Análisis de asegurados en el Perú

En el Perú, el sistema de aseguramiento es fundamental para brindar protección y cobertura de salud a la población. Este informe analiza la situación actual de los asegurados en el país, proporcionando un panorama general de sus características y tendencias.

ERICK ESTEFANO RAMOS ZAPATA

Motivación y Audiencia

1 Motivación

En este proyecto, analizaremos datos sobre los afiliados al Seguro Integral de Salud en Perú para identificar tendencias y patrones que puedan ayudar a mejorar el acceso y la calidad de la atención médica.

2 Audiencia

Está dirigido a profesionales de la salud, responsables de políticas públicas preocupados por el acceso a la atención médica en Perú. Nuestros Objetivo es proporcionar información útil que pueda informar la toma de decisiones y promover mejoras en la atención médica para todos los ciudadanos.

Resumen de Metadata

Población Total	1048575 asegurados
Región, provincia y distrito	Información geográfica del domicilio del afiliado
Nacional - Extranjero	Información sobre la nacionalidad del afiliado y el país de procedencia en caso de ser extranjero.
Documento de Identidad	Tipo de documento de identidad del afiliado.
Edad y sexo	Información demográfica del afiliado
Plan de Seguro y cobertura financiera	Tipo de seguro SIS y cobertura financiera asociada.

Preguntas e Hipótesis

Pregunta

1

¿Existen variaciones significativas en la distribución demográfica de los afiliados al Seguro Integral de Salud, evidenciadas por diferencias en la edad, distribución geográfica y tipo de seguro, lo que podría influir en la accesibilidad y el uso de los servicios de salud en diferentes regiones?

2

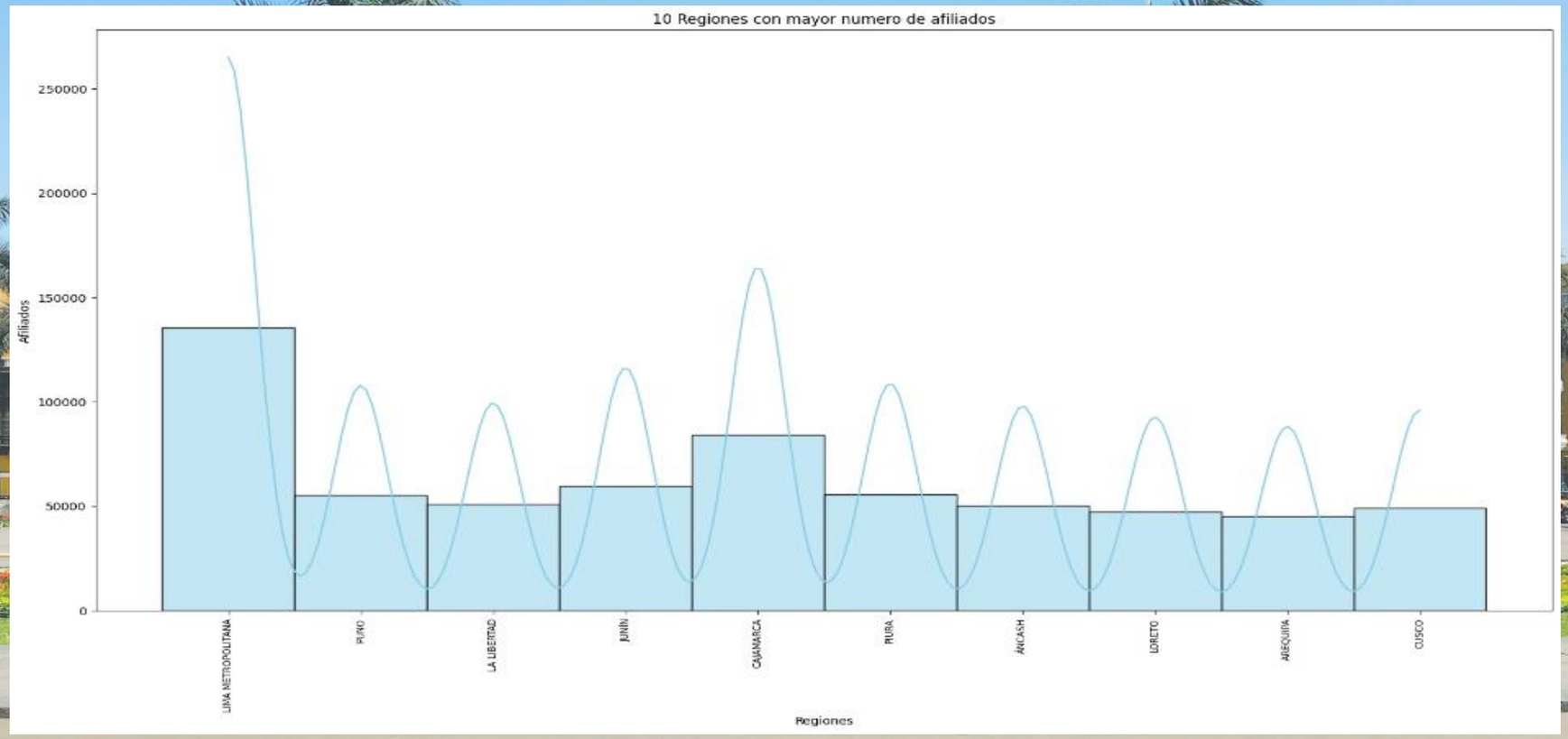
Hipótesis

Se espera que existan variaciones significativas en la distribución demográfica de los afiliados al Seguro Integral de Salud, evidenciadas por diferencias en la edad, distribución geográfica y tipo de seguro, lo que podría influir en la accesibilidad y el uso de los servicios de salud en diferentes regiones.

Análisis exploratorio(EDA)

En esta sección, realizaremos un análisis detallado de los datos del Seguro Integral de Salud.

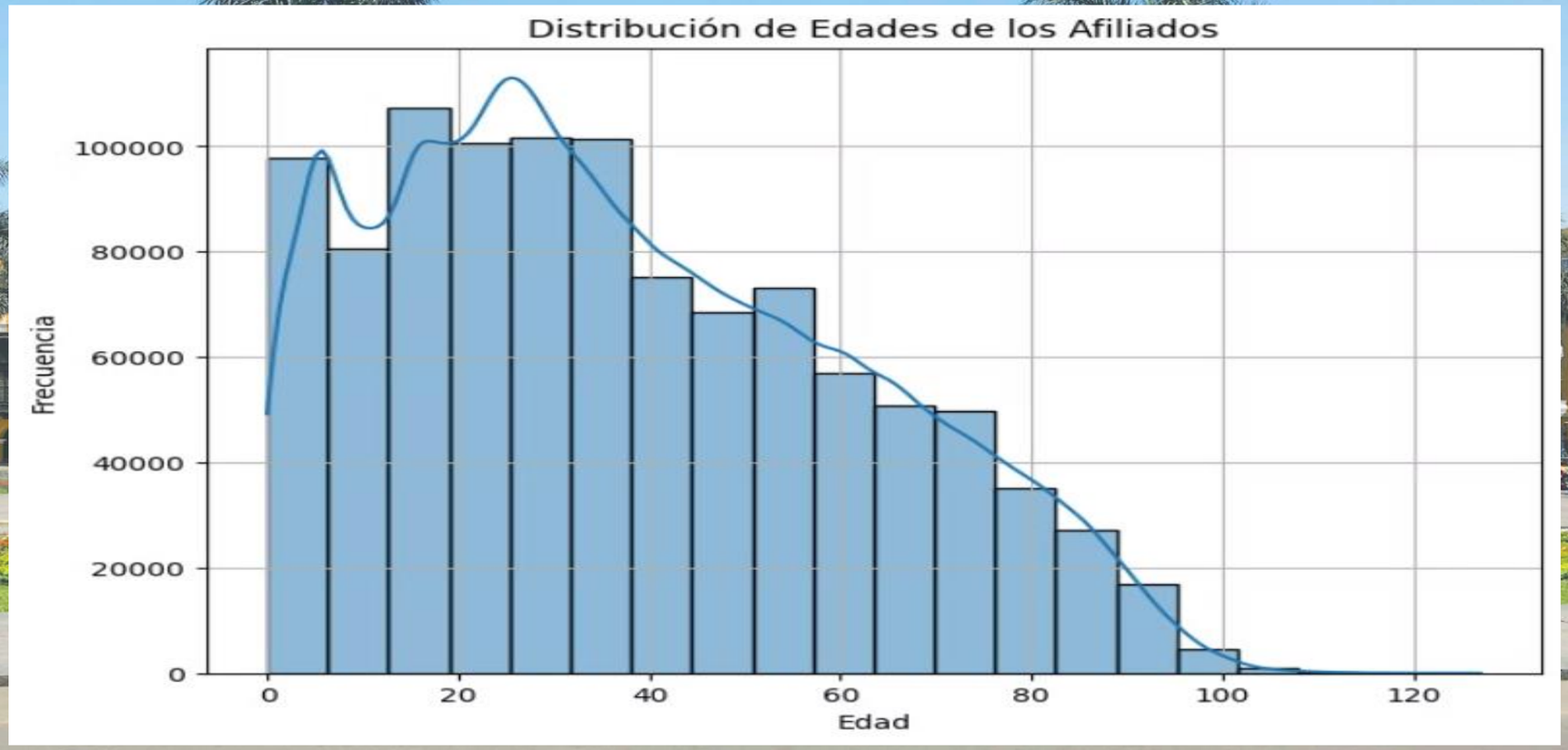
Región mas demandada



Se aprecia un gran numero de afiliados en general, en las regiones de Lima Metropolitana, Piura, Cajamarca.

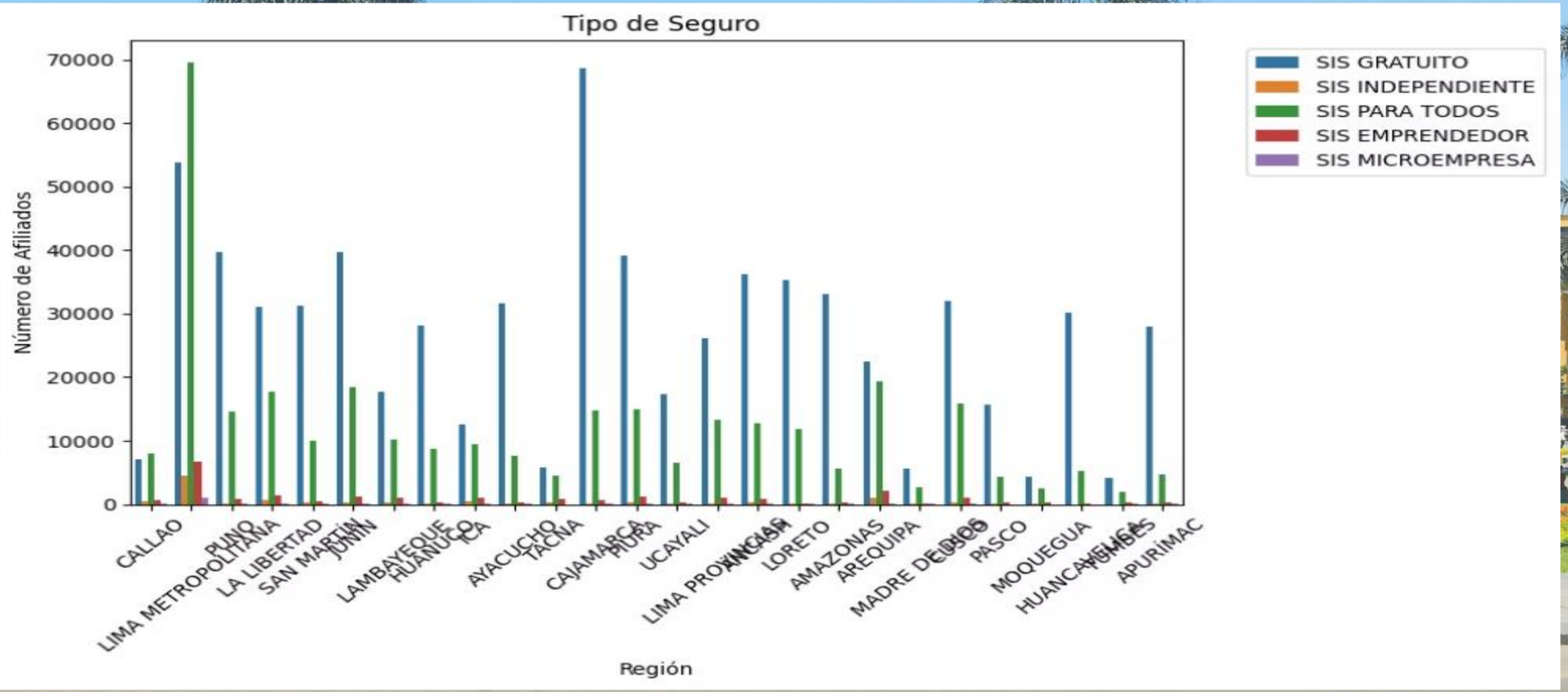
Esto puede deberse a que son las regiones con mas habitantes del Perú y, por lo tanto, es lógico esperar que también tengan más afiliados al Seguro Integral de Salud.

Distribución de afiliados por sus edades



Es importante considerar que hay una mayor cantidad de afiliados en tramos de edades más jóvenes, lo cual puede deberse a que este grupo poblacional suele tener menos acceso a servicios médicos privados y, por ende, al elegir el SIS como opción, aumentan las cifras de afiliados en el rango de 0-40 años.

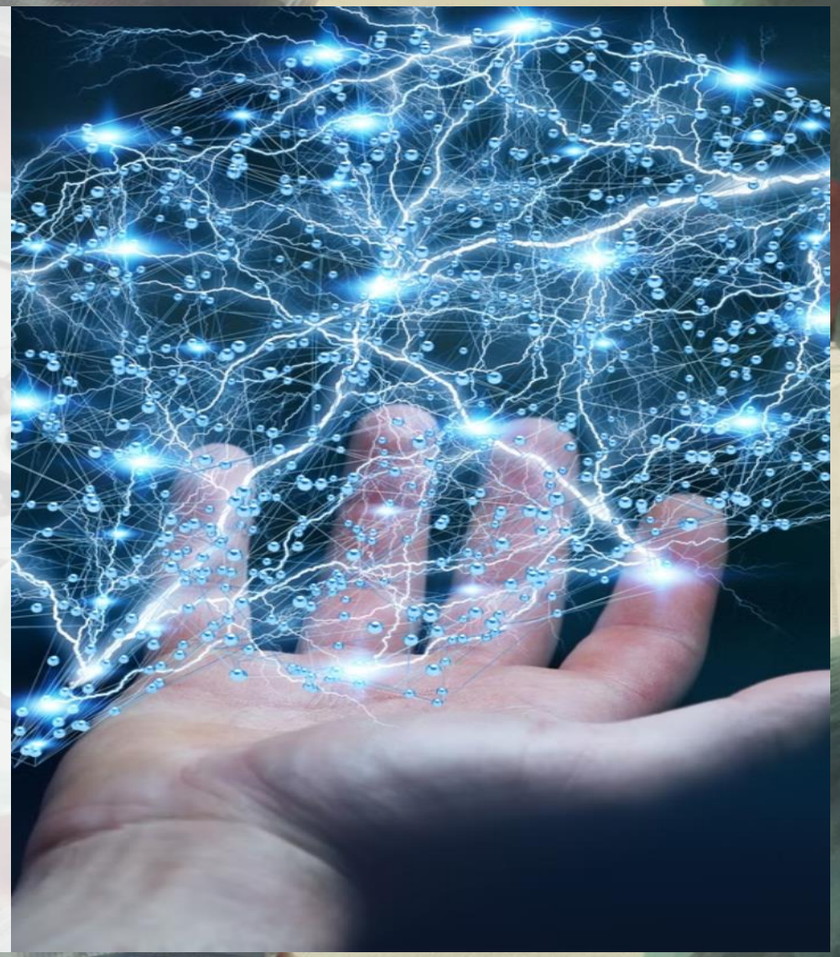
Tipos de seguro por región



El seguro que mas predomina en Cajamarca es el SIS GRATUITO, en Puno el SIS PARA TODOS. los seguros que menos representan son SIS MICROEMPRESA, SIS EMPRENDEDOR.

Estos datos sugieren que el acceso y la elección de seguros de salud varían significativamente según la región, lo que puede deberse a factores socioeconómicos y culturales.

Algoritmos de Aprendizaje Automático



Objetivos para nuestros modelos ML

Objetivo 1

1

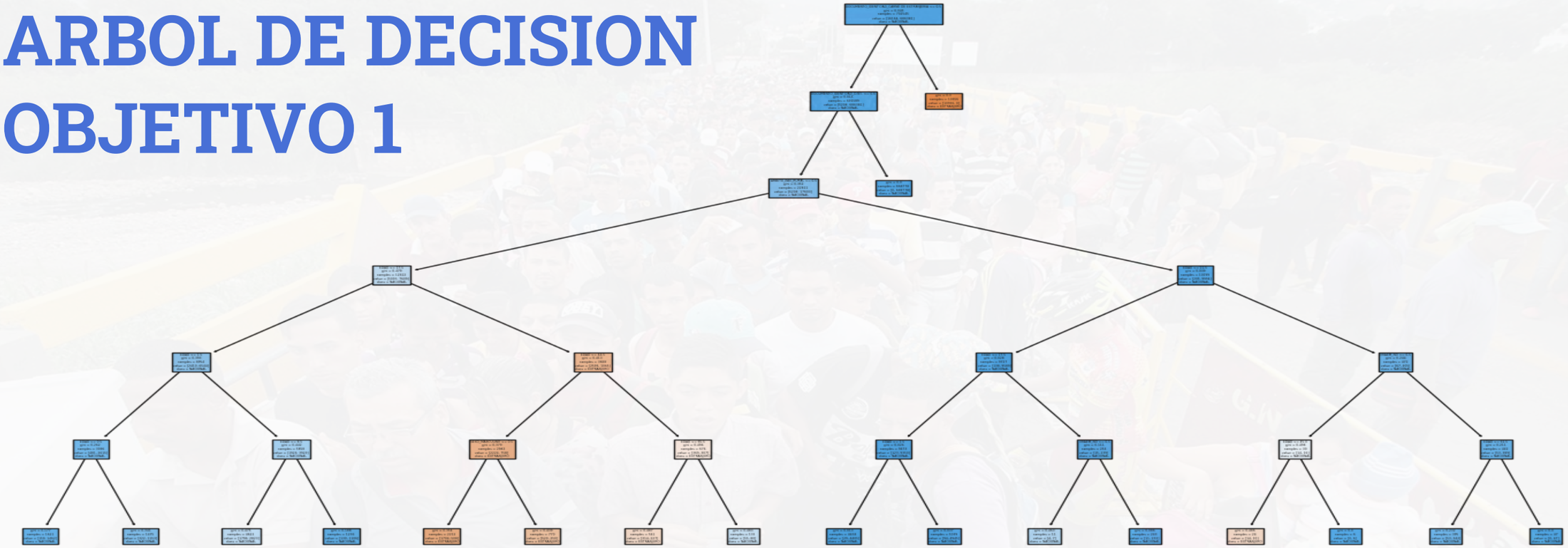
Uno de nuestros objetivos a futuro será implementar las técnicas vistas en clase para generar un modelo que nos ayude a predecir si el siguiente asegurado será de tipo Nacional o Extranjero

2

Objetivo 2

Nuestro segundo objetivo será analizar y entender la edad del asegurado y las predicciones a esta, para ello utilizaremos los modelos de Regresión logística, PCA.

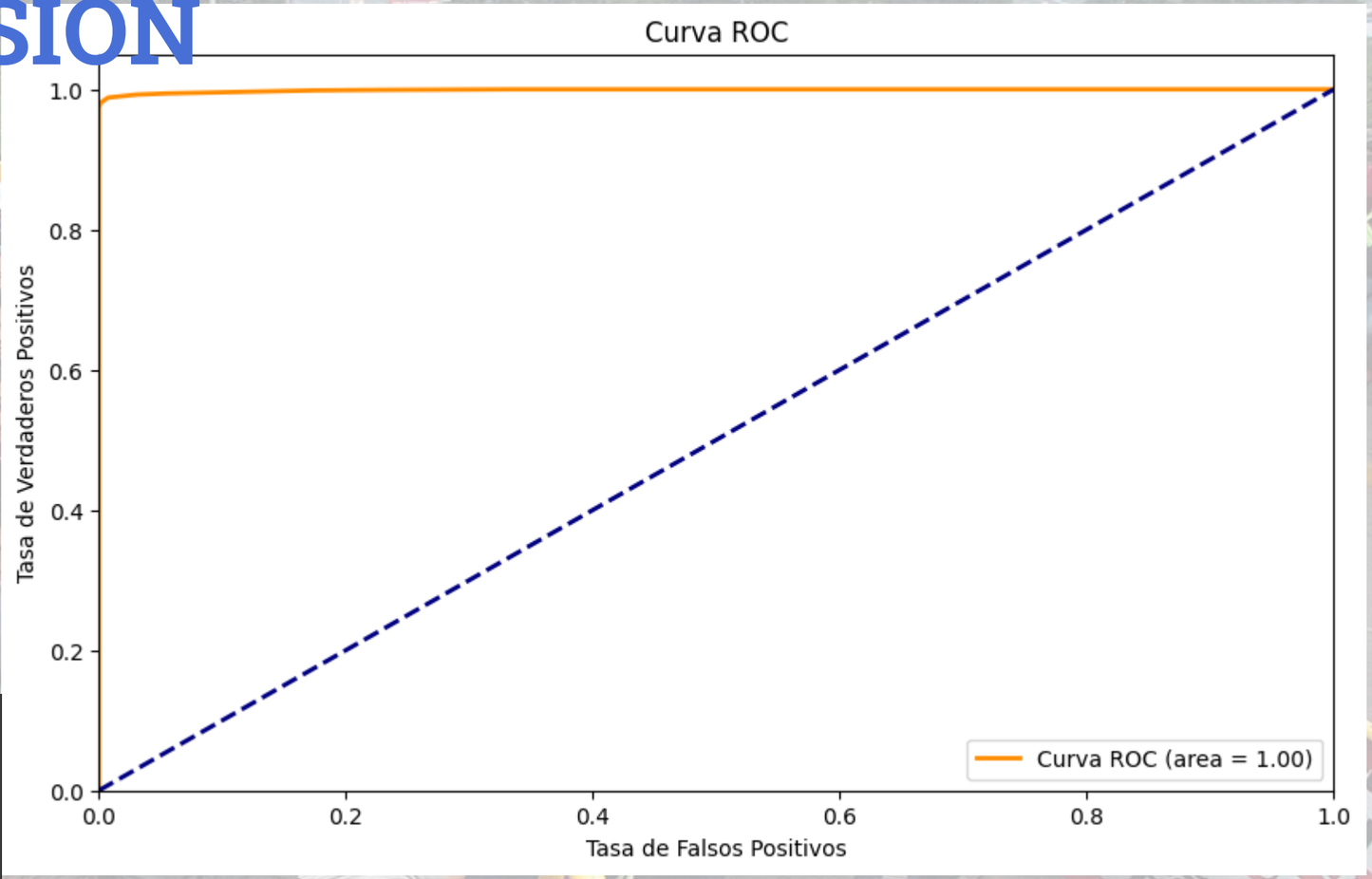
ARBOL DE DECISION OBJETIVO 1



```
# Dividir los datos en conjuntos de entrenamiento y prueba

X_train, X_test, y_train, y_test = train_test_split(X_NEW, Y_encoded, test_size=0.33, random_state=42)

# Entrenar el arbol

clf = tree.DecisionTreeClassifier(max_depth=6, min_samples_leaf=5, random_state=4)
clf = clf.fit(X_train, y_train)
```

El modelo aplicado Arbol de decision con una profundidad (depth=6) , y entrenado con 33% de nuestros datos. nos indica una precisión del 99% para predecir si el siguiente asegurado será Nacional o Extranjero. Esto significa que el algoritmo es altamente efectivo para clasificar y predecir datos en función de las características proporcionadas.

ARBOL DE DECISION




```
precisión: 0.994529376065659
Matriz de confusión:
[[  6551   1383]
 [   510 337586]]
Reporte de clasificacion:
              precision    recall  f1-score   support

           0       0.93      0.83      0.87      7934
           1       1.00      1.00      1.00    338096

    accuracy                           0.99    346030
   macro avg       0.96      0.91      0.94    346030
weighted avg       0.99      0.99      0.99    346030
```

	Model	Accuracy	Precision	Recall	ROC_curve
0	arbol_1	0.987287	0.999805	0.987181	0.997763

Podemos apreciar el area bajo la curva es 0.99 siendo nuestros modelo significativo para nuestros objetivos, Podemos obtener este resultado ya que nuestra investigacion cuenta solo con variables categoricas.

REGRESION LOGISTICA OBJETIVO 2

	Actual	Predicted	Sesgo	Error_porc
781974	6	40.656250	-34.656250	-577.604167
937737	29	33.523438	-4.523438	-15.598060
907828	33	40.578125	-7.578125	-22.964015
784628	65	40.875000	24.125000	37.115385
662460	11	40.578125	-29.578125	-268.892045
...	...	...	...	...
673443	46	33.523438	12.476562	27.122962
656736	15	33.617188	-18.617188	-124.114583
858501	25	40.656250	-15.656250	-62.625000
617079	56	33.304688	22.695312	40.527344
487559	53	40.671875	12.328125	23.260613

209715 rows × 4 columns

```
X1=df.drop(columns=['REGION','EDAD','FECHA_CORTE','PAIS_EXTRANJERO','PROVINCIA','DISTRITO','UBIGEO','NACIONAL_EXTRANJERO','CO
Y1=df['EDAD']

X_NEW1 = pd.get_dummies(X1, columns=['AMBITO_INEI','VRAEM','DOCUMENTO_IDENTIDAD','SEXO','REGIMEN_FINANCIAMIENTO','PLAN_DE_SEG

#Separacion en train y test
X_train, X_test, y_train, y_test = train_test_split(X_NEW1, Y1, test_size=0.3, random_state=42)

#Entrenamiento del modelo

regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

El modelo aplicado de regression logistica, para determinar la edad del asegurado, este modelo esta entrenado con 30% de datos para testear. Nos indica una precisión del 6% para predecir la edad del asegurado. Esto significa que el algoritmo no es efectivo para clasificar y predecir datos en función de las características proporcionadas.

El cual puede necesitar sobreajuste para intentar ajustar el modelo.

CROSS VALIDATION

```python
# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_NEW1, Y1, test_size=0.3, random_state=42)

# Entrenar el modelo de regresión lineal
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Realizar predicciones en el conjunto de prueba
y_pred = regressor.predict(X_test)

# Calcular R-squared
r_squared = r2_score(y_test, y_pred)
print("R-squared:", r_squared)

# Validación cruzada
kf = KFold(n_splits=5, shuffle=True, random_state=42)
cv_scores = cross_val_score(regressor, X_NEW1, Y1, cv=kf, scoring='r2')
```

```
R-squared: 0.06840988276425375
Cross-Validation R-squared Scores: [0.06755912 0.06944936 0.06766323 0.06875097 0.06739125]
Average Cross-Validation R-squared Score: 0.0681627859023547
```

El valor de $R2$ tanto en la evaluación directa como en la validación cruzada es bastante bajo (aproximadamente 6.8%), lo que indica que el modelo no está capturando bien la variabilidad de los datos. Un $R2$ bajo sugiere que el modelo no es muy eficaz para predecir la variable objetivo en este caso.

AJUSTE DE HIPERPARAMETROS DEL ARBOL DE DECISION

```python
# Definir los parámetros para el ajuste
param_dist = {
    'max_depth': [3, 5, 7, 10, 12],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 5],
    'criterion': ['gini', 'entropy']
}

# Realizar RandomizedSearchCV
random_search = RandomizedSearchCV(clf, param_distributions=param_dist, n_iter=50, cv=5, random_state=42, n_jobs=-1)
random_search.fit(X_train_bal, y_train_bal)
```

```
Mejores parámetros para Arbol de Decisión: {'min_samples_split'
              precision    recall  f1-score   support

           0       0.61      1.00      0.76      7934
           1       1.00      0.99      0.99    338096

    accuracy                           0.99    346030
   macro avg       0.81      0.99      0.88    346030
weighted avg       0.99      0.99      0.99    346030
```

Aplicando un ajuste de hiperparametros como la profundidad del arbol, numero minimo de muestras entre otros, se determina que el modelo tiene un accuracy del 99% un modelo muy ajustado para predecir.

AJUSTE DE HIPERPARAMETROS DEL REGRESION LOGISTICA

```python
# Definir el modelo
ridge_regressor = Ridge()

# Definir los parámetros para el ajuste
param_grid = {
    'alpha': [0.01, 0.1, 1, 10, 100]
}

# Realizar GridSearchCV
grid_search = GridSearchCV(ridge_regressor, param_grid, cv=5, n_jobs=-1)
grid_search.fit(X_train, y_train)
```

```
Mejores parámetros para Ridge Regression: {'alpha': 1}
R-squared: 0.06757848963579471
```

Aplicando un ajuste de hiperparametros como ALPHA, se determina que el modelo presenta un R2 de 6.7% siendo un modelo poco ajustado para las estimaciones y predicciones que nuetsra investigacion require. Dando por conclusion que ajustar los hiperparametros no mejora signifactivamente el modelo base.

COMPARACION DE MODELOS

ARBOL DE DECISION

	Model	Accuracy	Precision	Recall	ROC_curve
0	arbol_1	0.987287	0.999805	0.987181	0.997763
1	arbol_2	0.988122	0.999826	0.988015	0.998935

Nuestro modelo sin ajuste ya es muy optimo y aplicando un ajuste sigue contando con un 99% de precision. Aplicando este modelo podremos llevar nuestra investigacion para futuros procesos de predicciones

REGRESION LOGISTICA

	Metric	Modelo sin ajustar	Modelo ajustado
0	R-squared	0.067570	0.067578
1	MAE	19.409929	19.413530
2	MSE	550.292420	550.287156

Apreciamos que nuestros modelo no mejora significativamente al realizar un ajuste de hiper parametros, dejando como conclusion que el modelo no es optimo para nuestro studio.

Insights Clave

Distribución Geográfica de los Afiliados

La concentración de afiliados en Lima metropolitana sugiere una alta demanda deservicios de salud en esta área urbana densamente poblada

Tipo de Seguro y Cobertura Financiera

La popularidad del SIS PARA TODOS y el SIS GRATUITO sugiere una preferencia por los planes de seguro que ofrecen una cobertura más amplia y acceso gratuito a los servicios de salud.

Rangos de Edad

El predominio de afiliados en el rango de edad de 18 a 35 años destaca la importancia de garantizar la cobertura de seguro de salud para los jóvenes adultos.

Conclusiones y Recomendaciones

Mejora de la Infraestructura y Recursos en Lima Metropolitana

Dado el alto número de afiliados en Lima metropolitana.

Extensión de la Cobertura en Áreas Rurales

Para abordar las necesidades de áreas con grandes volúmenes de asegurados como Cajamarca, Junín y Piura.

Evaluación Continua y Adaptación de Políticas

para garantizar que estén cumpliendo con sus objetivos y abordando las necesidades cambiantes de la población

Ampliación de la Cobertura para Grupos Demográficos Específicos

Implementar programas específicos para ampliar la cobertura de seguro para grupos demográficos subrepresentados.