

# Resultados claves


## Integrantes

- Oscar Gomez A00394142
- Juan Sebastian Caviedez A00394958


## Repositorio

GitHub - LeonHartK/classificationWithSparkML

Contribute to LeonHartK/classificationWithSparkML development by creating an account on GitHub.

 <https://github.com/LeonHartK/classificationWithSparkML.git>

LeonHartK/  
**classificationWithSpark...**




2 Contributors

0 Issues

0 Stars

0 Forks



## Evidencias

### 1. Schema

```
Esquema del DataFrame:
root
|-- age: integer (nullable = true)
|-- sex: string (nullable = true)
|-- workclass: string (nullable = true)
|-- fnlwgt: integer (nullable = true)
|-- education: string (nullable = true)
|-- hours_per_week: integer (nullable = true)
|-- label: string (nullable = true)
```

Esquema y primeras filas del DataFrame. Muestra las columnas disponibles: age, sex, workclass, fnlwgt, education, hours\_per\_week, label. Estos son los datos de entrada usados para el modelo.

### 2. Primeros registros del Dataframe

```
Primeros 5 registros del DataFrame:
+---+---+---+---+---+---+
|age|sex|workclass|fnlwgt|education|hours_per_week|label|
+---+---+---+---+---+---+
|58|Male|Private|164194|HS-grad|34|>50K|
|65|Male|Gov|305929|Bachelors|57|<=50K|
|20|Male|Private|134629|HS-grad|52|>50K|
|53|Male|Gov|360726|Some-college|54|<=50K|
|32|Female|Gov|165852|Bachelors|30|<=50K|
```

### 3. Estadísticas

Estadísticas descriptivas de las columnas numéricas:

summary	age	fnlwgt	hours_per_week
count	2000	2000	2000
mean	41.7435	213280.405	40.0875
stddev	13.887769540909401	110668.86162896988	11.90465052414606
min	18	20129	20
max	65	399891	60

Estadísticas descriptivas de columnas numéricas. Observamos media de age ~41.7 y rango de hours\_per\_week entre 20 y 60.

#### 4. Predicciones

Para generar las predicciones, el conjunto de datos se divide en tres subconjuntos: **train**, **validation** y **test**.

- El conjunto de **train** se utiliza para entrenar el modelo.
- El conjunto de **validation** permite ajustar hiperparámetros y verificar la capacidad de generalización.
- Finalmente, el conjunto de **test** se reserva para evaluar el rendimiento real del modelo sobre datos nunca vistos.

Las métricas calculadas en **validation** muestran qué tan bien el modelo aprende sin sobreajustarse, mientras que las métricas en **test** reflejan su desempeño final y objetivo en un escenario real.

- Validation Results

age	sex	education	hours_per_week	label	prediction	probability
18	Female	Assoc	26	<=50K	0.0	[0.5526955698656754,0.44730443013432464]
18	Female	Assoc	51	<=50K	0.0	[0.5471886554335542,0.4528113445664458]
18	Female	Bachelors	35	<=50K	1.0	[0.44651042040500977,0.5534895795949902]
18	Female	Assoc	57	<=50K	0.0	[0.5498460423647564,0.4501539576352436]
18	Female	Bachelors	41	<=50K	1.0	[0.3943213053650058,0.6056786946349941]
18	Female	HS-grad	50	<=50K	0.0	[0.5662524074919937,0.4337475925080063]
18	Female	Bachelors	45	<=50K	1.0	[0.4118381118982134,0.5881618881017866]
18	Female	Bachelors	35	<=50K	1.0	[0.37224000990953177,0.6277599900904682]
18	Female	Some-college	45	>50K	1.0	[0.335198182045961,0.6648018179540389]
18	Female	Masters	39	<=50K	0.0	[0.5526734186475777,0.4473265813524223]

only showing top 10 rows

Predicciones: para cada registro se muestran la etiqueta real (label), la predicción (prediction) y la probabilidad asociada (probability).

- Test Results

age	sex	education	hours_per_week	label	prediction	probability
18	Female	11th	21	>50K	1.0	[0.3956138338668858,0.6043861661331142]
18	Female	HS-grad	39	>50K	1.0	[0.4970318507569324,0.5029681492430675]
18	Female	Bachelors	28	>50K	1.0	[0.4799461288098475,0.5200538711901526]
19	Female	Assoc	47	<=50K	0.0	[0.5143707925216413,0.4856292074783587]
19	Female	HS-grad	46	>50K	0.0	[0.6312660985576936,0.36873390144230644]
19	Male	Assoc	29	>50K	0.0	[0.5831675211558611,0.4168324788441389]
19	Male	Masters	27	>50K	1.0	[0.48472317165031414,0.5152768283496858]
19	Male	11th	26	<=50K	1.0	[0.3203741763310705,0.6796258236689294]
20	Female	Some-college	52	<=50K	0.0	[0.5909808669095199,0.4090191330904801]
20	Female	Masters	32	>50K	1.0	[0.38162425265265076,0.6183757473473492]

only showing top 10 rows

Predicciones: para cada registro se muestran la etiqueta real (label), la predicción (prediction) y la probabilidad asociada (probability).

- Colab

	real_label	predicted_label
0	>50K	>50K
1	>50K	<=50K
2	>50K	>50K
3	<=50K	<=50K
4	>50K	>50K
5	<=50K	<=50K
6	>50K	>50K
7	>50K	<=50K
8	<=50K	<=50K

Predicciones en ipynb: La columna real\_label muestra la clase real, predicted\_label la predicción del modelo.

## Pipeline

```
lr = LogisticRegression(
    featuresCol="features",
    labelCol="label_indexed",
    predictionCol="prediction",
    probabilityCol="probability",
    rawPredictionCol="rawPrediction",
    maxIter=100,
    regParam=0.01,
    elasticNetParam=0.0,
)

pipeline = Pipeline(stages=indexers + encoders + [assembler, lr])
```

```
model = pipeline.fit(df)
return model
```

- Primero, los indexadores y encoders preparan los datos (por ejemplo, convierten variables categóricas en numéricas).
- El ensamblador junta todas las características en un solo vector.
- Finalmente, se entrena el modelo de regresión logística.

## Comparación con otros modelos

- Logistic Regression (sin fnlwgt) fue el mejor

```
Métricas de evaluación:
Exactitud del modelo: 0.4882
+-----+-----+-----+
|label|prediction|count|
+-----+-----+-----+
|<=50K|      0.0|  408|
|<=50K|      1.0|  311|
|>50K|      0.0|  370|
|>50K|      1.0|  306|
+-----+-----+-----+
```

Validation Results: Se observa un **balance muy bajo**: el modelo confunde con frecuencia ambas clases.

- El modelo predice correctamente menos de la mitad de los casos, lo que indica que apenas está un poco por encima de un clasificador aleatorio.
- <=50K (verdaderos negativos): 408 predicciones correctas frente a 311 incorrectas.
- >50K (verdaderos positivos): 306 correctas frente a 370 incorrectas.

```
Métricas de evaluación:
Exactitud del modelo: 0.5066
+-----+-----+-----+
|label|prediction|count|
+-----+-----+-----+
|<=50K|      0.0|   65|
|<=50K|      1.0|   59|
|>50K|      0.0|   57|
|>50K|      1.0|   48|
+-----+-----+-----+
```

Test Results: El modelo predice las dos clases casi en proporciones similares, pero con mucha confusión.

- El rendimiento mejora apenas un poco en test, pero sigue siendo bajo (cercano al azar).
- <=50K: 65 correctos vs 59 incorrectos.
- >50K: 48 correctos vs 57 incorrectos.

Se observa que el modelo alcanza una exactitud cercana al 50%, lo que refleja bajo poder predictivo y gran confusión entre las clases <=50K y >50K. Las

probabilidades asignadas se concentran alrededor de 0.5, indicando indecisión del modelo.

En cuanto al modelo de regresión logística y otros modelos se observó que:

- Random Forest, SVM y GBT no superaron a la regresión logística, incluso se quedaron cerca de  $AUC = 0.5$
- Con `fnlwgt`, todos los modelos empeoraron

## Conclusiones

- El dataset es muy pequeño y con pocas features útiles.
- 'fnlwgt' no sirve en este problema, mejor descartarlo aunque no presenta un cambio significativo.
- Las variables **age, education, hours\_per\_week y sex** aportan algo, pero no lo suficiente para un modelo robusto.
- Los modelos no mejoran porque necesitan más datos y más diversidad de variables.