

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:

	Nombre	Región de AWS	Acceso	Fecha de creación
<input type="radio"/>	archivos-curados	EE. UU. Este (Norte de Virginia) us-east-1	Los objetos pueden ser públicos	2 Dec 2023 1:20:14 AM -05
<input type="radio"/>	aws-logs-355616091291-us-east-1	EE. UU. Este (Norte de Virginia) us-east-1	Bucket y objetos que no son públicos	29 Nov 2023 12:07:16 AM -05
<input type="radio"/>	bigdata10909	EE. UU. Este (Norte de Virginia) us-east-1	Los objetos pueden ser públicos	29 Nov 2023 12:06:29 AM -05
<input type="radio"/>	codigos-proyecto-bigdata	EE. UU. Este (Norte de Virginia) us-east-1	Los objetos pueden ser públicos	2 Dec 2023 4:50:16 PM -05

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	frame_0.csv	csv	2 Dec 2023 4:38:24 PM -05	2.2 MB	Estándar
<input type="checkbox"/>	frame_1.csv	csv	2 Dec 2023 4:38:25 PM -05	2.2 MB	Estándar

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	personas_2_1.csv	csv	2 Dec 2023 1:13:08 AM -05	6.3 MB	Estándar
<input type="checkbox"/>	raw-flight-data.csv	csv	30 Nov 2023 12:05:23 AM -05	69.1 MB	Estándar

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	gran_encuesta_de_hogares.py	py	2 Dec 2023 5:07:00 PM -05	8.8 KB	Estándar

Instancias:

Instancias (2) Información

Conectar

Estado de la instancia

Acciones

Lanzar instancias

Q

Buscar Instance por atributo o etiqueta (case-sensitive)

<

1

>

<input type="checkbox"/>	Name <div></div>	ID de la instancia	Estado de la i... <div></div>	Tipo de inst... <div></div>	Comprobación ...	Estado de la ...	Zona de dis
<input type="checkbox"/>	nueva_bigdata	i-04a9fccb3e2d33cbc	<div>En ejecución</div>	t2.micro	<div>2/2 comprobaci</div>	Sin alarmas +	us-east-1b
<input type="checkbox"/>	EC2-proyecto_bigdata	i-0454517467c2d1fa0	<div>Detenida</div>	t2.micro	-	Sin alarmas +	us-east-1b

Instancia: i-04a9fccb3e2d33cbc (nueva_bigdata)		
Detalles	Seguridad	Redes
<div>Resumen de instancia Información</div> <div><div><div>ID de la instancia</div><div>i-04a9fccb3e2d33cbc (nueva_bigdata)</div></div><div><div>Dirección IPv6</div><div>-</div></div><div><div>Tipo de nombre de anfitrión</div><div>Nombre de IP: ip-172-31-32-45.ec2.internal</div></div><div><div>Responder al nombre DNS de recurso privado IPv4 (A)</div><div></div></div></div> <div><div><div>Dirección IPv4 pública</div><div>3.81.92.45 dirección abierta</div></div><div><div>Estado de la instancia</div><div>En ejecución</div></div><div><div>Nombre DNS de IP privada (solo IPv4)</div><div>ip-172-31-32-45.ec2.internal</div></div><div><div>Tipo de instancia</div><div>t2.micro</div></div></div> <div><div><div>Direcciones IPv4 privadas</div><div>172.31.32.45</div></div><div><div>DNS de IPv4 pública</div><div>ec2-3-81-92-45.compute-1.amazonaws.com dirección abierta</div></div><div><div>Direcciones IP elásticas</div><div>-</div></div></div>		

Instancia: i-04a9fccb3e2d33cbc (nueva_bigdata)

Rol de IAM
AmazonEMR-InstanceProfile-20231129T000658

IMDSv2
Required

ID de subred
subnet-08a1bd0a79707455b

ID de AMI
ami-0fc5d935ebf8bc3bc

Nombre de AMI
ubuntu/images/hvm-ssd/ubuntu-jammy-22.04-amd64-server-20230919

Nombre del grupo de Auto Scaling
-

Monitoreo
desactivado

Protección de terminación
desactivado

Instancia: i-04a9fccb3e2d33cbc (nueva_bigdata)

Detalles

Seguridad

Redes

Almacenamiento

Comprobaciones de estado

Monitoreo

Etiquetas

▼ Detalles de seguridad

Rol de IAM
AmazonEMR-InstanceProfile-20231129T000658

Grupos de seguridad
sg-000c3f1ce4f808e3f (launch-wizard-2)

ID del propietario
355616091291

Hora de lanzamiento
Sat Dec 02 2023 14:16:02 GMT-0500 (hora estándar de Colombia)

▼ Reglas de entrada

Q Filtrar reglas

< 1 >

	Name	Security group ID	Nombre del grupo de seguridad	ID de la VPC
<input type="checkbox"/>	-	sg-027ae2f65261ce347	launch-wizard-1	vpc-05479049deabba376
<input type="checkbox"/>	-	sg-000c3f1ce4f808e3f	launch-wizard-2	vpc-05479049deabba376
<input type="checkbox"/>	-	sg-07444132d55c65096	default	vpc-05479049deabba376
<input type="checkbox"/>	-	sg-07e724ccd90405b8d	ec2-test-bigdata	vpc-05479049deabba376
<input checked="" type="checkbox"/>	-	sg-03c3a34b2075cd464	ElasticMapReduce-master	vpc-05479049deabba376

Cluster EMR:

Nombre

Mi_clúster3_3.4


Versión de Amazon EMR [Información](#)

Una versión contiene un conjunto de aplicaciones que se puede instalar en el clúster.


emr-6.15.0

Paquete de aplicaciones


Spark
Interactive




Core
Hadoop




Flink




HBase




Presto



Trino



Custom



- ☐ Flink 1.17.1

☐ HCatalog 3.1.3

☐ Hue 4.11.0

☒ Livy 0.7.1

☐ Phoenix 5.1.3

☒ Spark 3.4.1

☐ Tez 0.10.2

☐ ZooKeeper 3.5.10
- ☐ Ganglia 3.7.2

☒ Hadoop 3.3.6

☒ JupyterEnterpriseGateway 2.6.0

☐ MXNet 1.9.1

☐ Pig 0.17.0

☐ Sqoop 1.4.7

☐ Trino 426
- ☐ HBase 2.4.17

☒ Hive 3.1.3

☐ JupyterHub 1.5.0

☐ Oozie 5.2.1

☐ Presto 0.283

☐ TensorFlow 2.11.0

☐ Zeppelin 0.10.1

- Configuración del Catálogo de datos de AWS Glue
- Utilice el Catálogo de datos de AWS Glue para proporcionar un meta-almacén externo a la aplicación.
- ☐ Usar para metadatos de la tabla de Hive

☐ Usar para metadatos de la tabla de Spark

Opciones del sistema operativo [Información](#)

- ☒ Versión de Amazon Linux



Grupos de instancias

Elegir un tipo de instancia por grupo de nodos



Flotas de instancias

Elegir cualquier combinación de tipos de instancia dentro de cada grupo de nodos

Grupos de instancias

Principal

Elegir tipo de instancia de EC2

c1.medium

2 vCore 1.7 GiB memoria 350 GiB almacenamiento

Precio bajo demanda: 0.130 USD por instancia/h

Precio de spot más bajo: \$0.044 (us-east-1c)

Acciones ▼



Usar varios nodos principales

Para mejorar la disponibilidad del clúster, utilice 3 nodos principales con la misma configuración y acciones de arranque. No puede utilizar varios nodos principales con flotas de instancias.

► Configuración de nodo - *opcional*

Central

Elegir tipo de instancia de EC2

c1.xlarge

8 vCore 7 GiB memoria 1690 GiB almacenamiento

Precio bajo demanda: 0.520 USD por instancia/h

Precio de spot más bajo: \$0.147 (us-east-1c)

Acciones ▼

► Configuración de nodo - *opcional*

Tarea 1 de 1

Eliminar grupo de instancias

Nombre

Tarea - 1

Elegir tipo de instancia de EC2

c4.large

2 vCore 3.75 GiB memoria

Únicamente EBS almacenamiento

Precio bajo demanda: 0.100 USD por instancia/h

Precio de spot más bajo: \$0.047 (us-east-1e)

Acciones ▼

Elija una opción

- ☒ **Establecer el tamaño del clúster manualmente**
Utilice esta opción si conoce los patrones de la carga de trabajo de antemano.

- ☐ **Utilizar escalado administrado por EMR**
Supervise las métricas clave de la carga de trabajo de modo que EMR pueda optimizar el tamaño del clúster y la utilización de los recursos.

- ☐ **Utilizar el escalamiento automático personalizado**
Para escalar mediante programación los nodos principales y los nodos de tarea, cree políticas de escalamiento automático personalizadas.

Configuración de aprovisionamiento

Establezca el tamaño del principal y tarea grupos de instancias. Amazon EMR intenta aprovisionar esta capacidad al lanzar el clúster.

Nombre	Tipo de instancia	Tamaño de instancia(s)	Utilizar la opción de compra de spot
Tarea - 1	c4.large	<input type="text" value="1"/>	<input type="checkbox"/>
Central	c1.xlarge	<input type="text" value="1"/>	<input checked="" type="checkbox"/>

Redes [Información](#)

Virtual Private Cloud (VPC) [Información](#)

Examinar

Crear VPC [↗](#)

Subred [Información](#)

Examinar

Crear subred [↗](#)

► Grupos de seguridad de EC2 (firewall)

Terminación del clúster [Información](#)

- ☒ Terminar manualmente el clúster
- ☐ Terminar automáticamente el clúster después de que finalice el último paso
- ☐ Terminar el clúster después del tiempo de inactividad (recomendado)
- ☐ Use la protección contra la terminación
Proteja sus instancias de EC2 frente a la terminación accidental.

Configuración de seguridad y par de claves de EC2: *opcional* [Información](#)

Configuración de seguridad

Seleccione la configuración del servicio de cifrado, autenticación, autorización y metadatos de instancia del clúster.



Examinar [↗](#)

Crear configuración de seguridad [↗](#)

Par de claves de Amazon EC2 para el protocolo SSH al clúster [Información](#)



Examinar

Crear par de claves [↗](#)

Rol de servicio de Amazon EMR Información

El rol de servicio es un rol de IAM que Amazon EMR asume para aprovisionar recursos y realizar acciones de nivel de servicio con otros servicios de AWS.

☒ Elegir un rol de servicio existente

Seleccione un rol de servicio predeterminado o un rol personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con otros servicios de AWS.

☐ Crear un rol de servicio

Deje que Amazon EMR cree un nuevo rol de servicio para que pueda conceder y restringir el acceso a los recursos de otros servicios de AWS.

Rol de servicio

AmazonEMR-ServiceRole-20231202T134707 ▼



Perfil de instancia de EC2 para Amazon EMR

El perfil de instancia asigna un rol a cada instancia de EC2 de un clúster. El perfil de instancia debe especificar un rol que pueda acceder a los recursos de los pasos y las acciones de arranque.

☒ Elegir un perfil de instancia existente

Seleccione un rol predeterminado o un perfil de instancia personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con sus recursos de Amazon S3.

☐ Crear un perfil de instancia

Deje que Amazon EMR cree un nuevo perfil de instancia para que pueda especificar un conjunto personalizado de recursos a los que tendrá acceso en Amazon S3.

Perfil de instancia

AmazonEMR-InstanceProfile-20231129T000658 ▼



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.

```
1 # Evaluar el desempeño del modelo
2 evaluator = MulticlassClassificationEvaluator(labelCol="Sexo", predictionCol="prediction",
3                                              metricName="accuracy")
4 accuracy = evaluator.evaluate(predictions)
5 print("Test Error = %g" % (1.0 - accuracy))
```

Test Error = 0.315919

accuracy: 0.6840812681402099
precision: 0.6844682021163748
recall: 0.6840812681402099
f1_score: 0.6821405394889593

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.

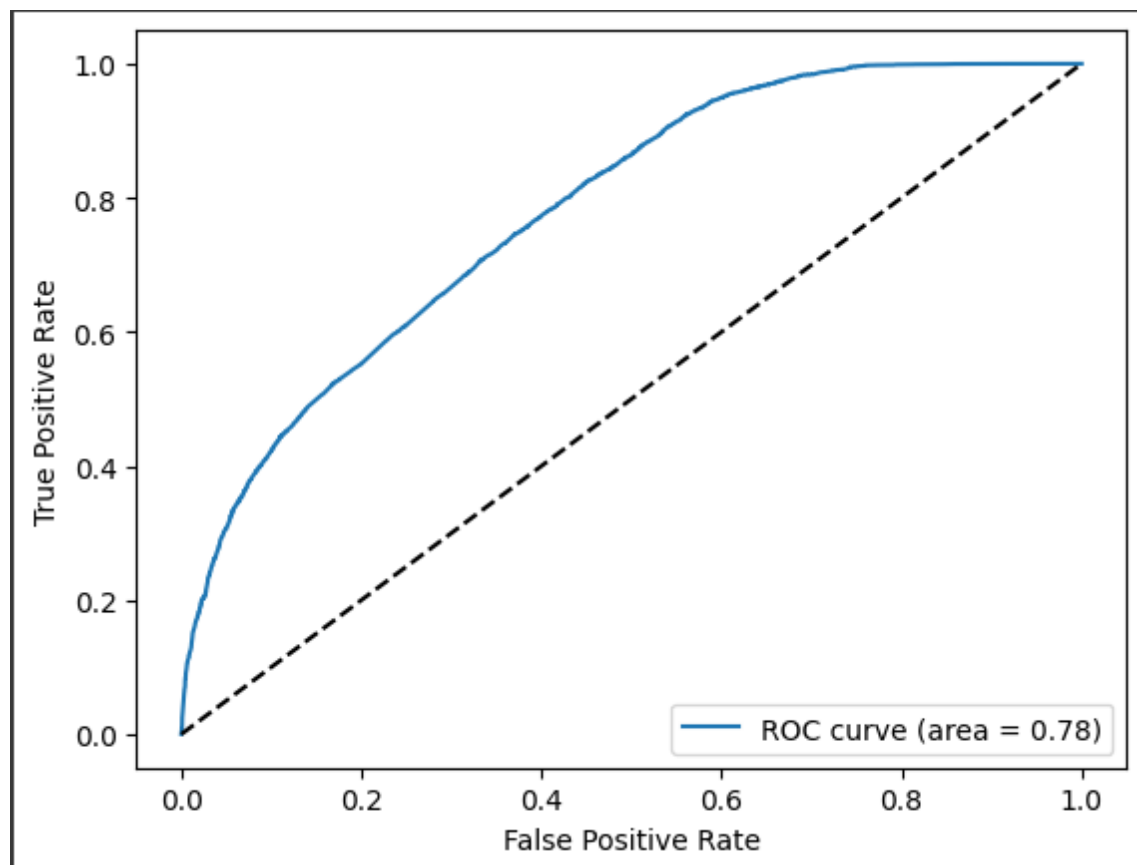
Recall (Sensibilidad)

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 puntuado debajo de la curva este tomara mejores predicciones en la clasificación de los datos binarios.