"ANTECEDENTES MÉDICOS Y RIESGOS OBSTÉTRICOS: UN ENFOQUE PREDICTIVO PARA IDENTIFICAR COMPLICACIONES DURANTE EL PARTO"

"MEDICAL HISTORY AND OBSTETRIC RISKS: A PREDICTIVE APPROACH TO IDENTIFY COMPLICATIONS DURING PREGNANCY"

[1] Hernan Camilo Triana Gutierrez
Correo electrónico: hernan.triana@campusucc.edu.co
[2] Mateo Zuluaga Roncancio
Correo electrónico: @campusucc.edu.co

Resumen

En Casanare, Colombia, donde se registra anualmente una cantidad significativa de partos, se busca comprender que variables pueden afectar a la madre durante el parto. Atreves de técnicas de machine learning no supervisado como K-Means, DBSCAN, KModes, métodos para escalar los datos, Clasificación con KNN entre otras técnicas. La atención materna adquiere una importancia crucial, especialmente para aquellas mujeres que presentan antecedentes médicos riesgosos o han tenido hábitos poco saludables, como el consumo de alcohol o sustancias alucinógenas. Es esencial brindar un seguimiento especializado a estas madres, reconociendo que su historial de salud puede influir en el desarrollo del embarazo y complicaciones durante el parto.

Palabras claves: unsupervised machine learning - atención materna- complicaciones durante el parto.

Abstract

In Casanare, Colombia, where a significant number of births are registered annually, we seek to understand what variables can affect the mother during childbirth. Using unsupervised machine learning techniques such as K-Means, DBSCAN, KModes, methods to scale data, Classification with KNN among other techniques. Maternal care becomes crucially important, especially for those women who have a risky medical history or have had unhealthy habits, such as the consumption of alcohol or hallucinogenic substances. It is essential to provide specialized follow-up to these mothers, recognizing that their health history can influence the development of pregnancy and complications during childbirth.

Keywords:

unsupervised machine learning - maternal care - complications during childbirth.

1. INTRODUCCIÓN

En la región de Casanare, Colombia, donde cada año se registra un notable número de partos, la atención materna se erige como una prioridad fundamental. La comprensión de las variables que inciden en el proceso de parto, especialmente para aquellas mujeres con historiales médicos complejos o hábitos poco saludables, se ha convertido en un área crucial de investigación. Con el objetivo de abordar este desafío, se ha empleado un enfoque analítico robusto que involucra tanto técnicas de machine learning no supervisado, como K-Means, DBSCAN y KModes, como métodos supervisados, incluyendo regresión logística y k-NN para clasificación. La aplicación de estas herramientas se ha complementado con técnicas de escalado de datos, selección cuidadosa de variables y un análisis exploratorio detallado. Este estudio busca identificar patrones, evaluar riesgos y, en última instancia, mejorar la calidad de la atención materna en la región. La presente introducción sienta las bases para explorar cómo el análisis integral de datos puede contribuir significativamente a la comprensión y mejora de la salud obstétrica en Casanare.

1.1 RELEVANCIA Y MOTIVACIÓN DE LA INVESTIGACIÓN

Este proyecto se adentra en un análisis detallado de los casos de partos en Casanare, fusionando la metodología CRISP-DM con innovadoras técnicas de minería de datos. La exploración minuciosa de los datos busca descubrir elementos clave y conexiones reveladoras que ofrecen una visión profunda de los determinantes en los resultados perinatales. La metodología CRISP-DM es un marco estándar para el desarrollo de proyectos de minería de datos. en este proyecto desempeña un papel crucial al proporcionar una estructura organizada

1.2 DEFINICIÓN DEL PROBLEMA

En el ámbito de la atención materno-infantil en Casanare, se reconoce la importancia de comprender y anticipar los posibles desafíos y resultados en los casos de partos. El objetivo central es mejorar la calidad de la atención perinatal al analizar los datos recopilados de manera integral. Este análisis busca identificar patrones y factores clave que puedan influir en los resultados perinatales, considerando variables como antecedentes médicos, hábitos de las madres y otros indicadores relevantes.

La anticipación de complicaciones durante el parto basada en análisis previos permitirá implementar estrategias personalizadas para el seguimiento y cuidado de las madres, abordando áreas específicas de riesgo. Este enfoque proactivo no solo tiene como objetivo mejorar la experiencia individual de las madres durante el parto, sino también contribuir a la excelencia en la atención materno-infantil en la región de Casanare. Con este análisis predictivo, se busca garantizar una atención de calidad, prevenir complicaciones, y consolidar la reputación de la región como referente en el cuidado perinatal.

2.TRABAJOS RELACIONADOS

2.1 El análisis de [1] se centró en identificar los factores de riesgo asociados a complicaciones durante el parto en adolescentes embarazadas. Al evaluar expedientes clínicos en el Hospital General de la Zona 6, Ciudad Juárez, se encontró que el 49.8% de las adolescentes experimentaron complicaciones perinatales durante el embarazo. El estudio destacó la falta de control prenatal adecuado como un factor significativo de riesgo para tales complicaciones, subrayando la importancia de estrategias preventivas y una atención prenatal más efectiva en este grupo vulnerable.

El análisis de [2] se enfocó en identificar los factores de riesgo asociados a complicaciones durante el parto en mujeres de edad materna avanzada. Al revisar historias clínicas en el Hospital Central de la Zona 4, Ciudad [Ciudad de mexico], se reveló que el 52.3% de las mujeres de edad materna avanzada experimentaron complicaciones perinatales durante el embarazo. El estudio resaltó la avanzada edad materna como un factor significativo de riesgo para dichas complicaciones, enfatizando la necesidad de un monitoreo prenatal más riguroso y estrategias preventivas específicas para este grupo demográfico.

La falta de control prenatal adecuado también se identificó como un componente clave en el aumento de las complicaciones perinatales en mujeres de edad materna avanzada, subrayando la importancia de mejorar la atención prenatal y la conciencia sobre los posibles riesgos asociados con la edad materna avanzada

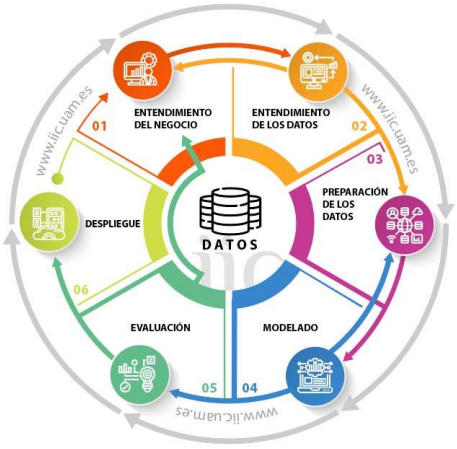
El análisis de [3] sobre el uso y abuso de drogas durante el embarazo se enfocó en la identificación de factores de riesgo asociados a complicaciones perinatales en mujeres gestantes que consumen sustancias psicoactivas. Al examinar historias clínicas en el Hospital General de la Zona 6, Ciudad Juárez, se reveló que el 37.2% de las mujeres embarazadas que abusaban de drogas experimentaron complicaciones durante el parto. El estudio resaltó la falta de atención prenatal adecuada como un factor significativo de riesgo para dichas complicaciones, enfatizando la necesidad de estrategias preventivas y una atención prenatal más efectiva para este grupo vulnerable. La investigación puso de manifiesto la urgencia de abordar el uso de sustancias durante el embarazo mediante enfoques integrales que incluyan tanto la prevención como la atención médica especializada para mitigar los riesgos asociados a esta problemática.

3.METODOLOGIA

3.1 justificación del enfoque de investigación y metodología CRISP-DM

En el estudio presentado, se adoptó una perspectiva cuantitativa para recuperar la información requerida y aplicar métodos de minería de datos con el propósito de evaluar el rendimiento y las métricas de los modelos. Se utilizó una metodología de desarrollo CRIPS-DM su objetivo principal es proporcionar una guía paso a paso que abarca desde la comprensión del problema hasta la implementación de soluciones.

3.2 PASOS DE CRISP-DM



Fases de la metodología CRIPS DM. Tomado de (1

La etapa de Preparación de Datos implicaría la limpieza y formateo de los datos para abordar posibles inconsistencias y asegurar su idoneidad para el análisis. Luego, en la fase de Modelado, se aplicarían técnicas avanzadas de minería de datos para descubrir patrones y relaciones en los datos perinatales, utilizando algoritmos de aprendizaje automático para identificar posibles riesgos y tendencias.

La Evaluación sería crucial para determinar la efectividad de los modelos y ajustarlos según sea necesario. Posteriormente, en la etapa de Implementación, se desarrollarían estrategias basadas en los resultados para mejorar la atención materno-infantil en la regió

La aplicación de la metodología CRISP-DM en el análisis de datos de partos en Casanare seguiría un enfoque estructurado y sistemático. En la fase inicial de Comprensión del Negocio, se definirían claramente los objetivos del análisis, centrándose en comprender los factores que influyen en los resultados perinatales. Posteriormente, en la etapa de Comprensión de los Datos, se explorarían y evaluarían las fuentes de datos disponibles, asegurándose de comprender la calidad y relevancia de la información recopilada

4.COMPRESIÓN DE LOS DATOS

El conjunto de los datos que vamos a trabar es de casos de parto en el municipio de Casanare, se obtuvo de datos abiertos

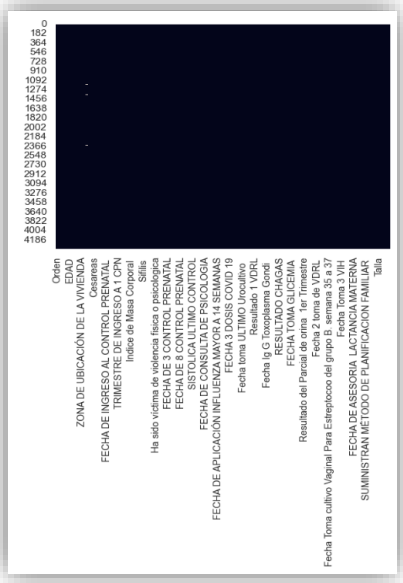
La tabla muestra los tipos de datos para cada columna del conjunto de datos con su respectiva descripción, en total hay 135 columnas, contiene etiquetas como riesgo obstétrico ultimo control prenatal, además de contar con variables categóricas y variables continuas como la edad.

Orden	int64
DEPARTAMENTO DE RESIDENCIA	object
NOMBRE IPS REPS	object
MUNICPIO DE RESIDENCIA	object
TIPO DE ID	object
EDAD	object
Ciclo_de_vida	object
SEXO	object
NACIONALIDAD	object
GRUPO POBLACIONAL	object
ZONA DE UBICACIÓN DE LA VIVIENDA	object
ESTADO ACTUAL DE USUARIA	object
DX1	object
Gestacions	object
Partos	object
Cesareas	object
Abortos	object
Muertos	object
Vivos	object
FECHA DE LA PRUEBA DE EMBARAZO	object
FECHA DE INGRESO AL CONTROL PRENATAL	object
FUM	object
EDAD GESTACIONAL A CORTE	object
FECHA PROBABLE DE PARTO POR FUR O ECO	object
EDAD GESTACIONAL AL INGRESO A CPN	object
TRIMESTRE DE INGRESO A 1 CPN	object
CAPTACION TEMPRANA	object
RIESGO OBSTETRICO AL INGRESO DEL CONTROL PRENATAL	object
Peso Inicial	object
Talla	object
Indice de Masa Corporal	object
Clasificación del IMC	object
Hipertension arterial	object
Diabetes	object
VIH	object
Sifilis	object
Tuberculosis	object
Apoyo familiar	object

Embarazo deseado	object
Hábitos de riesgo: FUMAR - SUSTANCIAS PSICOACTIVAS - ALCOHOL	object
Ha sido víctima de violencia fisica o psicologica	object
Ha sido víctima de abuso sexual	object
Información de causales para IVE de acuerdo a la Sentencia C-055-2022	object
FECHA DE 1 CONTROL PRENATAL	object
FECHA DE 2 CONTROL PRENATAL	object
FECHA DE 3 CONTROL PRENATAL	object
FECHA DE 4 CONTROL PRENATAL	object
FECHA DE 5 CONTROL PRENATAL	object
FECHA DE 6 CONTROL PRENATAL	object
FECHA DE 7 CONTROL PRENATAL	object
FECHA DE 8 CONTROL PRENATAL	object
FECHA DE 9 CONTROL PRENATAL	object
FECHA DE ULTIMO CONTROL	object
PESO EN EL ULTIMO CONTROL	object
RIESGO OBSTETRICO ULTIMO CONTROL PRENATAL	object
SISTOLICA ULTIMO CONTROL	object
DIASTOLICA ULTIMO CONTROL	object
FECHA DE CONSULTA POR NUTRICION	object
FECHA DE 1 CONSULTA DE ODONTOLOGIA	object
FECHA DE 2 CONSULTA DE ODONTOLOGIA	object
FECHA DE CONSULTA DE PSICOLOGIA	object
FECHA DE CONSULTA DE GINECOBSTETRICIA	object
ECOGRAFIA ENTRE LAS SEMANAS 10 SEMANAS + 6 DIAS Y 13 SEMANAS + 6 DIAS.	object
FECHA DE ECO DE DETALLE SEMANA 18 Y SEMANA 23 + 6 DIAS	object
FECHA ECO OBSTETRICA III TRIMESTRE	object
FECHA DE APLICACIÓN INFLUENZA MAYOR A 14 SEMANAS	object
FECHA DE APLICACIÓN TdaP ACELULAR MAYOR A 26 SEMANAS	object
NOMBRE BIOLOGICO CONTRA COVID 19	object
FECHA 1 DOSIS COVID 19	object
FECHA 2 DOSIS COVID 19	object
FECHA 3 DOSIS COVID 19	object
Fecha toma Hemograma	object
Resultado Hemograma	object
Fecha toma Urocultivo	object
Resultado Urocultivo con antibiograma	object
Fecha toma ULTIMO Urocultivo	object
Resultado ÚLTIMO Urocultivo con antibiograma	object
Fecha Toma 1 VIH	object
Resultado ELISA para detección de VIH	object
Fecha toma 1 VDRL	object
Resultado 1 VDRL	object
Fecha toma Ig G Rubeola	object
Resultado Ig G Rubeola	object
Fecha toma Ig M Rubeola	object
Resultado Ig M Rubeola	object
Fecha Ig G Toxoplasma Gondi	object
Resultado Ig G Toxoplasma Gondi	object

Fecha Ig M Toxoplasma Gondi	object
Resultado Ig M Toxoplasma Gondi	object
FECHA TOMA CHAGAS	object
RESULTADO CHAGAS	object
FECHA TOMA TSH	object
RESULTADO TSH	object
Fecha Toma Hemoclasificación	object
Resultado Hemoclasificación	object
FECHA TOMA GLICEMIA	object
REPORTE DE GLICEMIA	object
Fecha Toma Hepatitis B	object
Resultado Hepatitis B	object
Fecha toma de Parcial de orina	object
Resultado del Parcial de orina 1er Trimestre	object
Fecha toma Frotis vaginal	object
Reporte de Frotis de flujo vaginal	object
Fecha Toma prueba tolerancia a la glucosa	object
Reporte de Prueba de Tolerancia Oral a la glucosa	object
Fecha 2 toma de VDRL	Object
Resultado VDRL	object
Fecha de consejería para toma de 2 VIH	object
Fecha Toma 2 VIH	object
Resultado prueba VIH	object
Fecha Toma cultivo Vaginal Para Estreptocoo del grupo B. semana 35 a 37	object
Resusltado cutlivo vaginal	object
Fecha Toma 2 Hemograma	object
Resultado de Hemoglobina del 2 Hemograma	object
Fecha de consejería para toma de 3 VIH	object
Fecha Toma 3 VIH	object
Resultado Prueba VIH	object
Fecha toma de 3 VDRL	object
Resultado 3 VDRL	object
PLAN DE PARTO: VIA DE PARTO : VAGINAL-CESAREA	object
FECHA DE ASESORIA LACTANCIA MATERNA	object
Fecha de Parto (dia-mes-año)	object
Caracteristicas del parto VAGINAL-CESAREA	object
IPS ATENCION PARTO	object
No Semanas de gestación	object
SUMINISTRAN MÉTODO DE PLANIFICACION FAMILIAR	object
TIPO DE METODO	object
TIPO ID	object
Sexo	object
Peso al nacer	object
Talla	object
Toman TSH Neonatal	object
Resultado TSH	object
Fecha Vacunación con BCG	object
Fecha Vacunación hepatitis B	object
dtype	object

4.1 VISUALIZACION DE NULOS EN EL DATASET



DX1	34
Muertos	4
Vivos	2
FUM	1
FECHA DE CONSULTA POR NUTRICION	4
FECHA DE CONSULTA DE PSICOLOGIA	4

4.2 RESUMEN ESTADÍSTICO DE LAS COLUMNAS NUMÉRICAS DEL DATAFRAME

Visualizamos un resumen estadístico de las columnas numéricas. Proporciona varias estadísticas descriptivas que nos ayudan a entender la distribución de los datos en cada columna numérica.

	Orden
count	4359.000000
mean	2180.000000
std	1258.479241
min	1.000000
25%	1090.500000
50%	2180.000000
75%	3269.500000
max	4359.000000

4.3 PREPARACIÓN DE DATOS

Se verifico el conjunto de datos sobre los valores faltantes existentes en los registros y se encontraron valores nulos, se eliminaron duplicados para garantizar que no se presente registros repetidos de los datos. Se eliminaron columnas irrelevantes como sexo y fechas en las que se realizaron los exámenes algunos exámenes como: fecha vacunación hepatitis B, se unificaron datos de algunas columnas que contaban con valores similares como : Positivo y positivo para dejarlo con un solo valor, además, se corrigió el nombre de algunas variables y valores, para la etiqueta diabetes se realizó un if, el valor esta entre 100 y 125 va a tomar un valor de prediabetes y si es.

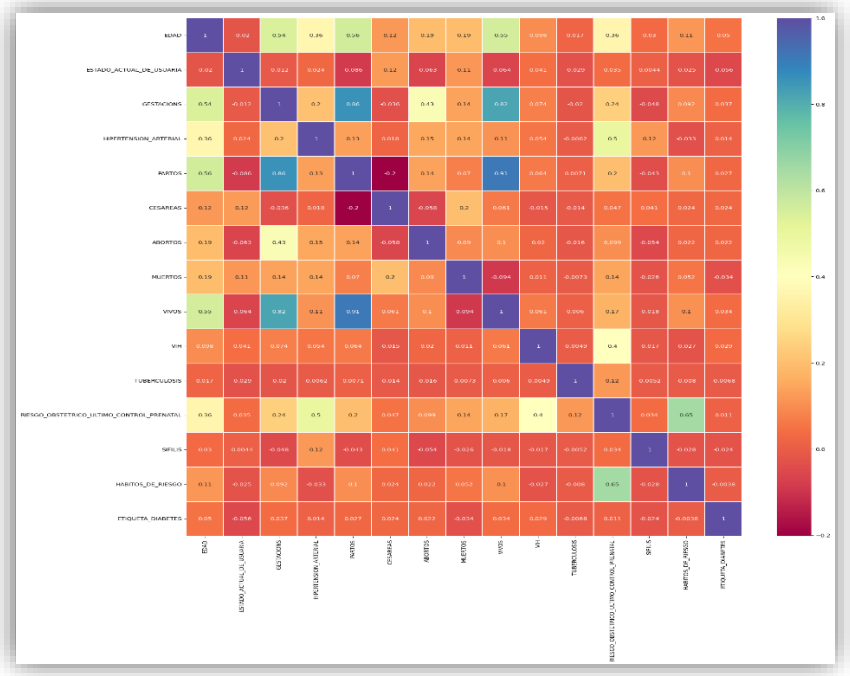
Valor de etiqueta	< 100 = normal
Valor de etiqueta	100 y 125 = prediabetes
Valor de etiqueta	>125 = diabetes

seguidamente hacemos selección de caracteristicas que son edad, estado actual de usuaria, gestacions, hipertension arterial, partos, cesareas, abortos, muertos, vivos, clasificacion del imc, diabetes, vih, sifilis, tuberculosis, riesgo obstétrico ultimo control prenatal, resultado vdrl, habitos de riesgo, etiqueta diabetes, un total de 15 columnas, seleccionamos estas columnas porque son los resultados de los exámenes médicos, igualmente para saber si han consumido alguna sustancia psicoactiva, y para sus gestaciones. Esto es fundamental ya que ayuda a los modelos a centrarse en los indicadores mas importantes en los que se esta trabajando, en este caso para predecir si van a tener complicaciones durante el parto, por último a algunas columnas le realizamos un mapeo a los datos como negativo 0 y positivo 1

.

0	EDAD	678 non-null	int64
1	ESTADO_ACTUAL_DE_USUARIA	678 non-null	int64
2	GESTACIONS	678 non-null	int64
3	HIPERTENSION_ARTERIAL	678 non-null	int64
4	PARTOS	678 non-null	int64
5	CESAREAS	678 non-null	int64
6	ABORTOS	678 non-null	int64
7	MUERTOS	678 non-null	int64
8	VIVOS	678 non-null	int64
9	VIH	678 non-null	int64
10	TUBERCULOSIS	678 non-null	int64
11	RIESGO_OBSTETRICO_ULTIMO_CONTROL_PRENATAL	678 non-null	int64
12	SIFILIS	678 non-null	float64
13	HABITOS_DE_RIESGO	678 non-null	int64
14	ETIQUETA_DIABETES	678 non-null	int64

4.4 MAPA DE CORRELACIÓN



4.5 NORMALIZACIÓN DE DATOS

Antes de ajustar los modelos de machine learning, se normalizan los datos para que las características tengan una media de 0 y una desviación estándar de 1, utilizando StandardScaler de Scikit-learn, Este proceso solo se le realizo a las columnas que lo requieren como, edad, gestaciones, partos, entre otras columnas.

Se recomienda utilizar StandardScaler en situaciones donde las características de entrada de un modelo de machine learning tienen diferentes escalas y se desea que todas tengan una

media de 0 y una desviación estándar de 1. Esto es particularmente útil para algoritmos que se ven afectados por la escala de las variables,

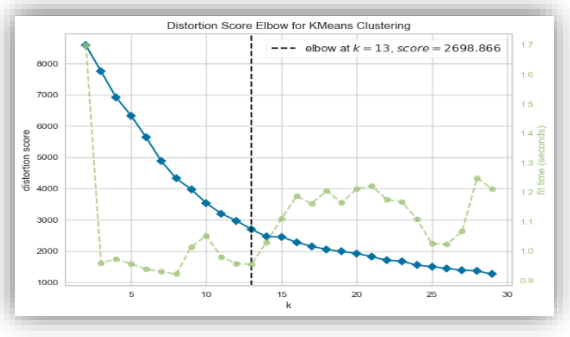
	EDAD	ESTADO_ACTUAL_DE_USUARIA	GESTACIONS	HIPERTENSION_ARTERIAL	PARTOS	CESAREAS	ABORTOS	MUERTOS	VIVOS	VIH
0	1.373205	0.425729	-0.507411	6.235572	-0.674871	-0.365367	-0.404932	-0.190893	-0.721349	-0.12842
1	0.278975	-0.752411	0.409988	-0.160370	1.042062	-0.365367	-0.404932	-0.190893	-0.721349	-0.12842
2	0.122657	0.425729	0.409988	-0.160370	-0.674871	5.019827	-0.404932	-0.190893	1.034754	-0.12842
3	-0.502617	0.425729	-0.507411	-0.160370	-0.674871	-0.365367	-0.404932	-0.190893	-0.721349	-0.12842
4	0.278975	-0.752411	-0.507411	-0.160370	-0.674871	-0.365367	-0.404932	-0.190893	-0.721349	-0.12842

Tras la aplicación de la técnica a los datos del conjunto, se observó una normalización efectiva de las variables, consiguiendo una media de 0 y una desviación estándar de 1, gracias a esto conseguimos un equilibrio en la data. Asi se evita que algunas características cuenten con más varianza que otras.

5. MACHINE LEARNING NO SUPERVISADO

El aprendizaje no supervisado en machine learning implica analizar datos no etiquetados para descubrir patrones y estructuras sin guía explícita. A diferencia del aprendizaje supervisado, no requiere

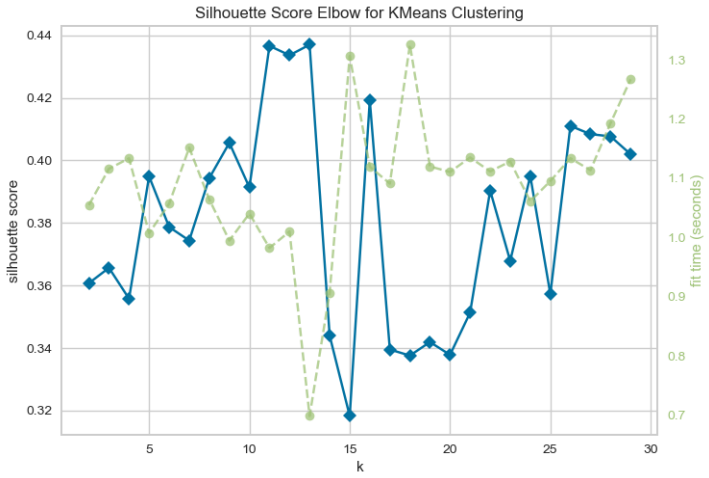
5.1 CODO DE PUNTUACIÓN DE DISTORSIÓN PARA LA AGRUPACIÓN DE KMEANS



etiquetas predefinidas. Este enfoque es esencial para tareas como la agrupación, detección de anomalías y reducción de dimensionalidad, permitiendo revelar información valiosa y realizar análisis exploratorios sin conocimiento previo de categorías

al elegir k = 13, se esta dividiendo los datos en 13 grupos de manera que los puntos dentro de cada grupo estén bastante cerca entre si.

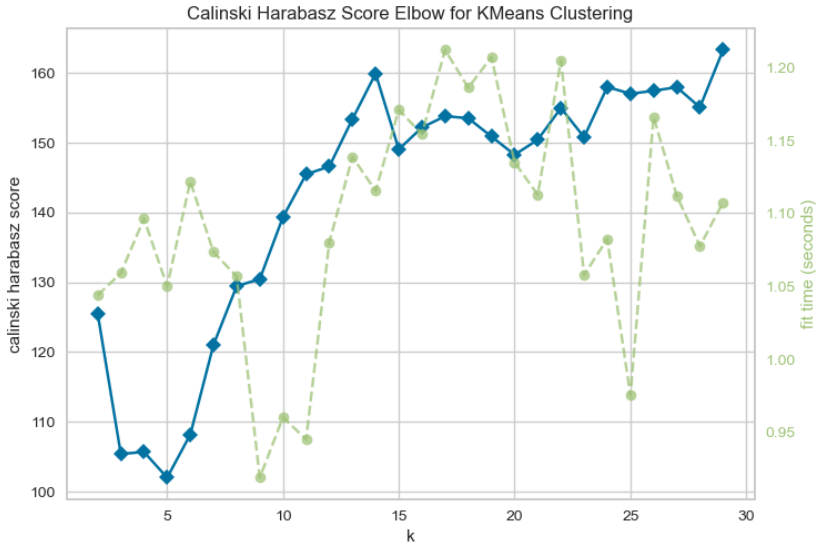
5.2 CODO DE PUNTUACIÓN DE SILUETA



"Optimal number of clusters based on silhouette score: [13]"

significa que, según la métrica de silueta, el número óptimo de clústeres para tus datos es 13.

5.3 MÉTRICA DE CALINSKI-HARABASZ



"Optimal number of clusters based on calinski harabasz: [29]"

significa que, según la métrica de Calinski-Harabasz, el número óptimo de clústeres para tus datos es 29.

5.4 MÉTRICAS DE VALIDACIÓN DE CLÚSTER

```
Silhouette Score: 0.3558
Calinski Harabasz Score: 105.6609
Davies Bouldin Score: 1.2028
```

Silhouette Score (Puntaje de Silueta):

Valor entre -1 y 1.

Cuanto más cercano a 1, mejor. Indica una buena separación entre clústeres.

Cuanto más cercano a -1, peor. Indica que las muestras pueden haber sido asignadas al clúster incorrecto.

Un Silhouette Score de 0.3558 sugiere que hay una separación significativa entre los clústeres.

Calinski Harabasz Score:

Cuanto mayor sea el puntaje, mejor.

Indica una mejor cohesión entre clústeres y una mayor separación entre ellos.

Se compara con otros valores posibles de k para determinar el número óptimo de clústeres.

Un Calinski Harabasz Score de 105.6609 sugiere que la separación entre clústeres es relativamente buena.

Davies Bouldin Score:

Cuanto menor sea el puntaje, mejor.

Indica una mejor separación y cohesión entre clústeres.

Un valor cercano a cero indica una buena partición.

Un Davies Bouldin Score de 1.2028 sugiere que hay una buena separación y cohesión entre los clústeres.

los resultados indican que la agrupación en 4 clústeres tiene una moderada a buena calidad de separación entre los clústeres según estas métricas

5.5 DBSCAN

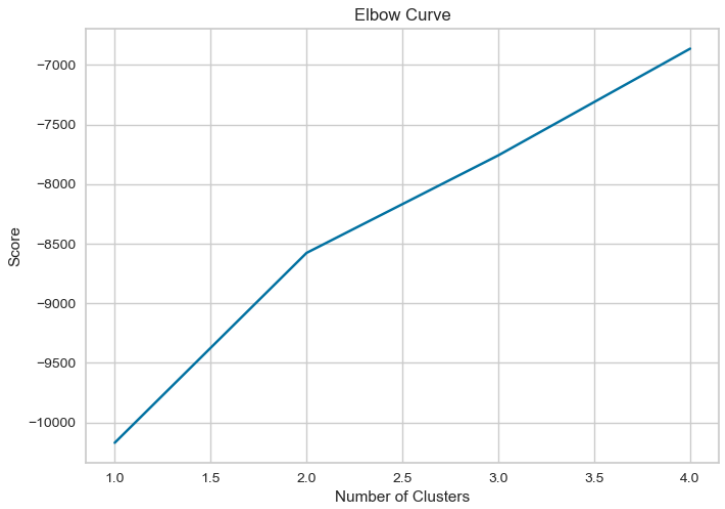
Se realiza una evaluación cuantitativa de la calidad de los clusters producidos por DBSCAN en términos de la métrica de silueta. La puntuación de silueta promedio se imprime en la consola como medida de la cohesión y separación de los clusters.

```
Silueta promedio (DBSCAN): 0.027780352332724616
```

sugiere que los clusters generados por el algoritmo DBSCAN tienen una calidad relativamente baja, ya que la separación entre los clusters es limitada en comparación con el solapamiento o la falta de estructura en los datos.

No supervisado	Situeta
KMeans	0.3558
DBSCAN	0.0277

5.5 MÉTODO DEL CODO

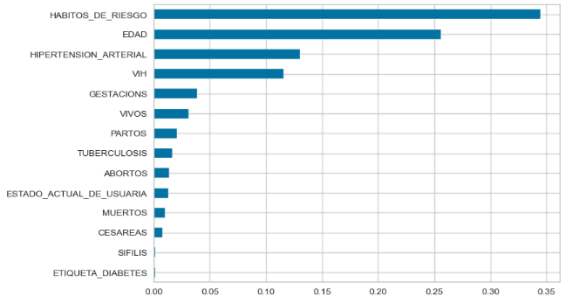


Por lo tanto, k=2 podría considerarse como el número óptimo de clústeres según el método de codo en tu conjunto de datos

6 RANDOM FOREST

RandomForestClassifier es una implementación del algoritmo Random Forest para clasificación en scikit-learn, y se utiliza para construir un conjunto de árboles de decisión y realizar predicciones basadas en el voto mayoritario de esos árboles.

Las características con puntuaciones de importancia más altas son más influyentes en el modelo.



7. REDUCCIÓN DE CARACTERÍSTICAS

La reducción de características, y en particular el Análisis de Componentes Principales (PCA), es una técnica crucial en el ámbito del aprendizaje automático. Esta estrategia permite simplificar conjuntos de datos al transformar las variables originales en un conjunto más pequeño de componentes principales, manteniendo la mayor parte de la información relevante

```
Using 4 components, we can explain 0.9877835985274785% of the variability in the original data.
```

En este caso, al reducir la dimensionalidad a 4 componentes principales, todavía se mantiene una proporción muy alta de la información original.

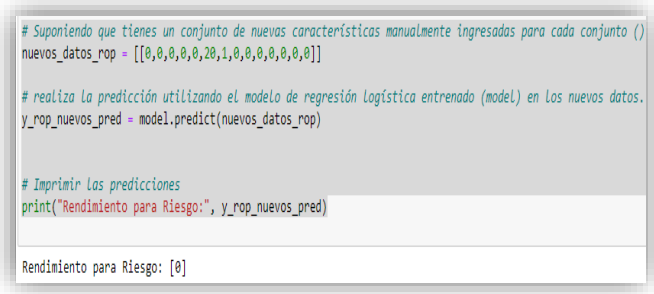
Esto sugiere que estos 4 componentes principales contienen la mayor parte de la variabilidad esencial en los datos originales y pueden ser considerados representativos del conjunto de datos.

8.APRENDIZAJE SUPERVISADO

El aprendizaje supervisado es un paradigma de aprendizaje automático en el que se enseña a un modelo utilizando un conjunto de datos etiquetado. En este enfoque, el algoritmo aprende a mapear las entradas a las salidas basándose en ejemplos de entrenamiento que incluyen tanto las características de entrada como las etiquetas correspondientes.

8.1 REGRESIÓN LOGÍSTICA

Se implemento un modelo de regresión lo gística para realizar una tarea de clasificac ión supervisada. Utiliza datos de entrada (X) que son: hipertension_arterial, vih, tub erculosis, sifilis, habitos_de_riesgo, edad, estado_actual_de_usuaria, gestacions, part os, cesareas, abortos, muertos,vivos, etiqu eta_diabetes (y) riesgo_obstetrico_ultimo _control_prenatal para entrenar el modelo, y luego evalúa su rendimiento en un conju nto de prueba. La precisión del modelo es del 98.53%. Esto significa que el 98.53% de las predicciones del modelo en el conju nto de prueba fueron correctas



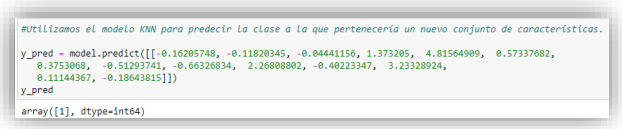
```
# Suponiendo que tienes un conjunto de nuevas características manualmente ingresadas para cada conjunto ()
nuevos_datos_rop = [[0,0,0,0,0,0,20,1,0,0,0,0,0,0,0,0]]

# realiza la predicción utilizando el modelo de regresión logística entrenado (model) en los nuevos datos.
y_rop_nuevos_pred = model.predict(nuevos_datos_rop)


# Imprimir las predicciones
print("Rendimiento para Riesgo:", y_rop_nuevos_pred)
```

```
Rendimiento para Riesgo: [0]
```

Realizamos una prueba de predicción, pro porcionamos los valores que nos solicita l a regresión, y nos arroja un resultado exito so sobre la predicción, 0 indicando que no tiene ningún riesgo y efectivamente le pro porcionamos valores de ningún riesgo.

8.2 CLASIFICACION DEL KNN

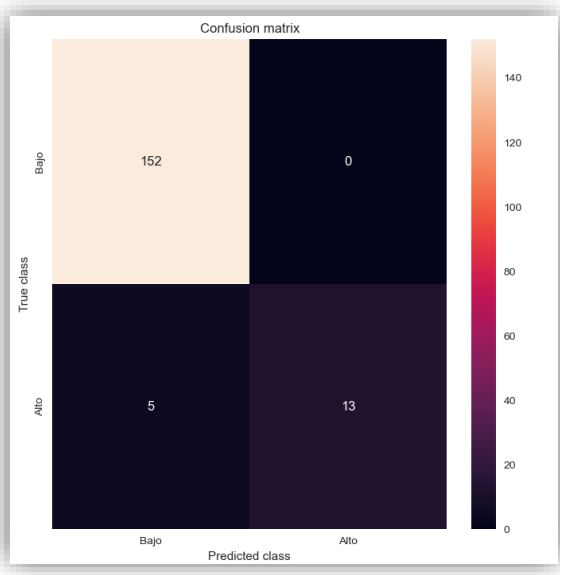
Se utiliza un conjunto de datos que incluye diversas características relacionadas con la salud obstétrica. Inicialmente, las columnas relevantes son seleccionadas y divididas en conjuntos de entrenamiento y prueba. Luego, se estandarizan las características para garantizar una escala consistente. Se crea y entrena un clasificador k-NN con 5 vecinos y la métrica de distancia euclidiana. Posteriormente, se utilizan las características del conjunto de prueba para realizar predicciones, y la precisión del modelo se evalúa mediante la métrica de precisión. Se genera y visualiza la matriz de confusión para analizar el rendimiento del clasificador en la clasificación binaria. Además, se utiliza el modelo entrenado para predecir la clase a la que pertenecería un nuevo conjunto de características ingresadas manualmente, le vamos a asignar valores de una mujer con problemas de hipertensión arterial, nos tiene que arrojar que la madre tendrá problemas al momento del parto ósea 1.



```
#Utilizamos el modelo KNN para predecir la clase a la que pertenecería un nuevo conjunto de características.

y_pred = model.predict([[-0.16205748, -0.11820345, -0.04441156, 1.373205,  4.81564909,  0.57337682,
   0.3753068,  -0.51293741, -0.66326834,  2.26808802, -0.40223347,  3.23328924,
   0.11144367, -0.18643815]])
y_pred
```

```
array([1], dtype=int64)
```

El resultado es satisfactorio, ya que nos indica que la mujer va a tener complicaciones al momento de dar a luz.



También generamos una matriz de confuncion y nos indica que

Verdaderos Negativos (TN):

152 instancias que fueron correctamente clasificadas como negativas.

Falsos Positivos (FP):

0 instancias que fueron incorrectamente clasificadas como positivas.

Falsos Negativos (FN):

5 instancias que fueron incorrectamente clasificadas como negativas

'pickle_model.pkl'. Posteriormente, el código carga este modelo guardado y calcula el accuracy promedio utilizando los datos de prueba.

```
# Encontramos el accuracy promedio usando datos de test
score = model.score(X_test, y_test)
print(score)
```

```
0.9926470588235294
```

```
x_in = np.asarray([1.373205, 0.425729,-0.507411,-0.674871, -0.365367,-0.404932,-0.190893, -0.721349,-0.178029, 1 , 0, 0, 0,0]).
predicts = model.predict(x_in)
predicts[0]
```

```
1
```

Verdaderos Positivos (TP):

13 instancias que fueron correctamente clasificadas como positivas.

8.3 PREDICION STREAMLIT

Se realiza un proceso completo de construcción, entrenamiento y evaluación de un modelo de Máquinas de Soporte Vectorial (SVM) para clasificación.

Se seleccionan columnas relevantes, tanto categóricas como numéricas, del conjunto de datos original y creando un nuevo DataFrame llamado ds_pred. Luego, este conjunto de datos se guarda en un archivo CSV para su posterior uso. Después de cargar los datos, se dividen en conjuntos de entrenamiento y prueba. Se crea un modelo SVM con un kernel lineal y se entrena utilizando los datos de entrenamiento. El modelo entrenado se guarda en un archivo .

Finalmente, se realiza una predicción utilizando un conjunto de nuevas características, demostrando así el proceso completo de construcción, entrenamiento y aplicación de un modelo de clasificación SVM en un contexto supervisado.

Aprendizaje supervisado	Accuracy
Regresion logistica	98.53
KNN	97.05
SVC	99.26

Streamlit

Para poder consumir el modelo en Streamlit es necesario usar estos códigos en anaconda prompt:

```
(base) C:\Users\hernan>$   conda create -n ApiCrop
```

Se utiliza para crear un nuevo entorno virtual en Conda (un gestor de paquetes y entornos para Python). El nombre del nuevo entorno es "ApiCrop". Este entorno virtual nos permite aislar las dependencias y configuraciones del proyecto.

```
(base) C:\Users\hernan>$   conda activate ApiCrop
```

Se utiliza para entrar en el entorno virtual llamado "ApiCrop". Una vez activado, se puede trabajar en ese entorno de forma aislada, instalando y ejecutando paquetes de Python sin afectar otros entornos o el sistema global.

```
(base) C:\Users\hernan>$   conda install python
```

Se usa para instalar la versión predeterminada de Python en el entorno virtual actual de Conda.

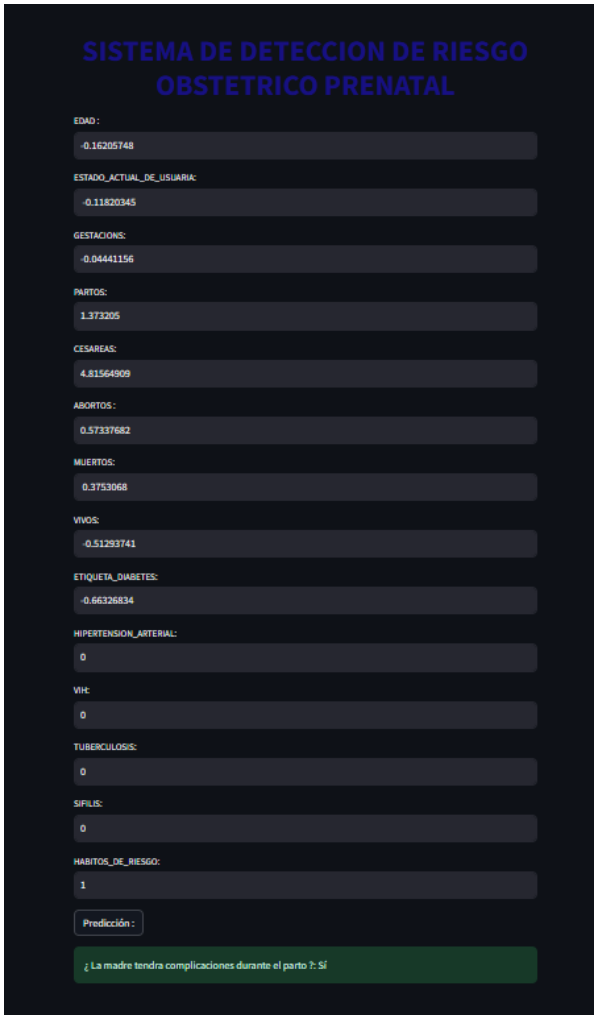
```
(base) C:\Users\hernan>$   pip install -r requirements.txt
```

Se usa para instalar las dependencias de un proyecto de Python, especificadas en un archivo llamado requirements.txt. Pip, el gestor de paquetes de Python, lee este archivo y procede a instalar cada paquete y sus versiones correspondientes

```
(base) C:\Users\hernan>$   streamlit run app.py
```

se utiliza para ejecutar una aplicación web desarrollada con Streamlit en Python. Streamlit es un marco de desarrollo rápido para crear aplicaciones web interactivas

Consumir el modelo en Streamlit



Como se puede visualizar esta funcionando correctamente el modelo consumido por **Streamlit**

CONCLUSIONES

En el contexto de Casanare, Colombia, donde la tasa de partos anuales es significativa, se ha llevado a cabo un análisis integral utilizando diversas técnicas de machine learning tanto supervisadas como no supervisadas. A través de métodos como K-Means, DBSCAN, y KModes, se ha buscado identificar patrones y estructuras en los datos relacionados con la salud obstétrica. La aplicación de técnicas de escalado de datos y la selección cuidadosa de variables, junto con un análisis exploratorio detallado, ha contribuido a la preparación eficaz de los datos. En el ámbito de aprendizaje supervisado, se implementaron modelos de regresión logística y k-NN para clasificación, destacando la importancia de seguir de cerca a las madres con historiales médicos riesgosos o hábitos poco saludables durante el embarazo. Se subraya la relevancia de proporcionar atención especializada a estas mujeres, reconociendo que su estado de salud puede influir de manera significativa en el desarrollo del embarazo y en las posibles complicaciones durante el parto. Además, la persistencia de modelos entrenados permite una aplicación futura, y la preocupación por la atención materna destaca la importancia de abordar factores de riesgo específicos para mejorar la salud materno-infantil en la región. Este enfoque integrador y analítico se presenta como una herramienta valiosa para la toma de decisiones informada y la mejora continua de la atención obstétrica en Casanare.

BIBLIOGRAFIA

[1] Accedido el 20 de noviembre de 2023. [En línea]. Disponible: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-56332022000200123

[2] Accedido el 20 de noviembre de 2023. [En línea]. Disponible: https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1870-72032018000200125

[3] "Uso y abuso de drogas durante el embarazo". SciELO - Scientific Electronic Library Online. Accedido el 20 de noviembre de 2023. [En línea]. Disponible: http://scielo.iics.una.py/scielo.php?script=sci_arttext&pid=S1812-95282009000200006

[4] Accedido el 20 de noviembre de 2023. [En línea]. Disponible: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-53072012000100011