Trabajo de grado en modalidad de aplicación

Diseño de un aplicativo para la identificación del estado de riesgo biopsicosocial en madres gestantes

María Alejandra Álvarez[a,c] , Daniel Andrés Buitrago[a,c] , María Fernanda Roa[a,c], Juan José Tapia[a,c]

Camilo Ernesto Martínez[b,c]

[a]*Estudiante de Ingeniería Industrial*
[b]*Profesor, Director del Proyecto de Grado, Departamento de Ingeniería Industrial*
[c]*Pontificia Universidad Javeriana, Bogotá, Colombia*

Summary

Fetal death in Colombia is considered a problem of high impact on public health, because it affects the health of the mother and produces emotional effects for her and her family. Over the years, different studies have been carry out that have shown that biopsychosocial factors have an impact on fetal death and that it is of vital importance to treat them throughout pregnancy. This work seeks to create a predictive model that identifies these variables and allows early identification of such negative conditions in the health of the mother. Use the Auto Machine Learning modeling method to agilely develop models similar to classification problems. For this, a Python tool called TPOT was implemented, which automatically creates and optimizes the automatic learning pipes through genetic programming, to find the best parameters and sets of algorithms. For the development of the model, the databases of birth and fetal deaths of DANE were used, from 2007 to 2016, and the result was Gradient Boosting, which showed as a result an accuracy of 80.7%. Finally, to identify the biopsychosocial risk status in mothers who periodically attend to prenatal control, an application was developed that works through an interface, which allows the doctor to know if the mother presents a risk or not.

Palabras claves: Biopsychosocial risk, Auto Machine Learning, TPOT, genetic programming, Gradient Boosting.

1. Justificación y planteamiento del problema

Según la Organización Mundial de la Salud, la muerte fetal es la defunción del producto de la concepción antes de su expulsión o extracción completa del cuerpo de la madre, independientemente de la duración del embarazo. (DANE, 2017). La muerte del feto y del recién nacido, están asociadas a diversos factores relacionados a los genes y al entorno de la madre (Huiza, Pacora, Ayala & Buzzio, 2003). Debido a esto, el análisis del proyecto se realizará sobre riesgos que abarcan factores biológicos, psicológicos y sociales que afectan el proceso de gestación. Estos se denominan: riesgos biopsicosociales. Cuando la paciente presenta dichos riesgos durante el embarazo, puede desencadenar en ella altos niveles de estrés y ansiedad de forma continua, ocasionando complicaciones en su proceso prenatal en aspectos como: aumento de la resistencia vascular, resistencia a la insulina, producción de citosinas proinflamatorias, entre otros. Causando finalmente preclamsia, parto prematuro, bajo peso o, en la peor circunstancia, muerte fetal (Herrera, Ersheng, Shahabuddin, Lixia, Wei, Faisal, Barua, & Akhtner, 2009).

En Colombia, la muerte fetal es considerada un problema de alto impacto en la salud pública, donde los riesgos biopsicosociales son factores que influyen en su desencadenamiento. Esto para las madres, además de presentar implicaciones en su salud, le genera efectos emocionales asociados al duelo posterior de la pérdida. Según las estadísticas vitales del DANE, para el 2016 se reportaron 223.078 muertes de las cuales 48.619 fueron fetales, lo que representa un 21,57% de la totalidad. Para el primer semestre de 2017, ya se llevaban registradas 132.504 muertes de las cuales 29.388 fueron muertes fetales (DANE, 2017). En la ciudad de Bogotá, para el año 2016, se presentaron 9893 muertes con esta patología. Con lo anterior se evidencia que la muerte fetal es representativa dentro de los indicadores de mortalidad en Colombia.

A nivel nacional, el ministerio de salud colombiano ha desarrollado diferentes sistemas y herramientas para reducir la mortalidad fetal en relación con factores biopsicosociales, como guías de práctica clínica para la prevención, detección temprana y tratamiento de las complicaciones de embarazo, parto o puerperio (Ministerio de Salud, 2013). Las guías de práctica clínica se implementan dentro del control prenatal al que está sujeta la madre durante su embarazo. Este control es el conjunto de actividades que se realizan en la mujer embarazada con el objetivo de lograr su buena salud, el desarrollo natural del feto y un recién nacido en óptimas condiciones a nivel físico, mental y emocional (Colombiana de Salud, 2016).

En Colombia, el control prenatal parte de una consulta de inscripción e identificación de la gestante, seguido de una consulta de primera vez, la cual evalúa las condiciones de salud de la madre. Su propósito es identificar los factores de riesgo biopsicosociales y enfermedades propias de la gestación y tiene una duración de 40 minutos. Posterior a esto, la madre debe asistir a consultas de seguimiento, mensualmente hasta la semana 36 y luego cada 15 días hasta el día del parto, estas tienen una duración de 20 minutos (Colombiana de Salud, 2016). Debido a que en la consulta de seguimiento se deben realizar revisiones tanto físicas como biopsicosociales, el tiempo estimado de la consulta debería ser suficiente para brindar un diagnóstico integral. Sin embargo, la mayoría de las EPS del régimen contributivo y subsidiado, no cumplen con este tiempo ya que las citas no duran más de 15 minutos. Esto debido a que el tiempo genera un costo representativo dentro de la institución, poniendo a los médicos en el grave riesgo del error médico con las correspondientes consecuencias adversas para la salud del paciente (Aguirre & Vázquez, 2016).

Se debe tener en cuenta, según la información anteriormente mencionada, la importancia de identificar de manera correcta y adecuada todos los factores de riesgo que pueda presentar la madre, sin dejar de lado los riesgos biopsicosociales. Ya que una alteración bioquímica no siempre se traduce en enfermedad, esta aparece por la interrelación de diversas causas, no solo moleculares, sino también psicológicas y sociales. Es decir, que de las alteraciones biopsicosociales pueden derivarse diferentes enfermedades y por esto deben tratarse con la misma importancia. Las variables de índole psicosocial suelen ser importantes a la hora de determinar la susceptibilidad, gravedad y curso del padecimiento más biológico que pudiera considerarse (Borrell, 2002).

Por lo anterior, el propósito del proyecto es desarrollar un aplicativo que permita identificar el estado de riesgo biopsicosocial en las madres que asisten periódicamente al control prenatal, utilizando un conjunto de técnicas de minería de datos que permiten la detección de información procesable de un volumen grande de información mediante un análisis matemático para deducir los patrones y tendencias que existen en los datos. (Microsoft, 2017).

A partir de lo anterior se procede a realizar la siguiente pregunta de investigación: *¿Cómo diseñar un aplicativo por medio de un modelo predictivo, para la identificación del estado de riesgo biopsicosocial en madres gestantes que asistan a controles prenatales?*

2. Antecedentes

En los últimos años se ha presentado un gran crecimiento en el uso de técnicas de minería de datos, para la identificación de factores médicos dentro de una población, que permiten mejorar la precisión de los diagnósticos y/o la anticipación de posibles riegos en la salud pública. Adicional a esto, se han vinculado nuevas herramientas para la solución de problemas predictivos. Por ejemplo, Forbes destaca en su listado de las 5 innovaciones más importantes del 2018, el uso de Machine Learning para el análisis de datos ya que permite abordar los problemas eficientemente y de una manera más simple, flexible y personalizada en comparación a los procesos normalmente utilizados en el análisis de datos (Forbes, 2018). A lo largo de los últimos años, el uso de Machine Learning ha tenido mayor relevancia. A pesar de ser anteriormente conocida, se ha logrado un uso más intensivo y evolucionado, ya que los costos de implementación se han disminuido y se tiene mayor acceso a la información. (Managment Solutions, 2018)

Teniendo en cuenta las ventajas que nos ofrecen estas metodologías para la solución de problemas de clasificación, se realizó una revisión a casos de estudios aplicados en el sector de la salud que se resumen a continuación:

Tabla 1. Antecedentes.

Autores	Factores	Resumen
(Ayoub, 2014)	Parkinson y enfermedades tumorales	**Técnica:** Minería de datos mediante redes neuronales, árboles de decisiones y métodos bayesianos **Resultados:** Mejor resultado mediante redes neuronales artificiales con una precisión del 90,76%, con respecto a árboles de decisión que logro el 80.51% y métodos bayesianos 69.23%. **Conclusiones:** Mejor técnica para predecir este tipo de enfermedades según las variables y datos del estudio fue redes neuronales artificiales.
(Mosquera, Parra, Castrillón, 2016)	Psicosociales	**Técnica:** Primera fase: Algoritmo de inteligencia artificial con técnicas de clasificación Naive Bayes y árboles de decisión. Segunda Fase: Machine learning con técnicas como Redes Neuronales Artificiales, Naive Bayes y Árboles de Decisión **Resultados:** Primera fase: Efectividad del 91% en comparación con el diagnóstico clínico. Segunda Fase: Precisión del 93% para redes neuronales artificiales, 86% para Naive Bayes y 90% para árboles de decisión. **Conclusiones:** Al utilizar algoritmos genéticos, se logra hallar el óptimo global y se logra ser más eficiente en comparación a solo aplicar técnicas de minería de datos tradicionales para clasificar el grado de riesgo psicosocial.
(Hernández, Morales, Casas, Pérez, González, Rodríguez, 2015)	Hipertensión arterial en niños	**Técnica:** Machine learning, técnicas utilizadas: Naives Bayes, Functions Logistic, Lazy IBK, Trees J48, Multilayer Perceptron, Trees AD Tree **Resultados:** Mejor resultado en primera instancia no supero el 67%. Dado esto a partir de un Algoritmo genético se buscó una combinación de clasificadores que mejoro en un 73% la exactitud del sistema multiclasificador. **Conclusiones:** Se concluyó que utilizando 6 clasificadores de manera individual no se logra alcanzar la máxima medida de exactitud mientras que, al usar un Algoritmo Genético, se obtiene un multiclasificador que logra mejorar en un 6% la clasificación anterior.

		Técnica: Machine learning, técnicas utilizadas: Random Forest.
(Menden, Iorio, Garnett, McDermott, Benes, Ballester, Rodríguez, 2013)	Machine learning para la predicción de la sensibilidad de las células cancerosas a Fármacos basados en propiedades genómicas y químicas.	**Resultados: Los** modelos predijeron los valores de IC50 en una validación cruzada y en una prueba ciega independiente con coeficiente de determinación R2 de 0.72 y 0.64 respectivamente. Además, los modelos fueron capaces de predecir con una precisión comparable (R2 de 0.61) IC50 de líneas celulares de un tejido no utilizado en la etapa de entrenamiento. **Conclusiones:** Se concluyó que los modelos pueden ser utilizados para optimizar el diseño experimental de pruebas de detección de células cancerígenas al fármaco mediante la estimación de una gran proporción de valores IC50 faltantes en lugar de realizar medición experimental.

Fuente: Creada por los autores.

3. Objetivos

Objetivo General

Diseñar un aplicativo que identifique el estado de riesgo biopsicosocial en madres gestantes que asisten a controles prenatales.

Objetivos específicos

a. Reconocer el contexto del problema y las necesidades del desarrollo del proyecto.
b. Identificar y analizar información de bases de datos asociadas a muertes fetales.
c. Detectar las variables biopsicosociales relevantes asociadas al problema.
d. Clasificar el estado de riesgo biopsicosocial de las madres gestantes aplicando técnicas de minería de datos para un modelo predictivo.
e. Diseñar un aplicativo para el área de ginecología y obstetricia en una institución hospitalaria.
f. Evaluar la solución propuesta en una muestra de madres gestantes que asistan a controles prenatales.

4. Metodología

Para el desarrollo del proyecto, se implementó la metodología Cross Industry Standard Process for Data Mining (CRISP-DM). Está proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga como se hace en los modelos de ciclo de vida de desarrollo de software (Villena, 2016). A continuación, se evidencia paso a paso el desarrollo de esta metodología:

4.1. Entender los datos

Para el desarrollo de la problemática, se seleccionaron las bases publicadas por el DANE desde el año 2007 hasta el 2016, las cuales proporcionan datos de natalidad y mortalidad fetal publicados anualmente, específicamente contienen 41 variables para nacimientos y 57 para muerte fetal. Cabe resaltar que las estadísticas de defunciones fetales y de nacimientos, se obtienen a partir de la información proveniente de los certificados de defunción y de nacidos vivos, diligenciados en medio físico o digital, por médicos y personal de salud autorizado. (DANE, 2017)

4.2. Preparar los datos

4.2.1. Selección de la muestra de entrenamiento

Para dar comienzo con la preparación de los datos, se seleccionaron 18 bases del DANE (9 de nacimientos y 9 de muertes fetales). Posterior a esto, para dar tratamiento a los datos, fue necesario consolidar la información en una única base de datos. En un principio, como criterio de unificación, se identificaron las variables comunes en las bases de datos (muertes fetales y nacimientos), para luego analizar la proporción de los datos de cada acontecimiento, y se determinó que por cada diez nacimientos se presentaba una muerte fetal. Con el fin de mejorar el entrenamiento del modelo que se construiría en fases posteriores, se modificó la proporción de los acontecimientos teniendo en cuenta una relación 1 a 1, es decir, por cada caso de nacimiento habría una muerte fetal. Para realizar dicha selección en la base de datos de nacimientos, se seleccionaron aleatoriamente los datos con el fin de eliminar posibles sesgos que se pudiesen presentar al tomar los datos arbitrariamente.

En resumen, se tomaron en promedio 90.000 registros en total por año, distribuidos en 50.000 registros de nacimientos y 40.000 de muertes fetales para un total de 816.354 registros aproximadamente.

4.2.2. Preprocesamiento

Una vez obtenida la base de datos unificada, se procedió a hacer una limpieza de los datos que se resume a continuación:

- *Valores perdidos:* En las variables categóricas, se presentaban datos que no contenían información relevante, es decir, se mostraban de forma indeterminada. Para esto, se procedió a eliminar dichos datos y reemplazarlos por espacios vacíos.
- *Imputación de valores:* Si la variable era de tipo escalar, se procedió a realizar un promedio, el cual obtiene el valor con mayor peso y mayor relevancia, para reemplazarlo por los espacios indeterminados o vacíos.

4.2.3. Selección de las variables

A partir de los registros obtenidos, se procedió a seleccionar las variables que fueran relevantes para el modelo. Se utilizaron dos métodos para dicha selección: El primero, consistió en eliminar las variables que no contenían mayor cantidad de información, es decir, que más de la mitad de sus datos fueran celdas vacías.

Una vez obtenidas las variables finales (Ver Tabla 2), se procedió a realizar una investigación profunda por medio de revisión bibliográfica sobre la justificación correspondiente al uso de cada variable dentro del modelo (Ver anexo #1). Adicional a esto, por medio de las investigaciones, se encontraron nuevas variables asociadas al estado de riesgo biopsicosocial que se podían agregar al modelo.

Tabla 2. Descripción de las variables[1].

DESCRIPCIÓN DE LAS VARIABLES FINALES				
No.	**Tipo**	**Nombre de la variable**	**Fuente**	**Descripción**
1	Social	NOMBRE DPTO	Base de datos	Nombre del departamento donde sucedió el hecho
2		MUNICIPIO	Creada por los autores	Nombre del municipio donde sucedió el hecho

[1] Dentro de la información de la tabla, se entiende "hecho" como: Nacimiento fetal o muerte fetal.

3		ALTITUD_MNCP	Creada por los autores	Altitud del municipio
4		RES_MADRE	Base de datos	Área de residencia habitual de la madre
5		SEGURIDAD SOCIAL	Base de datos	Seguridad Social de la madre
6		IDCLASADMI	Base de datos	EPS de la madre
7		NIVEL EDUCATIVO	Base de datos	Nivel educativo de la madre
8		ULTCURMAD	Base de datos	Último año cursado por la madre
9		RANKING	Creada por los autores	Ranking de la EPS de la madre
10		MES	Base de datos	Mes en el que sucedió el hecho
11		DESARROLLO_DEP	Creada por los autores	Desarrollo del departamento
12		REGION	Creada por los autores	Región en la que sucedió el hecho
13		TIPO_EMBARAZO	Base de datos	Tipo de embarazo de la madre
14		TIEMPO_GESTACION	Base de datos	Tiempo de gestación hasta que sucedió el hecho
15		EDAD_MADRE	Base de datos	Edad de la madre
16		N_HIJOSV	Base de datos	Número de hijos vivos de la madre
17	**Psicológica**	N_HIJOSM	Creada por los autores	Número de hijos muertos de la madre
18		ESTADO_SENTIMENTAL	Base de datos	Estado sentimental de la madre
19	**N/A**	ACONTECIMIENTO	Creada por los autores	Muerte Fetal: 1, Nacimiento: 0

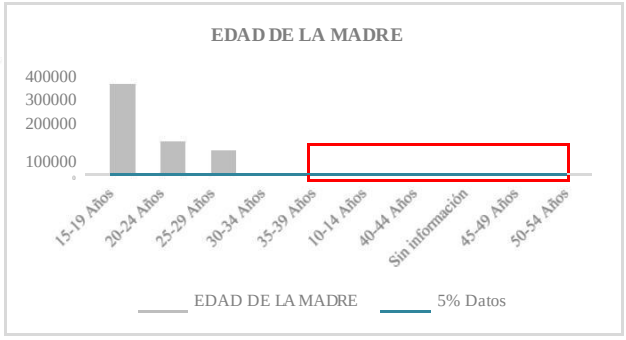
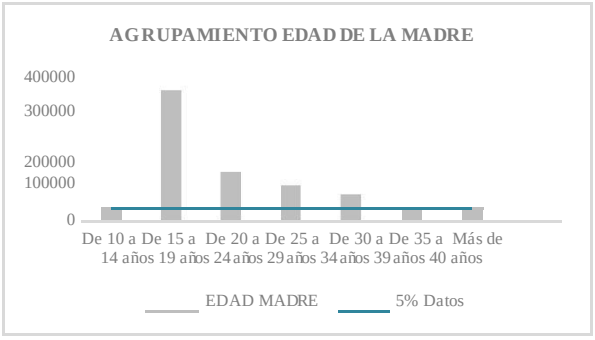
Fuente: Creada por los autores.

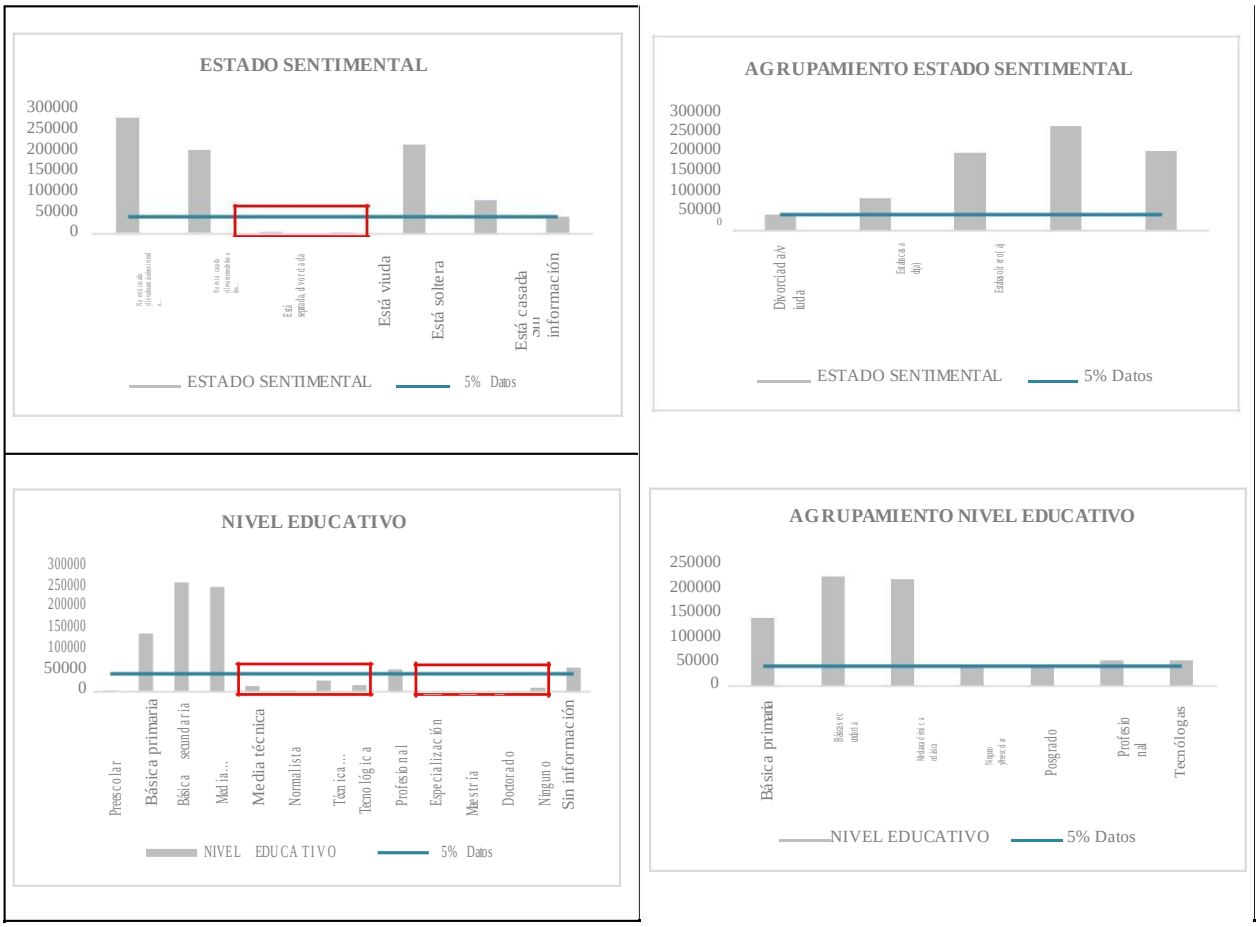
4.2.4. Construcción de las variables

Teniendo en cuenta que el modelo a construir depende de la calidad de las variables, se realizó un análisis descriptivo y estadístico, con el fin de evaluar la calidad de cada una de ellas. A continuación, se presentan las metodologías utilizadas para la construcción de las variables:

- Frecuencia: Realizando un análisis de frecuencia en todas las variables, se identificó que en las variables: Estado sentimental, Nivel educativo y Edad de la madre, ciertos intervalos de clase de cada una de estas no representaban un porcentaje significativo dentro de los datos, es decir, no superaban el 5% de la cantidad total de datos. Es por esto que se agruparon dichos intervalos de tal manera que pudieran ser representativos dentro de la variable:

Tabla 3. Resumen del agrupamiento de las clases.

Frecuencia variable	Agrupamiento de las clases
	



Fuente: Creada por los autores.

- Weight of Evidence (WOE): Se identificó que la variable Hijos muertos presentaba clases con poca significancia en datos en comparación con otras clases de la misma variable. Es decir, los datos agrupados no superaban más del 5% de los casos registrados (Ver Gráfica 1), es por esto que se determinó que la variable era sensible a agrupaciones de clases entre los intervalos 3 al 15. Sin embargo, dado que más del 80% de los datos se concentraban en las categorías 0, 1 y 2, se implementó Weight of evidence (Ver Imagen 1), que es una razón entre las proporciones de datos buenos y malos en cada clase. (Nieto, 2010)

Imagen 1. Ecuación WOE

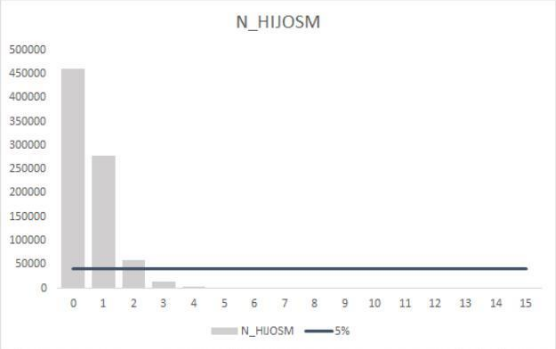
$$WOE_{ij} = 100 \cdot \ln\left(\frac{\text{Distribución de buenos en el atributo } j \text{ de la característica } i}{\text{Distribución de malos en el atributo } j \text{ de la característica } i}\right)$$
$$= 100 \cdot \ln\left(\frac{Pb_{ij}}{Pm_{ij}}\right)$$
$$= 100 \cdot \ln\left(\frac{b_{ij} \cdot m_i}{m_{ij} \cdot b_i}\right).$$

Fuente: (Nieto, 2010)

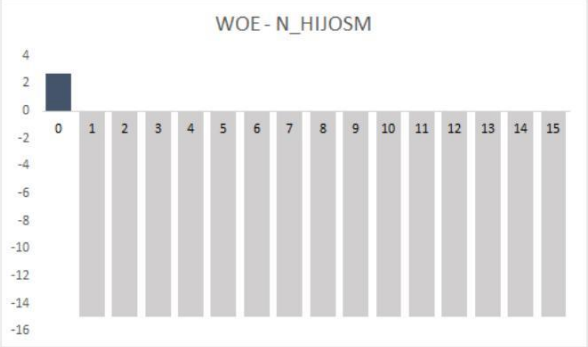
Cuando se obtienen valores negativos de WOE, significa que se tienen proporciones altas de datos malos sobre los buenos. Al observar la Gráfica 2, los datos del 1 al 15 presentan un WOE de valores negativos, mientras que la etiqueta de cero hijos muertos es la única que representa un valor positivo.

Lo anterior lleva a convertir la variable escalar en una categórica para clasificarla como: madres con antecedentes y sin antecedentes de hijos muertos.

Gráfica 1. Frecuencia de la variable Hijos Muertos.



Gráfica 2. WOE de la variable Hijos Muertos.



Fuente: Gráficas creadas por los autores.

4.3. Modelo

Para el desarrollo del modelo, se utilizaron las siguientes variables:

Tabla 4. Variables utilizadas para el modelo.

VARIABLES PARA EL MODELO				
No.	**Tipo**	**Nombre de la variable**	**Fuente**	**Descripción**
1	**Social**	RES_MADRE	Base de datos	Área de residencia habitual de la madre
2		SEGURIDAD SOCIAL	Base de datos	Seguridad Social de la madre
3		NIVEL EDUCATIVO	Base de datos	Nivel educativo de la madre
4		ULTCURMAD	Base de datos	Último año cursado por la madre
5		ALTITUD_MNCP	Creada por los autores	Altitud del municipio
6		RANKING	Creada por los autores	Ranking de la EPS de la madre
7		MES	Base de datos	Mes en el que sucedió el hecho
8		DESARROLLO_DEP	Creada por los autores	Desarrollo del departamento
9		REGION	Creada por los autores	Región en la que sucedió el hecho
10	**Biológica**	TIPO_EMBARAZO	Base de datos	Tipo de embarazo de la madre
11		EDAD_MADRE	Base de datos	Edad de la madre
12	**Psicológica**	ESTADO_SENTIMENTAL	Base de datos	Estado civil de la madre
13	**N/A**	ACONTECIMIENTO	Creada por los autores	Muerte Fetal: 1, Nacimiento: 0

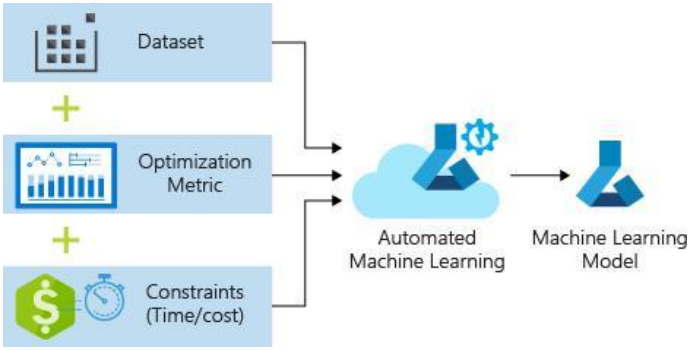
Fuente: Creada por los autores.

4.3.1. Selección de la metodología

El aprendizaje automático o también conocido como "Machine Learning", es un método de análisis de datos involucrado con la Inteligencia Artificial, que crea sistemas que identifican patrones dentro de una gran cantidad de datos, para luego tomar decisiones (SAS, 2018). Funciona de tal manera que el sistema aprende algoritmos, siendo capaz de predecir o clasificar comportamientos futuros. Sin embargo, uno de los desafíos de esta metodología consiste en diseñar tuberías de aprendizaje automáticas efectivas, ya que requiere una gran cantidad de recursos, conocimientos e inversión de tiempo para comparar los resultados de distintos modelos.

En respuesta a este desafío, desde el año 2015 se ha venido desarrollando un aprendizaje automático automatizado "Auto Machine Learning" (Olson, Moore, 2016), que consiste en tomar datos de entrenamiento con una función objetivo definida e iterar a través de combinaciones de algoritmos y funciones, para luego seleccionar automáticamente el mejor modelo y así obtener el mejor resultado. (Microsoft, 2018) Por ende, el objetivo del aprendizaje automático automatizado es hacer que el aprendizaje automático sea más accesible (Ver Imagen 2), al generar automáticamente un flujo de análisis de datos que puede incluir el preprocesamiento de datos y la selección de características (Olson, Moore, 2019).

Imagen 2. Funcionamiento del aprendizaje automático automatizado.



Fuente: (Microsoft, 2018)

Teniendo en cuenta lo anterior, para el proyecto se utilizará la metodología AutoML para automatizar modelos de aprendizaje supervisado de clasificación y operadores de preprocesamiento. A lo largo de este capítulo, se explicará el proceso mediante el cual los algoritmos de aprendizaje automático automatizado, pueden seleccionarse, aplicarse y evaluarse para el problema.

4.3.2. Selección del algoritmo

Para el desarrollo del problema fue necesario definir la herramienta adecuada para la ejecución del aprendizaje automático automatizado a través de Python. Para esto, se realizó un análisis comparativo entre TPOT, Auto Sckit Learn, Data robot y Pipeline, con el fin de seleccionar la herramienta que permitiera un resultado más eficiente y acertado.

- Auto Sckit learn: Esta librería incluye varios algoritmos de clasificación, regresión y análisis de grupos como SVM, Gradient boosting, Random Forest, entre otros. El objetivo de esta herramienta, es encontrar el mejor resultado entre los algoritmos disponibles, mediante la optimización de hiperparámetros. A partir de optimización bayesiana, la herramienta captura la relación entre la configuración de hiperparámetros y su rendimiento para luego seleccionar configuraciones útiles de los mismos para probar resultados. (KDnuggets, 2016)

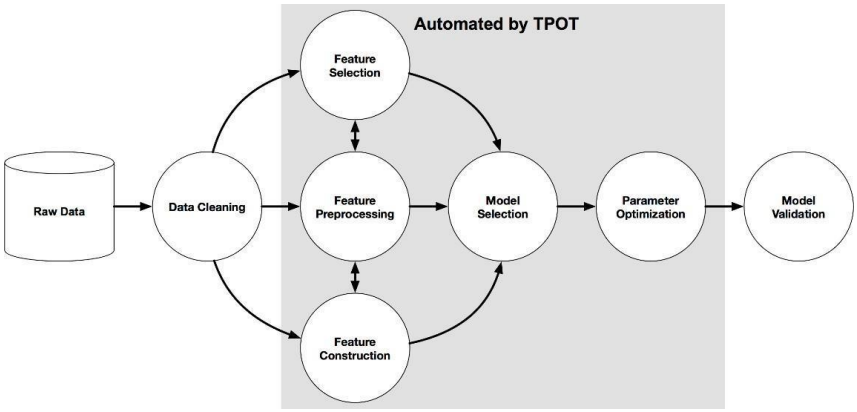
- Data Robot: Esta herramienta es ágil e intuitiva para el desarrollo de modelos predictivos ya que mediante el cargue de información y la selección de la variable objetivo, Data Robot ejecuta múltiples algoritmos para buscar un rango de valores para cada hiperparámetros y de esta manera llegar a un resultado preciso. Si en dado caso no se presenta la precisión esperada, es posible probar otro conjunto de hiperparámetros cambiando cada uno de sus valores con el fin de generar un nuevo modelo más preciso. (Data Robot, 2018)

- Pipeline: La herramienta Pipeline de Sckit-learn, facilita la elección de algoritmos mediante una metodología de tuberías. Esta metodología se basa en la utilización de hiperparámetros predeterminados con el fin de ejecutar cada uno de los algoritmos y dar como resultado el Accuracy de cada uno de estos para escoger el más acertado. (Medium, 2018)

- Tpot: Es una herramienta de Python que crea y optimiza automáticamente las tuberías de aprendizaje automático mediante la programación genética para encontrar los mejores parámetros y conjuntos de modelos (Ver Imagen 3). Tpot evalúa su rendimiento y cambia aleatoriamente partes de las tuberías en busca de algoritmos de mejor rendimiento (Data Science, 2018). Así mismo, detecta cuáles son las variables importantes en el modelo y encuentra la combinación de técnicas eficientes para poder maximizar la precisión de sus predicciones y

Teniendo en cuenta lo anterior, se analizaron cada una de las características de las herramientas mencionadas y se seleccionó Tpot, debido a que muestra resultados en un menor tiempo y existe un control del cómo se llega a una solución final, ya que depende de algoritmos genéticos. Así mismo, dicha selección se realizó teniendo en cuenta que Tpot proporciona no sólo un compilado de algoritmos si no también la preparación de datos, selección de características, elección y validación de modelos, ajuste de hiperparámetros y además, herramientas de ingeniería para la optimización de resultados.

Imagen 3. Automatización por TPOT



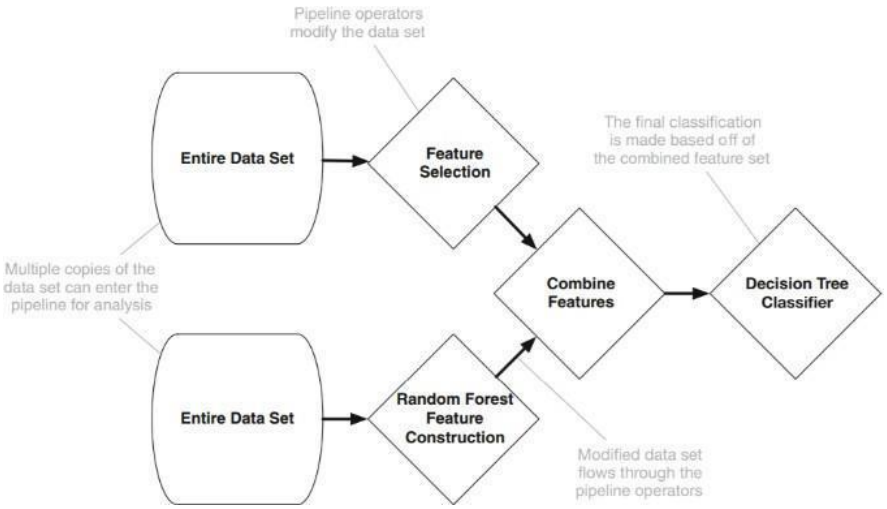
Fuente: (Data Science, 2018)

4.3.3. Optimización de tuberías con algoritmos genéticos

La Herramienta de Optimización de Tubería Basada en Árboles (TPOT), fue uno de los primeros métodos del aprendizaje automático automatizado, en el que mediante programación genética, evalúa diferentes algoritmos, generando soluciones a partir de la optimización. Estas soluciones las consigue teniendo en cuenta las propiedades básicas de la programación genética.

En Tpot, cada tubería corresponde a un algoritmo de aprendizaje automático automatizado. En la imagen 4 se puede visualizar una tubería basada en árboles, donde inicialmente se proporcionan dos copias del conjunto de datos a la tubería, modificadas posteriormente por operadores aleatorios de selección y construcción de características, luego se combinan las bases modificadas en un conjunto de datos y finalmente se utiliza para hacer clasificaciones (Olson, Moore, 2016). En la tabla 5 se muestran los modelos de clasificación y sus parámetros los cuales se prueban en las diferentes tuberías.

Imagen 4. Optimización de tuberías



Fuente: (Olson, Moore, 2016)

Para ejecutar automáticamente la optimización y generación de las tuberías, Tpot utiliza programación genética. En este caso el algoritmo genético (AG) construye tuberías evolucionando los diferentes operadores que actúan sobre el conjunto de datos al igual que los parámetros de estos operadores, teniendo como función objetivo: Maximizar la *precisión* de la clasificación final de la tubería. (Olson, Urbanowicz, Andrews, Lavender, Kidd, Moore, 2016).

Para generar la población inicial sobre la cual trabajará el algoritmo genético, inicialmente se crea 100 tuberías aleatorias las cuales serán validadas con validación cruzada y evaluadas a partir de la *precisión* (medida fitness). Posterior a esto, se seleccionan los individuos (tuberías) más aptos para generar la descendencia de la siguiente generación. Después de la selección, se procede a realizar el cruce entre los individuos seleccionados, que viene siendo la combinación de sus genes (parámetros de la tubería). Este cruce se realiza en un punto aleatorio de las tuberías, lo que se conoce como operador de cruce basado en un punto. Posterior a esto, la descendencia obtenida sufre un cambio o mutación de forma aleatoria, y de esta forma finalmente se obtiene la población de cada generación.

El proceso anteriormente mencionado (Evaluar-Seleccionar-Cruzar-Mutar), se repite en el algoritmo genético el número de generaciones que se establezca. Con el transcurso de las generaciones, el algoritmo va presentando modificaciones con el fin de afinar operadores que aumenten la función objetivo (precisión). Así mismo va eliminando aquellos que resultan redundantes o perjudiciales para la F.O.

Tabla 5. Modelos de clasificación que utiliza Tpot.

Modelo	**Parámetros/Valores**
Gaussian Naive Bayes (sklearn.naive_bayes.GaussianNB)	

Naive Bayes para modelos multivariados de Bernoulli. (sklearn.naive_bayes.BernoulliNB)	alfa: [1e-3 , 1e-2 , 1e-1 , 1. , 10. , 100.] fit_prior: [Verdadero, Falso]
Naive Bayes para modelos multinomiales (sklearn.naive_bayes.MultinomialNB)	alfa : [1e-3 , 1e-2 , 1e-1 , 1. , 10. , 100.] fit_prior : [Verdadero , Falso]
Árbol de decisión (sklearn.tree.DecisionTreeClassifier)	criterio : [" gini " , " entropía "] max_depth : rango (1 , 11) min_samples_split : range (2 , 21) min_samples_leaf : range (1 , 21)
Extra-árboles. (sklearn.ensemble.ExtraTreesClassifier)	n_estimadores : [100] criterio : [" gini " , " entropía "] max_features : np.arange (0.05 , 1.01 , 0.05) min_samples_split : range (2 , 21) min_samples_leaf : range (1 , 21) bootstrap : [Verdadero , Falso]
Random Forest (sklearn.ensemble.RandomForestClassifier)	n_estimadores : [100] criterio : [" gini " , " entropía "] max_features : np.arange (0.05 , 1.01 , 0.05) min_samples_split : range (2 , 21) min_samples_leaf : range (1 , 21) bootstrap : [Verdadero , Falso]
Gradient Boosting (sklearn.ensemble.GradientBoostingClassifier)	n_estimadores : [100], learning_rate : [1e-3 , 1e-2 , 1e-1 , 0.5 , 1] max_depth : rango (1 , 11) min_samples_split : range (2 , 21) min_samples_leaf : range (1 , 21) submuestra : np.arange (0.05 , 1.01 , 0.05) max_features : np.arange (0.05 , 1.01 , 0.05)
k de vecinos más cercanos. (sklearn.neighbors.KNeighborsClassifier)	n_neighbors : rango (1 , 101) weights: ["uniform", "distance"] p: [1, 2]
Soporte lineal de clasificación de vectores (sklearn.svm.LinearSVC)	penalty: ["l1", "l2"] loss: ["hinge", "squared_hinge"], dual: [True, False], tol: [1e-5, 1e-4, 1e-3, 1e-2, 1e-1], C: [1e-4, 1e-3, 1e-2, 1e-1, 0.5, 1, 5, 10, 15, 20, 25]
Regresión logística (sklearn.linear_model.LogisticRegression)	penalty: ["l1", "l2"] C: [1e-4, 1e-3, 1e-2, 1e-1, 0.5, 1, 5, 10, 15, 20, 25] dual: [True, False]

Fuente: (Data Camp, 2018)

4.3.4. Selección de variables para el algoritmo

TPOT utiliza la librería **sklearn.feature_selection,** ésta implementa algoritmos de selección de características e incluye métodos de selección univariable y el algoritmo de eliminación de características para mejorar la precisión de los estimadores y su rendimiento en conjuntos de datos de alta dimensión (Data Science, 2018). Las metodologías utilizadas para la selección de variables o características son[2]:

[2] Metodologías obtenidas de la referencia: (Data Science, 2018)

- **Eliminación de características con baja varianza:** En primera instancia, el selector de características VarianceThresholdes de Scikit Learn, elimina todas las características de baja varianza. Por ejemplo, en un conjunto de datos con características booleanas elimina todas las características que son cero o uno en más del 80% de las muestras. Las características booleanas son variables aleatorias de Bernoulli, y la varianza de tales variables está dada por: $[\ ] = (1 -)$. De esta manera se determina el umbral para realizar la selección de variables: $[\ ] = \%\,(1 - \%\ )$ Finalmente, las características que no superen el umbral son eliminadas.

- **Selección de características univariadas:** Esta metodología funciona seleccionando las mejores características basadas en pruebas estadísticas univariadas:

- **Chi cuadrado**: Mide la dependencia entre variables estocásticas, por lo que el uso de esta función elimina las características que tienen más probabilidades de ser independientes y, por lo tanto, irrelevantes para la clasificación.
- **F de Anova:** Los métodos basados en la prueba F, estiman el grado de dependencia lineal entre dos variables aleatorias. Esta prueba es utilizada para calcular la matriz de puntuación de las funciones selectBest, Select Fpr, SelectFdr, SelectFwe, Selectpercentile, que permiten seleccionar las 10 mejores características a partir de puntuaciones, prueba de datos falsos positivos, tasa estimada de descubrimiento falso, error familiar y percentiles, respectivamente.
- **Información mutua para una variable objetivo discreta:** La información mutua entre dos variables aleatorias, es un valor no negativo que mide la dependencia entre las variables. Es igual a cero si y solo si, dos variables aleatorias son independientes, y los valores más altos significan una mayor dependencia. La función mutual_info_classif, se basa en métodos no paramétricos para estimación de los k vecinos más cercanos, mediante esta metodología se seleccionan las variables que tienen correlación entre sí.

- **Selección de características ponderadas:** Dado un modelo lineal, se asignan pesos a las características. En Primera instancia, el estimador se entrena y se le da la importancia a cada una de estas. Luego, las características menos importantes se eliminan del conjunto actual de características y este procedimiento se repite hasta que finalmente se alcanza el número deseado de características para seleccionar.

4.3.5. Selección de los hiperparámetros de ejecución

De manera predeterminada, TPOTClassifier busca una gran cantidad de algoritmos de clasificación supervisados, transformadores e hiperparámetros (Ver Tabla 6), que pueden personalizarse completamente. Para el modelo, se modificaron dos hiperparámetros: El número de generaciones y Offspring_size. El primero, se modificó a 50 generaciones, las cuales proporcionaron cada una la eficiencia de entrenamiento del modelo. Cuando el algoritmo realiza un mayor número de iteraciones, da como resultado un modelo más óptimo para la data en ejecución. Por otro lado, se utilizaron 20 Ofspring**,** que como se explicó en el enunciado *4.3.*, mide la cantidad de veces que se cambiarán los parámetros por cada generación y por cada modelo. Para los demás hiperparámetros, se utilizaron los valores predeterminados.

Tabla 6. Parámetros TPOTClassifier

Parámetros	• **(Generaciones=50)**: Número de iteraciones para el proceso de optimización de la tubería de ejecución. • **(Population_size=100):** Número de individuos a retener en la población de programación genética de cada generación. • **(Offspring_size=20):** Número de descendientes a producir en cada generación de programación genética. • **(Mutation_rate=0.9):** Tasa de mutación para el algoritmo de programación genética en el rango [0.0, 1.0]. Este parámetro le dice al algoritmo de GP cuántas tuberías deben aplicar cambios aleatorios a cada generación. • **(Crossover_rate=0.1):** Velocidad de cruce para el algoritmo de programación genética en el rango [0.0, 1.0]. Este parámetro le dice al algoritmo de programación genética cuántas canalizaciones debe "criar" cada generación. • **(Puntuación="Precisión")**: Función utilizada para evaluar la calidad de una tubería para un problema de clasificación. • **(Cv=5):** Estrategia de validación cruzada utilizada al evaluar tuberías. • **(Submuestra=1.0):** Fracción de las muestras de entrenamiento que se utilizan durante el proceso de optimización de TPOT. • **(Ajuste de submuestra=1):** Le dice a TPOT que use una submuestra aleatoria de la mitad de los datos de entrenamiento. Esta submuestra permanecerá igual durante todo el proceso de optimización de la tubería: 0.5 n_jobs.

Fuente: (Data Camp, 2018)

4.3.6. Resultado final

El resultado de la búsqueda es un conocido método de clasificación llamado **Gradient Boosting**. Este tipo de clasificador es un método de aprendizaje supervisado que se utiliza para: Clasificación y regresión. Así mismo, éste produce un modelo de predicción en forma conjunta a partir de modelos de predicción débiles como lo son los árboles de decisión. Mediante un gradiente, se busca aprovechar al máximo la reducción del error cuadrático medio. Este método maneja los siguientes hiperparámetros:

Tabla 7. Hiperparámetros de Gradient Boosting.

Parámetros	• **(Learning_rate=0,5)**: La tasa de aprendizaje reduce la contribución de cada árbol por learning_rate. • **(Max_depth=7):** Profundidad máxima de los estimadores de regresión individuales. La profundidad máxima limita el número de nodos en el árbol. • **(Max_features=0,2):** La cantidad de características a considerar cuando se busca la mejor división. • **(Subsample=0,9):** La fracción de muestras que se utilizarán para ajustar a los aprendices base individuales. • **(Nestimators=100):** El número de etapas de impulso a realizar

	• **(Min_samples_split=4):** se refiere a la cantidad mínima de muestras que debe tener un nodo para poder subdividir. Especifica el número mínimo de muestras requerido para dividir un *nodo interno*. • **(Min_samples_leaf=14):** Especifica el número mínimo de muestras requeridas para estar en un nodo *hoja*. Si tuviera menos, no se formaría esa hoja y "subiría" un nivel su antecesor

Fuente: (Data Camp, 2018)

GradientBoostingClassifier toma como entrada dos matrices: Una matriz X_train, que contiene las muestras de entrenamiento y una matriz Y_train, que contiene las etiquetas de clase para las muestras de entrenamiento: [n_samples, n_features] [n_samples]. Una vez entrenado el modelo, se imprimió la matriz de confusión y el Accuracy, para identificar el nivel de predicción del modelo. Por otro lado, para evitar el sobre entrenamiento, se utilizó el 80% de la base para ejecutar TPOT. El modelo resultante, fue entrenado y probado con el 20% restante. A continuación, se presenta la matriz de confusión y el Accuracy:

Tabla 8. Matriz de confusión.

Matriz de Confusión		PREDICHO	
		Negativo	**Positivo**
REAL	**Negativo**	16500	2807
	Positivo	3853	11384

Fuente: Creada por los autores.

La anterior matriz muestra que se predijeron correctamente 11384 casos donde las mujeres en embarazo se encontraban en riesgo biopsicosocial y 16500 donde no presentaban el riesgo. Adicional, se presentaron 6660 casos de falsos positivos, de los cuales 3853 son mujeres que presentan el riesgo, pero el modelo predijo que no. Los 2807 casos restantes, son mujeres que no presentan el riesgo, pero el modelo predijo que si se encuentran en riesgo biopsicosocial. La anterior información se predijo con un Accuracy del 80,7%.

Por otro lado, para verificar que el modelo no tuviera un sobreajuste, se realizó el proceso de validación con otra base que el modelo desconociera, en este caso, la utilizada en TPOT. A continuación, se presenta la matriz de confusión y el Accuracy de la predicción de esta prueba:

Tabla 9. Matriz de confusión.

Matriz de Confusión		PREDICHO	
		Negativo	**Positivo**
REAL	**Negativo**	265384	45427
	Positivo	61689	180194

Fuente: Creada por los autores.

Esta matriz muestra que se predijeron correctamente 180194 casos donde las mujeres en embarazo se encontraban en riesgo biopsicosocial y 265384 casos donde no presentaban el riesgo. Adicional, se presentaron 107116 casos de falsos positivos, lo cual se predijo con un Accuracy del 80,6%. De acuerdo con esto, el nivel de predicción se mantiene.

Tabla 10. Importancia de las variables.

#	Variables	Importancia
1	TIPO_EMBARAZO	76%
2	REGIÓN	12%
3	EDAD_MADRE	5%
4	MES	4%
5	SITUACIÓN_SENTIMENTAL	3%

Fuente: Creada por los autores.

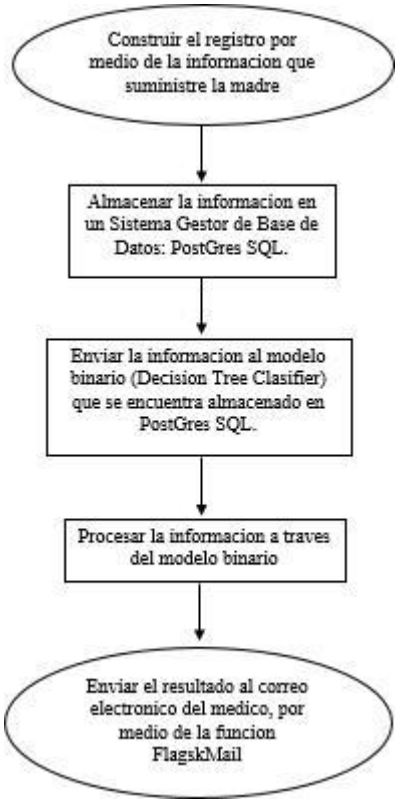
Como resultado se obtuvieron 5 variables importantes para el modelo: Tipo de embarazo, región, edad de la madre, mes y situación sentimental, con 76%, 12%, 5%, 4% y 4% respectivamente. De acuerdo con lo anterior, las variables restantes no generan un impacto dentro de este.

4.4. Aplicativo

4.4.1. ¿Cómo funciona el aplicativo?

La principal función del aplicativo es adquirir la información biopsicosocial de la madre que asiste a la entidad de salud, para que ésta sea procesada por el modelo de clasificación, el cuál envía al correo electrónico del médico el estado de riesgo en el que se encuentra la paciente, que en este caso es: Fuera de Riesgo, En Riesgo. En la Figura 6, se evidencia el proceso que ejecuta el aplicativo.

Imagen 6. Proceso de ejecución del aplicativo.



Fuente: Creada por los autores.

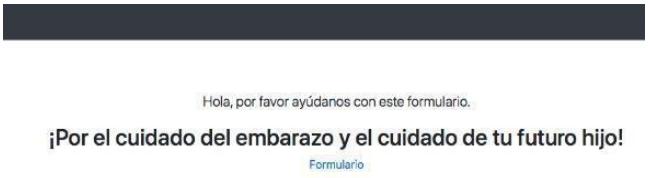
4.4.2. Interfaz

La interfaz con la cual interactuarán los usuarios (Ver Anexo #2), se realizó mediante Bootstrap, un marco de trabajo para interfaz web que se visualiza de la siguiente manera:

- **Interacción con la madre:**

1. La madre visualizará en primera instancia un saludo.

Imagen 7. Captura de pantalla página 1 del aplicativo.



Fuente: Creada por los autores.

2. Seguido a esto, deberá diligenciar los datos requeridos.

Imagen 8. Captura de pantalla página 2 del aplicativo.

Fuente: Creada por los autores.

Dentro de los campos requeridos por la encuesta (Ver Imagen 8), se busca así mismo identificar qué enfermedades padece la madre, ya que esto podría afectar el desarrollo del embarazo. Por ejemplo, enfermedades como la diabetes, hipertensión, obesidad, enfermedad renal, etc., son factores biológicos que se deben tener en cuenta a la hora de evaluar los factores biopsicosociales de la madre. (Molina, Alfonso, 2010)

3. Finalmente se agradece a la madre por desarrollar la encuesta.

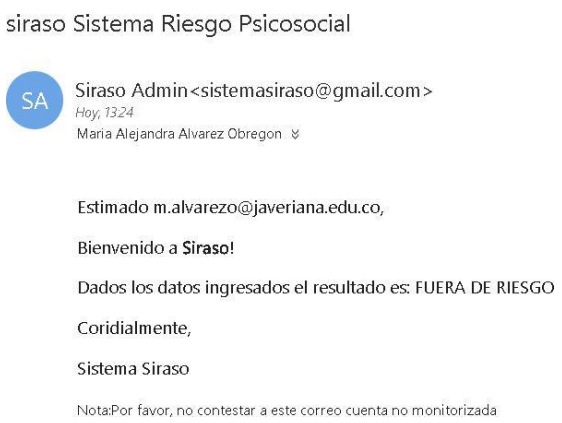
Imagen 9. Captura de pantalla página 3 del aplicativo.

Fuente: Creada por los autores.

- **Interacción del médico:**

1. Posterior al diligenciamiento de la encuesta, el médico recibirá un e-mail que le permitirá conocer el resultado obtenido por el modelo binario.

Imagen 10. Captura de pantalla correo electrónico.



Fuente: Creada por los autores.

4.4.3. Pruebas de desempeño

El aplicativo fue probado en una muestra de 10 madres en la ciudad de Pasto en un consultorio ginecológico de la EPS Emsanar, en donde se identificó que el modelo de predicción cuenta con las siguientes características:

- Es funcional, ya que permite la recolección de información de manera inmediata y es procesada al instante.
- Determina y clasifica el estado de riesgo biopsicosocial a través de la información que la madre digita antes de la consulta.
- El tiempo de procesamiento y ejecución fue de 10 segundos, por lo tanto, es eficiente.

4.4.4. Cumplimento y testeo

- Para garantizar los requerimientos de desempeño, se comprobaron los resultados del aplicativo con el criterio del médico y estos en su totalidad fueron acertados.
- En cuanto a la funcionalidad, no se obtuvo ningún tipo de error o falla en la ejecución del aplicativo.
- Se desarrolló el aplicativo bajo los lineamientos de la Norma ISO 9126 como estándar para la creación y desarrollo de éste. (Ver Tabla 11)

Tabla 11. Norma ISO 9126

Características	Aplicación
Funcionalidad	- **Idoneidad:** El aplicativo permite que la madre suministre la información necesaria, pero la información es visualizada por el médico tratante, reduciendo el impacto del conocimiento de la predicción por parte de la madre. - **Exactitud:** Según las pruebas de desempeño, el aplicativo coincidió con el diagnóstico del médico en un 100%. - **Interoperabilidad:** Existe información compartida entre la madre, el médico y la institución de salud.

	- **Seguridad:** El medico tendrá que entrar autenticado a su cuenta de correo electrónico y será de uso institucional. - **Cumplimiento de normas:** La interfaz cumple con la Resolución Nº 823 de 23 de Marzo de 2017, restringiendo la manipulación y publicación de la información utilizada en cada predicción.
Fiabilidad	- **Madurez:** La interfaz está en fase piloto y ha sido testeada, se espera realizar una segunda fase de implementación - **Recuperabilidad:** Los resultados de cada predicción, son almacenados en la nube del servidor de correo y pueden ser recuperados. Adicional a esto, el modelo utilizado puede ser consultado y modificado utilizando la herramienta Python. - **Tolerancia a fallos:** El modelo predictor fue entrenado y testeado a partir de datos de procesamiento y de entrenamiento.
Utilidad	- **Comprensión:** La interfaz es intuitiva y presenta preguntas de fácil entendimiento. - **Operatividad:** La interfaz no requiere de interacción adicional a la programada, se entregan los resultados en línea al médico tratante. - **Atractividad:** La interfaz es de gran utilidad dado la importancia de herramientas que faciliten la identificación del riesgo y teniendo en cuenta que en el sistema de salud colombiano no se ha desarrollado teniendo en cuenta la metodología usada.
Eficiencia	- **Comportamiento en el tiempo:** La herramienta tarda aproximadamente 10 segundos en procesar y entregar un resultado. - **Comportamiento de recursos:** La interfaz y el modelo de machine learning opera en óptimas condiciones teniendo en cuenta que es un modelo que fue optimizado, testeado y validado por médicos antes de emitir un diagnóstico.
Mantenibilidad	- **Estabilidad:** El sistema no presenta inconvenientes al momento de las consultas, opera normalmente. - **Facilidad de análisis:** Se entregan los resultados de manera concreta. - **Facilidad de cambio:** La interfaz puede ser cambiada sin necesidad del desarrollo de un proyecto de larga duración. - **Facilidad de pruebas:** Puede ser testeado tantas veces se requiera.
Portabilidad	- **Adaptabilidad:** Puede ser ejecutado en cualquier entidad de salud que tenga conexión a internet.

Fuente: (Abud, 2000)

4.5. Conclusiones

- A partir de la información recolectada de las bases de datos, fue posible comprender y analizar la problemática de riesgo de muerte fetal a causa de factores biopsicosociales, esto a partir de la implementación de nuevas metodologías e innovaciones en Auto Machine Learning. Teniendo en cuenta la metodología utilizada, se logró tener un proceso robusto y completo de análisis de datos en comparación a estudios realizados anteriormente en el desarrollo del problema.

- Con el uso de la metodología planteada (AutoML), se logró obtener la mejor tubería junto con el modelo de aprendizaje automático y la configuración de parámetros más óptima. En este caso, se obtuvo como mejor modelo de clasificación llamado Gradient Boosting con un Accuracy del 80,7%. De acuerdo con los resultados obtenidos, se logró identificar que al aplicar Auto Machine Learning, se obtienen resultados más óptimos y eficientes en menor tiempo, permitiendo abordar problemáticas de manera proactiva evitando la proliferación de las mismas.

- Teniendo en cuenta los resultados obtenidos por el modelo, se logró identificar que el "Tipo de embarazo" es la variable con mayor importancia para la predicción del riesgo biopsicosocial. Esto implica que madres con un embarazo múltiple sean vulnerables a presentar preclamsia, diabetes gestacional, entre otros, ya que el cuerpo de la madre debe soportar una sobre carga durante el embarazo.

- "Región" es la segunda variable más importante para el modelo. Lo anterior puede tener una estrecha relación con la alta proporción de madres gestantes con necesidades básicas insatisfechas y el alto índice de pobreza en ciertas regiones de Colombia. Esto puede causar deficiencia nutricional e infecciones que generan un ambiente intrauterino sub-óptimo que limita el desarrollo fetal.

- Se logró desarrollar un aplicativo funcional, el cual se implementó en una entidad de salud, donde se comprobó su funcionalidad y su precisión en la predicción, además de conocer la opinión de las madres y médicos que hicieron parte de la implementación. A partir de esto, se logró dar a conocer la importancia del uso de este tipo de herramientas multidisciplinares que aportan a problemas tanto de ingeniería como de la medicina y que contribuyen a la prevención de este tipo de problemáticas sociales.

- Las ventajas del uso de modelos de Machine Learning que posteriormente son automatizados con algoritmos de Auto Machine Learning, permite que el aplicativo desarrollado pueda manejar cualquier tipo de algoritmos. Es decir, si se vuelve a correr el AutoML y este arroja un modelo diferente al que hoy es implementado por la interfaz, ésta no requerirá ser modificada, debido a que con el mismo input de información se puede acomodar cualquier tipo de algoritmo.

4.6. Recomendaciones

- El modelo de predicción creado puede ser modificado para futuros proyectos. Sin embargo, se recomienda hacer uso de procesadores de mayor potencia con el fin de minimizar el tiempo de ejecución del modelo dado que este puede tardar entre horas y semanas, entre mayor sea el número de generaciones a evaluar en el algoritmo genético, mayor será el tiempo de procesamiento.
- Se recomienda dar uso de la herramienta con el fin de crear campañas de salud pública lideradas por entes privados o públicos.

- Se recomienda para futuros desarrollos, incluir en el resultado una predicción analítica prescriptiva. Por ejemplo, si la madre presenta riesgo a causa de su estado sentimental, se debería sugerir un acompañamiento con un especialista del área de psiquiatría o acompañamiento de sus familiares cercanos o de su pareja.

5. Glosario

- **Factores de riesgo biopsicosociales**: Se define a los factores que afectan el embarazo de la madre debido a causas biológicas, psicológicas y sociales.
- **Auto machine learning:** Consiste en tomar datos de entrenamiento con una función objetivo definida e iterar a través de combinaciones de algoritmos y funciones, para luego seleccionar automáticamente el mejor modelo y así obtener el mejor resultado.
- **TPOT**: Es una herramienta de auto machine learning de Python que optimiza los procesos de aprendizaje automático automatizado mediante la programación genética.
- **Programación Genética**: Son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización. Están basados en el proceso genético de los organismos vivos
- **Gradient Boosting:** Es un método de aprendizaje supervisado que se utiliza para: Clasificación y regresión. Así mismo, éste produce un modelo de predicción en forma conjunta a partir de modelos de predicción débiles como lo son los árboles de decisión.

6. Referencias

Abud M., (2000). Calidad en la industria del Software, La norma ISO-9126. Recuperado el 25 de noviembre de 2018 de http://recursosbiblioteca.utp.edu.co/tesisd/textoyanexos/0053L864e_anexo.pdf

Aguirre H., Vázquez F. (2006). El error médico, eventos adversos. *Medigraphic*, 495-503. Recuperado el 1 de abril de 2018 de http://www.medigraphic.com/pdfs/circir/cc-2006/cc066n.pdf

Ayoub T. (2014). International Journal of Engineering Science Invention ISSN, *(Online): 2319 – 6734, ISSN (Print): 2319 – 6726*, Recuperado el 20 de marzo de https://www.researchgate.net/profile/Tawseef_Shaikh/publication/303919880_A_Prototype_of_Parkinson's_and_Primary_Tumor_Diseases_Prediction_Using_Data_Mining_Techniques/links/575d91ad08aec91374aef5c3.pdf

Borrell F. (2002). El modelo biopsicosocial en evolución. *Medicina clínica*, *119*(5), 175-179. Recuperado el 5 de abril de 2018 de http://altascapacidades.es/portalEducacion/html/otrosmedios/13034093_S300_es.pdf

Colombiana de Salud S.A. Ciencia y amor nuestra solución. (2016). Protocolo Control Prenatal. *Manual de Calidad*. Recuperado el 20 de marzo de 2018, de http://www.colombianadesalud.org.co/PROMOCION_PREVENCION/INSTRUCTIVOS%20PYP/okPROTOCOLO%20CONTROL%20PRENATAL.pdf

DataCamp (2018). TpoT in Paython. Recuperado el 15 de Noviembre del 2018, de https://www.datacamp.com/community/tutorials/tpot-machine-learning-python

DataCamp (2018). TpoT automated Machine Learning in Paython. Recuperado el 15 de Noviembre del 2018, de https://towardsdatascience.com/tpot-automated-machine-learning-in-python-4c063b3e5de9

Data Robot. (2018). Cómo elegir un proveedor de inteligencia artificial. Recuperado el 09 de diciembre de 2018 de https://www.datarobot.com/wiki/tuning/

DataScience. (2018) TPOT Python Recuperado el 09 de diciembre de 2018 de http://epistasislab.github.io/tpot/using/

Departamento Administrativo Nacional de Estadística. (2017). Boletín Técnico, Estadísticas vitales - EEVV (2016-2017). Recuperado el 22 de marzo de 2018, de https://www.dane.gov.co/files/investigaciones/poblacion/bt_estadisticasvitales_2016def-2017pre.pdf

Forbes. (2018) 5 Artificial Intelligence Trends To Watch Out For In 2019 Recuperado el 01 de enero de 2019 de https://www.forbes.com/sites/janakirammsv/2018/12/09/5-artificial-intelligence-trends-to-watch-out-for-in-2019/#73ec03db5618

Hernández L, Morales A. Casas G, Denoda L, González E y Rodríguez J. (2015). Algoritmos genéticos con medidas de diversidad para el diagnóstico del riesgo de HTA en escolares. Recuperado el 28 de noviembre de 2018 de https://www.researchgate.net/publication/228708779

Herrera, J., Ersheng, G., Shahabuddin, A., Lixia, D., Wei, Y., Faisal, M., Barua, P., & Akhtner, H. (2009). Periodical assessment of the prenatal biopsychosocial risk to predict obstetric and perinatal complications in Asian countries 2002-2003. *Colombia Médica, 37*(2 Supl 1), 6-14. Recuperado de http://colombiamedica.univalle.edu.co/index.php/comedica/article/view/431/1067

Huiza, Lilia, Pacora, Percy, Ayala, Máximo, & Buzzio, Ytala. (2003). La muerte fetal y la muerte neonatal tienen origen multifactorial. *Anales de la Facultad de Medicina*, *64*(1), 13-20. Recuperado el 2 de abril de 2018, de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1025-55832003000100003&lng=es&tlng=es

KDnuggets. (2017). El estado actual del aprendizaje automático automatizado. Recuperado el 09 de diciembre de 2018 de https://www.kdnuggets.com/2017/01/current-state-automated-machine-learning.html

Marsland S. (2015). Machine Learning An Algorithmic Perspective. Second Edition. Recuperado el 16 de Agosto de 2018 de https://doc.lagout.org/science/Artificial%20Intelligence/Machine%20learning/Machine%20Learning_%20An%20Algorithmic%20Perspective%20%282nd%20ed.%29%20%5BMarsland%202014-10-08%5D.pdf

Management Solutions. (2018) Machine learning una pieza clave en la transformación del necogico, Recuperado el 01 de enero de 2019 de https://www.managementsolutions.com/sites/default/files/publicaciones/esp/machine-learning.pdf

Medium. (2018). Gestión de flujos de trabajo de aprendizaje automático con pipelines de Scikit-Learn Parte 2: Integración de la búsqueda de cuadrículas. Recuperado el 09 de diciembre de 2018 de https://medium.com/datos-y-ciencia/gesti%C3%B3n-de-flujos-de-trabajo-de-aprendizaje-autom%C3%A1tico-con-pipelines-de-scikit-learn-parte-2-eeecab194d83

Menden M, Iorio F, Garnett, M, McDermott U, Benes C, Ballester P, Saenz Rodriguez J (2013), Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties Recuperado el 19 de diciembre de 2018 de https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061318

Microsoft. (2017). Conceptos de minería de datos. Recuperado el 15 de marzo de 2018 de https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts

Microsoft. (2018). ¿Qué es el aprendizaje automático de máquinas?. Recuperado el 15 de marzo de 2018 de https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-automated-ml

Ministerio de Salud y Protección Social, Colciencias. (2013). Guías de Práctica Clínica para la prevención, detección temprana y tratamiento de las complicaciones del embarazo, parto o puerperio. *Centro Nacional de Investigación en evidencia y tecnologías en Salud CINETS*. Recuperado el 20 de marzo de 2018 de http://saludpublicavirtual.udea.edu.co/eva/pluginfile.php/5789/mod_resource/content/1/GPC_Completa_Embarazo%2011-15%20de%202013.pdf

Mosquera R., Parra L. y Castrillón O. (2016). Metodología para la Predicción del Grado de Riesgo Psicosocial en Docentes de Colegios Colombianos utilizando Técnicas de Minería de Datos. *Vol. 27(6), 259-27*. Recuperado el 3 de abril de 2018 de https://scielo.conicyt.cl/pdf/infotec/v27n6/art26.pdf

Nieto S. (2010). Crédito al consumo: La estadística aplicada a un problema de riesgo crediticio. Universidad Autónoma Metropolitana. Recuperado el 10 de Diciembre de 2018 de http://mat.izt.uam.mx/mcmai/documentos/tesis/Gen.07-O/Nieto-S-Tesis.pdf

Olson, R.S., Moore, J.H. (2016). TPOT: A Tree-based Pipeline Optimization Tool for Automating Machine Learning. JMLR **64**, 66–74. Recuperado el 24 de noviembre de 2018 de http://proceedings.mlr.press/v64/olson_tpot_2016.pdf

Olson R., Urbanowicz R., Andrews P., Lavender N., Kidd L., Moore H. 2016. Automating Biomedical Data Science Through. Tree-Based Pipeline Optimization. University of Louisville. Philadelphia, PA 19104, USA

Olson, R.S., Moore, J.H. (2019). Information about Automated Machine Learning, Recuperado el 30 de noviembre de 2018 de http://automl.info/about-us/

SAS. (2018). Aprendizaje Automático, Qué es y para qué sirve. Recuperado el 19 de noviembre de 2018 de https://www.sas.com/es_co/insights/analytics/machine-learning.html

Villena J. (2016). CRISP-DM: La metodología para poner orden en los proyectos de Data Science. Singular *Data & Analytics*. Recuperado el 25 de marzo de 2018 de https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science