**Predicción del APGAR en Recién Nacidos: Análisis de Nacimientos en Medellín**

**INTEGRANTES: María Alexandra Vasco Lopera – ID: 000483569**

**Manuela Gómez Gallego - ID: 000447364**

1. **ENTENDIMIENTO DEL NEGOCIO**
   1. **DESCRIPCIÓN DEL NEGOCIO**

El Hospital General de Medellín es una institución en donde se atienden miles de partos cada año, por ende, este es una fuente clave de datos clínicos neonatales. Dentro de estos datos se incluyen diferentes indicadores para evaluar el estado clínico de un recién nacido como por ejemplo el APGAR el cual por medio de un puntaje determina la condición de un bebé al primer y quinto minuto de vida, este indicador por medio de su puntaje ayuda a identificar las posibles complicaciones o riesgos a los cuales el recién nacido puede estar expuesto para así brindarle la atención necesaria.

Por esta razón el hospital busca explotar el potencial de los datos históricos de anteriores nacimientos para identificar patrones del APGAR, para poder anticipar las condiciones clínicas de los recién nacidos, para así prevenir posibles complicaciones y mejorar la toma de decisiones.

* 1. **DESCRIPCIÓN DEL PROBLEMA**

El puntaje del indicador APGAR, específicamente el que se toma al minuto 5 de vida es crítico, ya que influye directamente a la adaptación del bebé al entrono extrauterino. En Medellín en la actualidad el puntaje para este indicador es tomado de manera manual, ósea, es tomado mediante observación clínica inmediatamente después del parto pues no se cuenta con una herramienta que automatice el resultado o que con base en factores previos y durante el parto permita anticipar los valores del indicador, o si el bebé puede tener o no complicaciones médicas, lo que hace que existan demoras en la toma de decisiones y de posibles intervenciones.

* 1. **OBJETIVOS DE LA MINERÍA**

Desarrollar un modelo de machine learning, el cual permita predecir el puntaje de del APGAR 2, empleando los datos históricos de nacimientos anteriores en el Hospital General de Medellín

Desarrollar un modelo de machine learning, el cual permita predecir si un recién nacido puede sufrir de complicaciones medicas teniendo en cuenta los datos históricos de nacimientos ocurridos en el Hospital General de Medellín

* 1. **DISEÑO DE SOLUCIÓN**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Problema** | **Tipo de Minería** | **Tipo de aprendizaje** | **Requerimiento datos** | **Métodos** | **Evaluación** |
| Predecir la condición del recién nacido a los 5 minutos de nacido (APGAR2) | Minería Predictiva | Aprendizaje Supervisado | Variables clínicas del parto y datos sociodemográficos del bebé, madre y padre | Regresión (si se predice el valor exacto de APGAR2) o Clasificación (si se agrupan los valores en categorías: bajo, normal, alto) | Métricas como MAE, RMSE (si es regresión) o Accuracy, F1-score, Matriz de confusión (si es clasificación) |

* 1. **RECURSOS PARA CREACIÓN DEL MODELO Y PARA DESPLIEGUE**
* Hw
* Sw

1. **ENTENDIMIENTO DE LOS DATOS (Datos específicos del problema)**
   1. **CICLO DE LOS DATOS:** Generación, Almacenamiento, Modificación (ruta), Periodicidad

* **Generación:** Los datos son generados inicialmente por el personal médico del Hospital General de Medellín a través del sistema de registro de nacimientos, durante el proceso prenatal, de parto, evaluación y análisis inmediata del recién nacido ya que cada registro tiene consigo tanto información sociodemográfica y clínica de los padres como información clínica del recién nacido.
* **Almacenamiento:** La información es almacenada y publicada en el portal oficial de datos abiertos del gobierno de Colombia, para uso de este trabajo el archivo se utilizó en formato csv y fue almacenado localmente para realizar el modelo
* **Modificación:** Durante la fase de preparación de los datos se realizaron las siguientes modificaciones:
* Se verificaron cuantas columnas duplicadas existían en el dataset para eliminarlas, en este caso solo existía una fila duplicada la cual fue eliminada.
* Se seleccionaron las variables más relevantes para la predicción, las cuales fueron: SEXO, PESO (Gramos), TALLA (Centimetros), TIEMPO DE GESTACION, NUMERO CONSULTAS PRENATALES, TIPO PARTO, MULTIPLICIDAD EMBARAZO, APGAR1, APGAR2 EDAD MADRE, NUMERO HIJOS NACIDOS VIVOS, NUMERO EMBARAZOS.
* Se crea un nuevo dataframe para no alterar al original.
* Se realiza la conversión de las columnas categóricas de tipo object a category
* Se realiza una limpieza de nulos, como se tiene se tiene un solo registro con sexo como indeterminado se reemplaza por un nan
* Se realiza un conteo por columna para verificar cuales tienen valores nulos, como APGAR2 es nuestra variable objetivo, se eliminan las filas en donde APGAR2 es nulo, ya que aun no podemos predecir esta variable, por otro lado podemos observar que tenemos datos nulos en las variables de SEXO, PESO (Gramos), TALLA (Centimetros), TIEMPO DE GESTACION y APGAR1, en donde el sexo se imputa por la moda, para el tiempo de gestación se usa KNN imputer basado en PESO y TALLA, las cuales son biológicamente relacionadas con el tiempo de gestación.
* Se crean nuevas variables (dummies), también se crea la variable INDICE\_MASA\_NEONATAL en donde se combinan las variables PESO y TALLA
* Se realizó también un análisis de correlación entre las variables en donde PESO y TALLA esta altamente correlacionadas por ende se eliminan ambas variables para evitar redundancias con la nueva variable.
* **Periodicidad:** El dataset generado por el Hospital General de Medellín, es actualizado periódicamente por los registros históricos de los nacimientos ocurridos, para este caso, la versión del archivo que se tiene contiene 10.038 registros, los cuales corresponden a un periodo acumulado
  1. **DICCIONARIO DE DATOS**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descripción** | **Tipo** |
| **SEXO** | Informa el sexo del recién nacido | Category |
| **PESO (Gramos)** | Informa sobre el peso del bebé al momento de nacer | Float |
| **TALLA (Centímetros)** | Informa sobre la talla del bebé al momento de nacer | Float |
| **TIEMPO DE GESTACION** | Informa sobre el tiempo de gestación que tuvo el bebé recién nacido | Float |
| **NUMERO DE CONSULTAS PREMATALES** | Informa sobre el número de consultas prenatales que tuvo la madre del recién nacido | Int |
| **TIPO PARTO** | Informa sobre el tipo de parto del recién nacido | Category |
| **MULTIPICIDAD EMBARAZO** | Informa sobre si existen dos o más fetos dentro de la madre | Category |
| **APGAR1** | Informa sobre el APGAR del nacido al minuto de nacido | Float |
| **APGAR2** | Informa sobre el APGAR del nacido a los cinco minutos de nacido | Float |
| **EDAD MADRE** | Informa sobre la edad de la madre a la fecha del parto | Float |
| **NUMERO HIJOS NACIDOS VIVOS** | Informa sobre el número de hijos nacidos vivos que ha tenido la madre incluido el presente | Int |
| **NUMERO DE EMBARAZOS** | Informa el número de embarazos que ha tenido la madre incluido el presente | Int |

* 1. **REGLAS DE CALIDAD**

|  |  |
| --- | --- |
| Variable | Regla calidad (valores válidos) |
| Edad | 18-80 |
| Sexo | F,M |

1. **PREPARACIÓN DE DATOS**
   1. **INTEGRACIÓN**

El conjunto de datos utilizado para este proyecto fue descargado del portal oficial de datos abiertos del Gobierno de Colombia, correspondiente a los nacimientos registrados en el Hospital General de Medellín. El archivo contenía datos históricos de nacimientos en formato .csv, por lo que no se requirió integración de múltiples fuentes.

Durante la etapa de integración, se realizaron las siguientes acciones para asegurar la coherencia y limpieza estructural inicial del conjunto de datos:

* **Revisión de duplicados exactos:** Se identificó una fila completamente duplicada, la cual fue eliminada para evitar sesgos o redundancias en el análisis.
* **Revisión de filas completamente vacías**: Se verificó que no existieran filas sin ningún dato (todos los campos nulos). El resultado fue cero filas completamente vacías, por lo cual no fue necesario aplicar acciones adicionales.
  1. **SELECCIÓN DE VARIABLES**

Se seleccionaron las variables con mayor relevancia clínica y estadística para la predicción del puntaje APGAR a los 5 minutos (APGAR2). Las variables iniciales seleccionadas fueron:

* SEXO
* PESO (Gramos)
* TALLA (Centímetros)
* TIEMPO DE GESTACIÓN
* NÚMERO DE CONSULTAS PRENATALES
* TIPO DE PARTO
* MULTIPLICIDAD DE EMBARAZO
* APGAR1
* EDAD MADRE
* NÚMERO DE HIJOS NACIDOS VIVOS
* NÚMERO DE EMBARAZOS

Para facilitar la limpieza y manipulación del conjunto de datos sin afectar la fuente original, se creó un subconjunto de trabajo a partir del DataFrame original solo con estas variables.

Esto permitió aplicar los pasos de limpieza exclusivamente para las variables relevantes, por consiguiente, se realizó la conversión de las variables categóricas, originalmente en formato object, al tipo category.

* 1. **DESCRIPCIÓN ESTADÍSTICA**

Para facilitar la limpieza y manipulación del conjunto de datos sin afectar la fuente original, se creó un subconjunto de trabajo a partir del DataFrame original, que incluyó exclusivamente las variables seleccionadas.

Se realizó la conversión de variables categóricas de tipo object a category para optimizar el manejo de datos.

Posteriormente, se llevó a cabo un análisis descriptivo inicial para entender la distribución de los datos:

• Cálculo del conteo y proporción de categorías para variables categóricas relevantes como SEXO, TIPO PARTO y MULTIPLICIDAD DE EMBARAZO.

• Estadísticas de tendencia central y dispersión para variables numéricas mediante describe(), evaluando medias, desviaciones estándar, mínimos y máximos.

• Visualización de distribuciones numéricas clave con histogramas y boxplots para identificar valores atípicos y sesgos.

• Análisis de frecuencias y detección de errores de codificación en variables categóricas, como en TIPO PARTO, que fue corregido manualmente.

• Observación de que la variable objetivo APGAR2 presenta una fuerte concentración en valores 9 y 10 (más del 85%), indicando que la mayoría de recién nacidos se encuentran en buena condición.

• Identificación de distribuciones aproximadamente normales para PESO y TALLA; concentración de TIEMPO DE GESTACIÓN entre 37 y 40 semanas con ligero sesgo; y asimetrías positivas en CONSULTAS PRENATALES, HIJOS NACIDOS VIVOS y EMBARAZOS.

• Confirmación de que la variable SEXO se distribuye equilibradamente entre masculino y femenino, con un caso aislado de valor indeterminado; predominancia del parto espontáneo y cesárea en TIPO DE PARTO; y mayoría de embarazos simples en MULTIPLICIDAD DEL EMBARAZO, acorde a expectativas clínicas.

* 1. **LIMPIEZA DE ATÍPICOS**

Se utilizaron gráficos tipo boxplot para identificar posibles valores atípicos en variables numéricas clave. Sin embargo, tras un análisis contextual, se concluyó que:

• En número de consultas prenatales, aunque hay valores altos, es clínicamente válido que algunas mujeres tengan muchas o pocas consultas según su situación particular.

• En APGAR1 y APGAR2, los valores bajos aparecen como outliers debido a la fuerte concentración en puntajes altos (9-10), pero son casos reales y clínicamente relevantes.

• La edad de la madre por encima de 40 años, aunque poco frecuente, es biológicamente posible y no representa un error.

• En número de hijos nacidos vivos y número de embarazos, los valores altos reflejan casos extremos pero válidos, considerando la distribución sesgada hacia valores bajos.

Para validar la coherencia multivariable, se analizaron específicamente los bebés con peso menor a 1000 gramos, cruzando con su talla y tiempo de gestación para detectar posibles inconsistencias. No se encontraron combinaciones anómalas que justificaran eliminación de registros.

Se evaluó además la relación entre edad de la madre y número de embarazos:

• Se identificaron 9 casos de madres menores de 18 años con más de 2 embarazos. Aunque infrecuente, es biológicamente posible.

• Se detectaron 4 casos con número de embarazos potencialmente incompatible con la edad (considerando inicio de fertilidad a los 12 años y un embarazo por año). Sin embargo, la literatura médica indica que el inicio precoz del ciclo menstrual puede justificar estos casos, por lo que no se descartaron.

* 1. **LIMPIEZA DE NULOS**
* SEXO: Se imputó utilizando la moda, ya que solo existía un registro con valor indeterminado.
* TIEMPO DE GESTACIÓN: Se aplicó KNN Imputer basado en las variables PESO y TALLA, debido a su relación biológica con el tiempo de gestación.
* APGAR1 y APGAR2: Se eliminaron las filas donde APGAR2 era nulo (22 casos), ya que es la variable objetivo y no se puede predecir sin su valor.
* PESO y TALLA: Los registros con valores nulos en estas variables coincidían con los eliminados por tener APGAR2 nulo.
  1. **CREACIÓN DE NUEVAS VARIABLES**

Se creó una nueva variable denominada Índice de Masa Neonatal (IMN), calculada como el peso del recién nacido dividido por el cuadrado de su talla (IMN = Peso / Talla²). Esta métrica es análoga al Índice de Masa Corporal (IMC) utilizado en adultos y permite evaluar de forma más estandarizada la constitución física del recién nacido, independientemente de su tamaño, facilitando una mejor interpretación clínica y predictiva.

* 1. **ANÁLISIS DE CORRELACIONES PARA REDUNDANCIA**

Se calculó la matriz de correlación para evaluar la relación entre variables numéricas. Se encontró que las variables PESO y TALLA presentan una alta correlación (r = 0.87). Dado que se había creado la variable Índice de Masa Neonatal (IMN) que combina ambas, se decidió eliminar las variables originales PESO y TALLA para evitar problemas de multicolinealidad y reducir la dimensionalidad del modelo sin perder información relevante.

* 1. **ANÁLISIS DE CORRELACIONES PARA IRRELEVANCIA (PREDICCIONES)**

Se evaluó la correlación entre las variables predictoras y la variable objetivo APGAR2 para identificar variables irrelevantes. Se estableció un umbral de correlación lineal de 0.05 (r < 0.05) para considerar una variable como irrelevante, en lugar de un umbral más alto como 0.1, por las siguientes razones:

• En contextos clínicos, incluso relaciones débiles pueden ser relevantes; sin embargo, un umbral más alto podría eliminar variables con aportes indirectos o no lineales importantes.

• El umbral de 0.05 permite conservar únicamente variables con al menos una ligera relación lineal con la variable objetivo, reduciendo el riesgo de incluir ruido en el modelo.

Con base en este criterio, se eliminaron las siguientes variables por su baja correlación con APGAR2:

• Edad madre

• Número de hijos nacidos vivos

• Número de embarazos

• Tipo de parto instrumentado

• Multiplicidad del embarazo (doble, simple, triple)

• Sexo masculino

Esta depuración busca mejorar la eficiencia del modelo sin sacrificar variables con valor clínico o predictivo significativo.

* 1. **REDUCCIÓN DE DIMENSIÓN (OPCIONAL EN PREDICCIONES)**

En este proyecto no se aplicaron técnicas adicionales de reducción de dimensión como Análisis de Componentes Principales (PCA) o selección automática de variables, ya que, tras la limpieza, creación de variables y eliminación de redundancias, el conjunto de datos quedó suficientemente optimizado para el modelado. Por lo tanto, esta etapa no fue implementada en el proceso de preparación de datos.

* 1. **BALANCEO (CLASIFICACIÓN)**

Dado que el problema planteado corresponde a una tarea de regresión para predecir el valor continuo del puntaje APGAR2, no fue necesario aplicar técnicas de balanceo de clases. Estas técnicas son propias de problemas de clasificación con clases desbalanceadas, lo cual no aplica en este caso.  
  
Por lo tanto, esta etapa no fue implementada en el proceso de preparación de datos.

* 1. **TRANSFORMACIONES**

Para preparar los datos para los modelos predictivos, se aplicaron las siguientes transformaciones:

• Escalado de variables numéricas: Se utilizó la estandarización con StandardScaler para normalizar las variables numéricas, asegurando que tengan media cero y desviación estándar uno. Esto es crucial para algoritmos sensibles a la escala como Support Vector Machines, K-Nearest Neighbors y redes neuronales.

• Codificación de variables categóricas: Las variables categóricas TIPO PARTO, MULTIPLICIDAD EMBARAZO y SEXO fueron transformadas a variables dummy mediante la función pd.get\_dummies de pandas para permitir que los modelos interpreten correctamente estas variables.

1. **MODELAMIENTO, EVALUACIÓN E INTERPRETACIÓN**
   1. **CONFIGURACIÓN MÉTODOS DE MACHINE LEARNING**

Para la predicción del puntaje APGAR a los 5 minutos (APGAR2), se seleccionaron y configuraron siete modelos de regresión supervisada, divididos en dos grupos:

* **Modelos supervisados individuales:**
  + Support Vector Regressor (SVR)
  + Multilayer Perceptron Regressor (MLP)
  + K-Nearest Neighbors (KNN)
  + Árbol de Decisión (Decision Tree)
* **Modelos de ensamble:**
  + Random Forest
  + Gradient Boosting
  + Bagging con Árboles de Decisión

Se utilizó validación cruzada tipo K-Fold con 10 particiones, con particionado estratificado, para evaluar el desempeño en el conjunto de entrenamiento. Cada modelo fue entrenado con los datos escalados y evaluado tanto en validación cruzada como en el conjunto de prueba independiente.

Se definió una función para automatizar la evaluación, calculando métricas clave y generando gráficos de predicción vs. valores reales.

* 1. **ANALISIS DE MEDIDAS DE CALIDAD**

Se emplearon las siguientes métricas para evaluar el desempeño de los modelos:

* **MAE (Mean Absolute Error):** error absoluto promedio entre las predicciones y valores reales. Indica cuán lejos se encuentra en promedio la predicción del valor real.
* **RMSE (Root Mean Squared Error):** raíz del error cuadrático medio. Penaliza más los errores grandes, reflejando la dispersión o variabilidad de los errores.
* **R² (Coeficiente de determinación):** proporción de la varianza en la variable objetivo explicada por el modelo. Un valor más alto indica mejor capacidad predictiva.

Los resultados en el conjunto de prueba mostraron:

* **MLP Regressor:** mejor desempeño global, explicando el 68% de la variabilidad (R² = 0.6795), con error promedio moderado (MAE = 0.31) y buena consistencia (RMSE = 0.50).
* **Gradient Boosting:** segundo mejor modelo (R² = 0.6398), con errores promedio bajos (MAE = 0.30) y buena robustez (RMSE = 0.53).
* **KNN Regressor:** menor error absoluto promedio (MAE = 0.30), aunque con errores más dispersos (RMSE = 0.54), explica un 62.6% de la variabilidad.
* **Random Forest y Bagging:** muestran buen desempeño, con R² en torno al 0.60, manteniendo un equilibrio entre precisión y estabilidad.
* **SVR:** desempeño similar a Bagging, con R² alrededor de 0.59 y error moderado.
* **Decision Tree:** menor desempeño (R² = 0.34), con errores mayores y menor capacidad para generalizar.
  1. **SELECCIÓN DEL MEJOR MODELO**
* **Comparación de calidad mediante pruebas estadística ANOVA, Tukey**
* **Tiempo computacional de despliegue (algunos casos creación)**

1. **DESPLIEGUE**
   1. PREDICCIÓN DE DATOS FUTUROS: almacenar modelo, pipes para el despliegue, servicio web/local
   2. MONITOREO
   3. CRONOGRAMA DE MANTENIMIENTO/RE-ENTRENAMIENTO