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

**INTEGRANTES: María Alexandra Vasco Lopera – Manuela Gómez Gallego**

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
  2. 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 alex

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

1. **PREPARACIÓN DE DATOS**
   1. INTEGRACIÓN
   2. SELECCIÓN DE VARIABLES
   3. DESCRIPCIÓN ESTADÍSTICA
   4. LIMPIEZA DE ATÍPICOS
   5. LIMPIEZA DE NULOS
   6. CREACIÓN DE NUEVAS VARIABLES
   7. ANÁLISIS DE CORRELACIONES PARA REDUNDANCIA
   8. ANÁLISIS DE CORRELACIONES PARA IRRELEVANCIA (PREDICCIONES)
   9. REDUCCIÓN DE DIMENSIÓN (OPCIONAL EN PREDICCIONES)
   10. BALANCEO (CLASIFICACIÓN)
   11. TRANSFORMACIONES
2. **MODELAMIENTO, EVALUACIÓN E INTERPRETACIÓN**
   1. CONFIGURACIÓN MÉTODOS DE MACHINE LEARNING
   2. ANALISIS DE MEDIDAS DE CALIDAD
   3. 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