Informe del Proyecto de Análisis y Modelado de Datos: HEALTHE CARE

Integrantes: Ramiro sturla, Martiniano lema, William Taylor, Juan Ignacio Gerber

Breve Introducción

El presente proyecto aborda el análisis y modelado de un conjunto de datos de atención médica utilizando la herramienta Orange Data Mining. El objetivo principal es predecir los resultados de las pruebas médicas (`Test Results`) de los pacientes, una tarea de crucial importancia en el ámbito de la salud. Se ha trabajado con un dataset que contiene información detallada sobre pacientes, admisiones hospitalarias y servicios de atención médica.

1. Definición del Problema

El problema de negocio o investigación que se busca resolver es la **predicción del tipo de resultado de las pruebas médicas (`Test Results`) de los pacientes**. Específicamente, se busca clasificar los resultados en "Normal", "Anormal" o "inconcluso".

La importancia de esta predicción radica en el impacto directo sobre la salud del paciente. En el contexto de la atención médica, no identificar una condición "Anormal" o "Inconclusiva" que realmente existe (un falso negativo) puede llevar a peores resultados para el paciente, ya que podría significar no detectar una enfermedad grave a tiempo. Por lo tanto, un modelo que pueda predecir con alta fiabilidad los resultados de las pruebas es fundamental para mejorar la detección temprana y la intervención médica.

2. Análisis Exploratorio de Datos

Se realizó un análisis exploratorio inicial para familiarizarse con la estructura y el contenido del dataset.

- **Variables Clave:** El conjunto de datos incluye variables como `Name`, `Age`, `Gender`, `Blood Type`, `Medical Condition`, `Date of Admission`, `Doctor`, `Hospital`, `Insurance Provider`, `Billing Amount`, `Room Number`, `Admission Type`, `Discharge Date`, `Medication`, y `Test Results`. La variable objetivo para la predicción es `Test Results`.
- **Distribución de la Variable Objetivo:** Se observó la distribución de los `Test Results`. El número de instancias para cada resultado es el siguiente: "Abnormal" cuenta con 18677 casos, "Inconclusive" con 10356, y "Normal" con 18517.
- **Hipótesis Inicial:** Se planteó la hipótesis de que el `Admission Type` (Tipo de Admisión) está directamente relacionado con los `Test Results`. La justificación es que un paciente que ingresa por "Emergency" (Emergencia) o "Urgent" (Urgencia) probablemente presenta un estado de salud más crítico o una condición aguda que justifique la urgencia. Esto podría correlacionarse con una mayor probabilidad de resultados "Anormales" o "inconclusos" en sus pruebas iniciales, en comparación

con admisiones "Elective" (Electivas) que suelen ser planificadas y para condiciones estables.
- **Visualización de la Hipótesis:** Gráficos de distribución y mosaicos fueron utilizados para explorar visualmente esta relación. Por ejemplo, la distribución de los resultados de las pruebas (`Test Results`) se puede desagregar por tipo de admisión (`Admission Type`) para observar si hay patrones visibles que respalden la hipótesis. Aunque una inspección visual en el gráfico muestra que las proporciones de "Abnormal", "Inconclusive" y "Normal" son relativamente similares entre los tipos de admisión, el mosaico de la fuente sugiere una relación más compleja, posiblemente influenciada por otras variables como la edad, y destaca la presencia de "Abnormal" e "Inconclusive" resultados en todas las categorías de admisión, con "Elective" mostrando una proporción visible de "Normal" en algunos segmentos de edad.

3. Preprocesamiento y Selección de Variables

Para preparar los datos para el modelado, se realizaron los siguientes pasos de preprocesamiento:

- **Selección de Columnas:** Se identificaron y seleccionaron las variables relevantes para el problema de predicción. Se ignoraron o excluyeron las columnas que no aportaban valor predictivo o eran metadatos, tales como `Hospital`, `Billing Amount`, `Name`, `Doctor`, y `Room Number`.
- **Definición de Características (Features) y Objetivo (Target):** Las variables restantes (`Age`, `Gender`, `Medical Condition`, `Admission Type`, `Insurance Provider`, `Blood Type`, `Medication`, `Date of Admission`, `Discharge Date`) fueron designadas como características (features), y `Test Results` como la variable objetivo (target).
- **Tratamiento de Variables Categóricas y Numéricas:** Se aplicó la transformación `one-hot encoding` a variables categóricas como `Gender` y `Admission Type` para que los modelos pudieran procesarlas adecuadamente. Las variables numéricas, como `Age`, se mantuvieron en su formato original (`Keep as it is`).
- **Muestreo de Datos:** Se utilizó un `Data Sampler` para dividir el dataset en conjuntos de entrenamiento y prueba. Se optó por una proporción fija de datos, asignando el 70% de las instancias para el entrenamiento (`Train Data`) y el 30% restante para la evaluación (`Test Data`), garantizando un muestreo replicable (determinístico).

4. Modelado

Se experimentó con varios modelos de aprendizaje automático disponibles en Orange Data Mining para predecir los `Test Results`. Se seleccionaron y entrenaron los siguientes modelos, utilizando el conjunto de `Train Data` y evaluándolos con el `Test Data`:

- **Random Forest:** Este es un algoritmo de conjunto que construye múltiples árboles de decisión y combina sus predicciones. Es conocido por su robustez y buen

rendimiento en problemas de clasificación, ayudando a mejorar la recall y reducir el sobreajuste.
- **Tree (Árbol de Decisión):** Un modelo que divide el espacio de datos en regiones más pequeñas, creando un árbol de decisiones basado en las características de entrada para llegar a una predicción de clase.
- **Naive Bayes:** Un clasificador probabilístico simple pero a menudo efectivo, basado en el teorema de Bayes. Asume una fuerte independencia entre las características, lo que lo hace rápido y eficiente, especialmente con grandes conjuntos de datos.

5. Evaluación del Modelo

La evaluación de los modelos fue un paso crítico, priorizando métricas que se alinearan con los requisitos específicos del problema de salud.

- **Métrica Principal - Recall (Sensibilidad):** A diferencia de otros contextos donde el costo de un falso positivo puede ser alto, en salud, el costo de un falso negativo (no identificar una condición "Anormal" o "Inconclusiva" que realmente existe) es significativamente mayor. Un falso negativo podría acarrear la omisión de una enfermedad grave, llevando a peores resultados para el paciente. Por lo tanto, la métrica principal de evaluación fue el **Recall (Sensibilidad o Tasa de Verdaderos Positivos)** para las clases "Anormal" e "inconcluso". El Recall mide la proporción de casos positivos reales que fueron correctamente identificados por el modelo.
- **Uso de la Matriz de Confusión:** La Matriz de Confusión (`Confusion Matrix`) fue una herramienta fundamental para evaluar el rendimiento de clasificación de cada modelo de manera granular, proporcionando una vista detallada para todas las clases simultáneamente. Para cada modelo, se analizaron los `Verdaderos Positivos (VP)`, `Falsos Negativos (FN)` y `Falsos Positivos (FP)` para las clases "Abnormal" e "Inconclusive".

A continuación, se presentan los resultados de Recall para las clases "Anormal" e "inconcluso" calculados a partir de las matrices de confusión de cada modelo:

- **Random Forest:**
  - Recall para "Abnormal": `VP / (VP + FN)` = 12175 / (12175 + 397 + 413) = 12175 / 12985 ≈ **0.9376**.
  - Recall para "Inconclusive": `VP / (VP + FN)` = 11985 / (11985 + 465 + 445) = 11985 / 12895 ≈ **0.9294**.
- **Tree (Árbol de Decisión):**
  - Recall para "Abnormal": `VP / (VP + FN)` = 12175 / (12175 + 397 + 413) = 12175 / 12985 ≈ **0.9376**.
  - Recall para "inconcluso": `VP / (VP + FN)` = 11031 / (11031 + 1395 + 469) = 11031 / 12895 ≈ **0.8554**.
- **Naive Bayes:**
  - Recall para "Abnormal": `VP / (VP + FN)` = 4625 / (4625 + 3577 + 4783) = 4625 / 12985 ≈ **0.3562**.
  - Recall para "inconcluso": `VP / (VP + FN)` = 4775 / (4775 + 4354 + 3766) = 4775 / 12895 ≈ **0.3703**.

Comparación y Selección del Modelo: Al comparar los resultados, se observa que el modelo **Random Forest** presenta el Recall más alto para ambas clases críticas ("Abnormal" e "inconcluso"). El modelo `Tree` tuvo un Recall excelente para "Abnormal" similar a Random Forest, pero ligeramente inferior para "inconcluso". Naive Bayes, en contraste, mostró un rendimiento considerablemente bajo en ambas clases, lo que lo hace inadecuado para este problema.

Dados los requisitos del problema de salud, donde minimizar los falsos negativos es la máxima prioridad, el modelo **Random Forest** es el más adecuado, ya que asegura la detección de la mayor proporción de casos con condiciones críticas.

Adicionalmente, se puede observar en las predicciones que Random Forest tiene un error menor en general comparado con Naive Bayes. El rendimiento general (`AUC`, `CA`, `F1`, `Prec`, `Recall`) para Random Forest fue de 0.987, 0.927, 0.927, 0.927, 0.927 respectivamente, mientras que Naive Bayes tuvo valores mucho más bajos.

6. Interpretación de las Predicciones

Se analizó la importancia de las variables para los modelos, lo cual permite entender qué características son más influyentes en las predicciones y cómo se relacionan con las hipótesis iniciales.

- **Importancia de las Características (Feature Importance):**
    - Para **Random Forest**, las variables más importantes en la predicción fueron `Date of Admission` (Fecha de Admisión), `Discharge Date` (Fecha de Alta), y `Age` (Edad). `Admission Type: Elective` también figura como una característica relevante.
    - Para **Logistic Regression**, las variables más importantes incluyeron `Medical Condition: Asthma` (Condición Médica: Asma), `Blood Type: AB-` (Grupo Sanguíneo: AB-), y nuevamente `Admission Type: Elective`.
    - Para **Naive Bayes**, `Date of Admission`, `Discharge Date`, `Age`, `Insurance Provider: UnitedHealthcare` y `Admission Type: Elective` fueron las características más influyentes.
- **Relación con la Hipótesis Inicial:** Nuestra hipótesis inicial planteaba que el `Admission Type` estaría relacionado con los `Test Results`. Los análisis de importancia de características para los tres modelos principales (`Random Forest`, `Logistic Regression`, `Naive Bayes`) confirmaron que `Admission Type: Elective` es una variable relevante en la predicción. Esto apoya la idea de que el contexto de la admisión del paciente es un factor influyente en los resultados de sus pruebas. Las fechas de admisión y alta, así como la edad, también mostraron ser muy influyentes en varios modelos, lo que sugiere una complejidad en la relación que va más allá de la hipótesis inicial simple.

7. Conclusiones

A lo largo de este proyecto, se ha desarrollado un modelo predictivo para los resultados de pruebas médicas utilizando un conjunto de datos de atención hospitalaria y la herramienta Orange Data Mining. El objetivo primordial fue identificar con alta fiabilidad los resultados "Anormales" o "Inconclusos" para prevenir falsos negativos que podrían tener consecuencias graves para la salud de los pacientes.

Se partió de la hipótesis de que el tipo de admisión (`Admission Type`) influiría en los resultados de las pruebas. Tras un riguroso análisis exploratorio, preprocesamiento de datos y modelado con Random Forest, Tree, y Naive Bayes (además de la evaluación de Logistic Regression), se determinó que **Random Forest es el modelo más adecuado** para este problema. Esto se basa en su superior rendimiento en el Recall para las clases "Abnormal" e "inconcluso" (aproximadamente 0.938 y 0.929 respectivamente), lo que asegura una minimización crucial de los falsos negativos, que son de alto riesgo en el contexto de salud.

La interpretación de las predicciones reveló que el `Admission Type`, especialmente la categoría "Elective", es una característica importante en la predicción de los `Test Results` para todos los modelos, lo que respalda parcialmente nuestra hipótesis inicial. Además, variables como la fecha de admisión, la fecha de alta y la edad emergieron como factores altamente influyentes, sugiriendo que la temporalidad y las características demográficas del paciente son cruciales para entender los resultados de las pruebas.

En retrospectiva, la importancia de elegir la métrica de evaluación adecuada (Recall en este caso) fue fundamental para el éxito del proyecto, dado el alto costo de los falsos negativos en salud. Este proyecto demuestra la capacidad de Orange Data Mining para llevar a cabo un análisis completo, desde la formulación del problema hasta la interpretación de modelos complejos, ofreciendo una solución valiosa para la detección temprana en el ámbito médico.