a sufrir enfermedades del corazón, y esto lo hace por medio de las características de la persona, una vez realizada la práctica se observó que cuando seleccionamos y entrenamos los datos se evalúan y es de esa evaluación que se observa en una confusión matrix los datos seleccionados y la precisión que este modelo tiene. Una vez evaluado los dos modelos y del experimento respectivo, no se pudo decir cual debe seleccionarse, pero lo que si se tuvo muy en claro fue que el modelo Logistic Regression es el que tiene más desempeño en comparación al modelo Tree, para llegar a una conclusión final graficamos los modelos en Tree Viewer y muestra un árbol con las características escogidas y por otra parte el modelo Logistic Regression seleccionamos ROC Analysis pero antes realizamos un test and score con el que se obtuvo un mejor resultado.

7 Recomendaciones

Recomiendo cambiar los modelos de clasificación, es importante sean implementados en las demás carreras puesto que, usando estos algoritmos, tenemos la facilidad y el control de poder tener respuestas a nuestras dudas y así mismo aprender a manejar todo este tema. También al usar estos algoritmos nos hacen tener todo el conocimiento para poder adentrarnos en el uso de estas herramientas y comenzar a estudiar más a fondo todo lo que conlleva.

8 Bibliografia

References

[1] A. M. Vázquez, "Introducción a machine learning," 2018.

[2] A. T. Cantero, H. M. P. Meana, and M. Nakano-Miyatake, "Algoritmos de aprendizaje supervisado para la clasificación de géneros musicales caracterizados mediante modelos estadísticos.," *Res. Comput. Sci.*, vol. 147, no. 5, pp. 119–128, 2018.