**Selección de la problemática.**

**Introducción**

La Gran Encuesta de Hogares (GEH) llevada a cabo por el Departamento Administrativo Nacional de Estadística (DANE) de Colombia es una fuente fundamental de datos que proporciona una visión detallada de las condiciones sociodemográficas, económicas y educativas de la población colombiana. En este informe, se explorará el uso de la GEH para realizar un estudio de clasificación que segmente a hombres y mujeres según sus condiciones sociodemográficas, educativas, salariales y la presencia de discapacidades.

**Datos y Metodología**

La GEH recopila información de una muestra representativa de hogares colombianos, abarcando una amplia gama de variables socioeconómicas y demográficas. Estos datos se recopilan a través de entrevistas realizadas en los hogares seleccionados, abordando aspectos como la edad, género, nivel educativo, situación laboral, ingresos, entre otros.

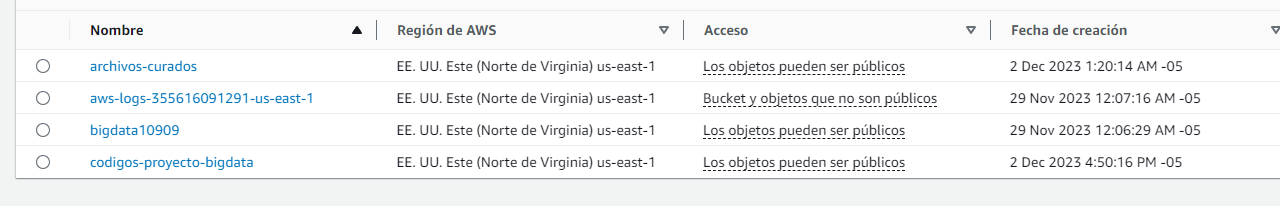
Para el estudio de clasificación propuesto, se utilizarán técnicas de Machine Learning y análisis estadístico. Se considerarán algoritmos de clasificación, como modelos de regresión logística, árboles de decisión o máquinas de soporte vectorial (SVM), para predecir la clasificación por género y las condiciones sociodemográficas y educativas.

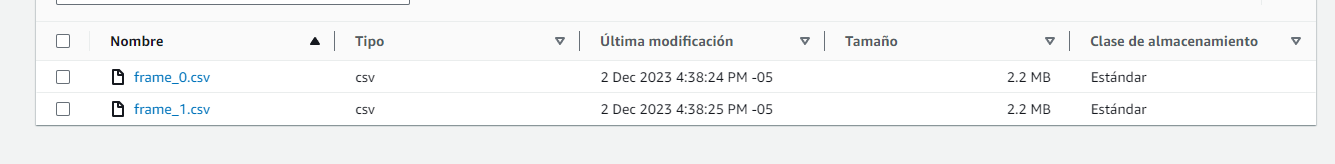
**Variables de Interés**

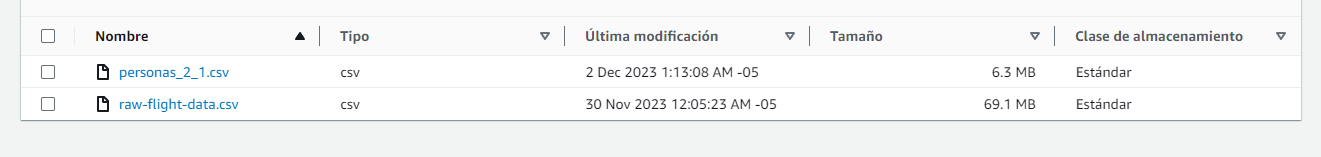
* Género: Segmentación de la población en hombres y mujeres para realizar un análisis diferenciado según estas categorías.
* Condición Sociodemográfica: Variables como la edad, estado civil, ubicación geográfica, tamaño del hogar y tipo de vivienda.
* Educación: Nivel educativo, asistencia escolar y acceso a la educación formal.
* Situación Laboral y Salarial: Estado laboral, ingresos, tipo de empleo y condiciones salariales.
* Discapacidad: Identificación de personas que presenten algún tipo de discapacidad física o cognitiva.

Se anexarán los scripts del proceso de ETL y el modelamiento para que se pueda contemplar el uso de Spark y la paralelización de datos.

**Infraestructura:**







Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Instancias:**

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

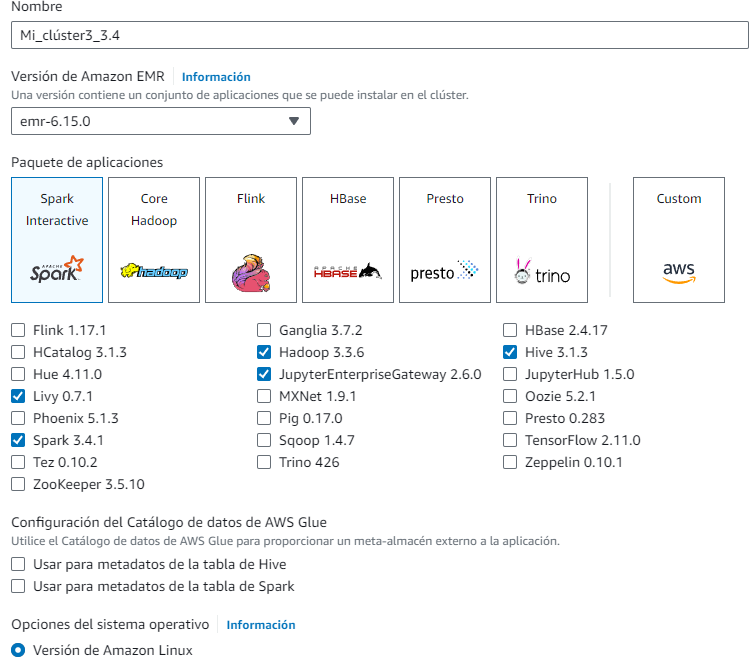
**Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente**

**Cluster EMR:**

****

**Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

**Descripción y detalle de la arquitectura:**

La creación de la arquitectura de este proyecto fue un proceso estratégico que comenzó con la configuración de buckets en Amazon S3. Estos buckets fueron diseñados para almacenar datos en su estado crudo, dando inicio a la estructura fundamental para una sólida infraestructura ETL (Extract, Transform, Load). Desde allí, la información se extrajo, procesó y transformó, llevándola a un bucket destinado para los datos depurados y listos para la etapa de modelado.

Para facilitar el desarrollo, se implementó otro bucket para almacenar los códigos esenciales. Esto permitió acceder a ellos desde la consola de la shell, integrándolos de manera eficiente en el futuro cluster EMR (Elastic MapReduce) que se crearía más adelante.

El siguiente paso consistió en la creación de una instancia con características básicas para gestionar el consumo de recursos de manera eficiente, evitando gastos innecesarios durante la ejecución. Esta instancia, basada en la última versión de Ubuntu, asegura un sólido soporte técnico y facilita la ejecución de tareas mediante la descarga de claves esenciales para operaciones desde la shell en la máquina virtual.

En cuanto a la infraestructura, se dio paso a la creación de un cluster EMR -6.15.0 con Spark, orientado a la ejecución paralela de los datos. Aunque se intentó utilizar Ubuntu como base del sistema operativo para el cluster, debido a conflictos, se optó por Amazon Linux para garantizar una configuración más estable. La distribución de la memoria se ajustó con precisión: el nodo maestro, un c1.medium con 1.7 GB de RAM, asignado a tareas menos exigentes, mientras que para los dos nodos restantes se destinaron 7 GB de memoria y 8 vCPU cada uno, permitiendo un reparto equitativo de trabajos.

Inicialmente concebido como un cluster pseudo distribuido, la inclusión de dos nodos justificó su configuración como distribuido, anticipando una posible expansión del proyecto en etapas posteriores. Próximamente se llevará a cabo la configuración de la instancia con el cluster, permitiendo la ejecución eficaz de los procesos planificados. Este diseño, meticuloso y preciso, establece una base sólida y adaptable para futuras fases del proyecto.

**Resultados del modelamiento:**

En el proceso de evaluación de datos de hogares, se llevó a cabo la construcción y evaluación de un modelo utilizando el algoritmo de Random Forest. El objetivo principal fue comprender y predecir ciertas características basadas en los datos disponibles.

Proceso de Modelado

Se utilizó el algoritmo Random Forest, conocido por su capacidad para manejar conjuntos de datos complejos y predecir con precisión, incluso en presencia de muchas características. Se implementó este modelo sobre el conjunto de datos de hogares, utilizando características específicas para obtener insights relevantes.

Resultados de la Evaluación

El modelo entrenado mostró un nivel de precisión (accuracy) de 0.32 en la evaluación. Si bien este resultado podría considerarse modesto, es esencial destacar que representa un buen comienzo para la comprensión inicial de los datos y sus relaciones.

Análisis del Rendimiento

El accuracy del 0.32 indica que el modelo clasifica correctamente aproximadamente el 32% de los casos. Si bien esto puede considerarse bajo, para un primer acercamiento y una comprensión inicial de los datos, es un resultado alentador. Este puntaje proporciona una base sobre la cual continuar mejorando el modelo, ajustando parámetros y explorando características adicionales.

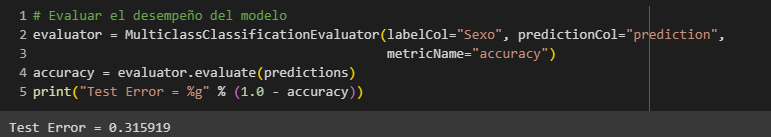
Uso del Cluster EMR y Almacenamiento de Resultados

El script del modelo se implementó en el entorno del cluster EMR. Para facilitar su ejecución y mantener un flujo de trabajo organizado, se subió al bucket correspondiente para acceder y ejecutar el script desde la máquina virtual.

Los resultados obtenidos también se almacenaron en el bucket destinado a la curación de datos, asegurando la trazabilidad y permitiendo un análisis más detallado posteriormente.

Conclusiones y Pasos Futuros

Aunque el modelo inicial proporcionó una visión introductoria, se identifica como un punto de partida sólido para profundizar en el análisis. Se planea realizar futuras iteraciones en el modelo, explorar más características relevantes y ajustar los hiperparámetros para mejorar la precisión y el entendimiento del comportamiento de los datos de hogares.



Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Accuracy (Precisión)

El accuracy es una medida que indica la proporción de predicciones correctas que realiza un modelo sobre el total de predicciones realizadas. Se calcula dividiendo el número de predicciones correctas por el número total de predicciones. En este caso, el valor de accuracy es aproximadamente 0.6841, lo que significa que alrededor del 68.41% de las predicciones realizadas por el modelo fueron correctas.

Precision (Precisión)

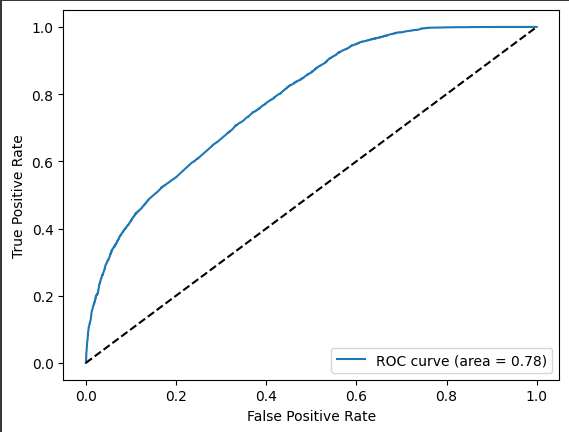
La precisión mide la proporción de predicciones positivas correctas respecto al total de predicciones positivas realizadas por el modelo. Se calcula dividiendo el número de verdaderos positivos (predicciones correctas de la clase positiva) entre la suma de verdaderos positivos y falsos positivos (predicciones incorrectas de la clase positiva). En este caso, el valor de precision es aproximadamente 0.6845, lo que indica que alrededor del 68.45% de las predicciones positivas fueron verdaderamente positivas.

Recall (Sensibilidad)

El recall mide la proporción de predicciones positivas correctas respecto al total de casos positivos reales en los datos. Se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos (instancias positivas que fueron clasificadas incorrectamente como negativas). En este caso, el valor de recall es aproximadamente 0.6841, lo que significa que alrededor del 68.41% de los casos positivos reales fueron identificados correctamente por el modelo.

F1-score

El F1-score es una medida que combina la precisión y el recall en un solo número. Se calcula como la media armónica de la precisión y el recall. Es útil cuando hay un desequilibrio entre las clases (diferente cantidad de muestras por clase) porque considera tanto falsos positivos como falsos negativos. En este caso, el valor de F1-score es aproximadamente 0.6821, lo que indica una buena combinación entre precisión y recall para el modelo en cuestión.



Nuevamente se evidencia en esta imagen la métrica de rendimiento del modelo el cual tiene una medida de 0.78 teniendo en cuenta que la curva Roc entre más se aproxime al valor punteado debajo de la curva este tomara mejores predicciones en la clasificación de los datos binarios.