Comparación de algoritmos de clasificación para la predicción de ingresos a partir del conjunto de datos *Census Income*

Bryan Buiza Mogrovejo
Escuela Profesional de Ingeniería Industrial
Universidad Católica de Santa María
Arequipa, Perú
bryan.buiza.6@gmail.com

Luis Guanilo Quiñones
Escuela Profesional de Ingeniería Industrial
Universidad Nacional de Trujillo
Trujillo, Perú
luigqfer@gmail.com

José Alfredo Tuyo Llipita
Escuela Profesional de Ingenieria en Informática y Sistemas
Universidad Nacional Jorge Basadre Grohmann
Tacna, Perú
jose.tuyo90@gmail.com

Elias Antonio Apaza Ramos
Departamento Académico de Estadística e Informática
Universidad Nacional Agraria la Molina
Lima, Perú
apazaramose@gmail.com

Roberto Alfredo Méndez Vilca
Escuela Profesional de Ingeniería de Computación y Sistemas
Universidad de San Martín de Porres
Lima, Perú
robertomendez.vilca@gmail.com

***Abstract*—Hoy en día las ganancias anuales obtenidas por la población asalariada es un punto a analizar, existe una marcada desigualdad respecto a la riqueza e ingresos, el presente artículo toma como punto de partida un dataset de un censo realizado en los EE. UU. En el cual se analiza diversas variables de personas que ganan más o menos de 50k anualmente.**

Partiendo de esa premisa, se analizarán diversos algoritmos de clasificación y así obtener el que mejor se ajuste a nuestros datos. Una vez aplicados estos algoritmos tendremos como resultado la predicción de ver si el ingreso anual de una persona en EE.UU. está en la categoría de más de 50k o menor o igual al 50k en función a las variables que se analizarán.

Como resultado final se obtuvo que el algoritmo de Random Forest Classifier y Support Vector Classifier obtuvieron una precisión del 84% respecto a los demás que obtuvieron menores porcentajes.

Keywords—Machine learning, accuracy, logistic regression, support vector classifier, random forest classifier, decisión tree, Gaussian Naive Bayes, K-nearest neighbors, adult income dataset.

I. Introducción

El conjunto de datos a analizar será del censo realizado en los EE.UU. el cual cuenta con 48,842 instancias y 15 atributos.

En esta investigación se realizará el análisis de que algoritmo de clasificación es más certero respecto a otros.

Como primer paso realizaremos el análisis exploratorio de los datos seleccionado las variables que más influyen sobre nuestra variable dependiente (***income***), segundo, a partir de ello formularemos el problema y la hipótesis que probaremos, seguidamente verificaremos trabajos anteriores similares a nuestra investigación los cuales plasmaremos en los antecedentes.

Continuando con el paper, realizaremos el pre procesamiento de los datos, para que estén limpios y así tratarlos adecuadamente. En la sección de modelos propuestos se encontrarán todos los algoritmos de clasificación que se compararán para encontrar la que mejor se ajuste a nuestra data, seguidamente haremos la comparación de dichas técnicas y finalmente veremos las conclusiones, recomendaciones y referencias bibliográficas.

II. Analisis exploratorio

A. *Descripción de las variables*

Primero realizamos un análisis descriptivo de las variables, de esta forma logramos entender qué tipo de datos son y cómo se comportan. En total hay 15 variables, 6 de ellas son numéricas (age, fnlwgt, educational-num, capital-gain, capital-loss, hours-per-week) y el resto son categóricas (workclass, education, marital-status, occupation, relationship, race, gender, native-country, income). A continuación una descripción de cada columna:

1. Age: Representa la edad en números enteros.
2. Fnlwgt: Es una variable numérica continua, no se tiene más información de ella.
3. Educational-num: Representa el grado académico obtenido o el último grado conseguido. Es del tipo ordinal.
4. Capital-gain: Representa las ganancias de capital debido al ahorro.
5. Capital-loss: Son las pérdidas de capital que tiene una persona.
6. Hours-per-week: Variable numérica entera que representa la cantidad de horas trabajadas a la semana.
7. Workclass: Es el sector en el que trabaja la persona, por ejemplo el sector público.
8. Education: Es igual a la variable educational-num, en este caso está expresada en categorías.
9. Marital-status: Expresa el estado civil de una persona. Variable categórica.

10. Occupation: Es la actividad que realiza la persona. Categorías.
11. Relationship: Representa el tipo de relación que tiene la persona, además de incluir a su pareja sentimental puede incluir información sobre sus hijos.
12. Race: Es la raza o etnia.
13. Gender: Género de la persona. Puede ser hombre o mujer.
14. Native-country: Representa el país de origen de la persona. Tiene 42 categorías posibles.
15. Income: Es nuestra variable objetivo y explica si una persona gana una cantidad mayor o menor igual que 50 mil dólares al año.

B. Visualización de datos

Dentro del dataset de tienen distintos tipos de datos, tanto cuantitativos como cualitativos.

Para los datos cualitativos, se hizo uso del histograma para conocer que distribución existe (Fig 1).

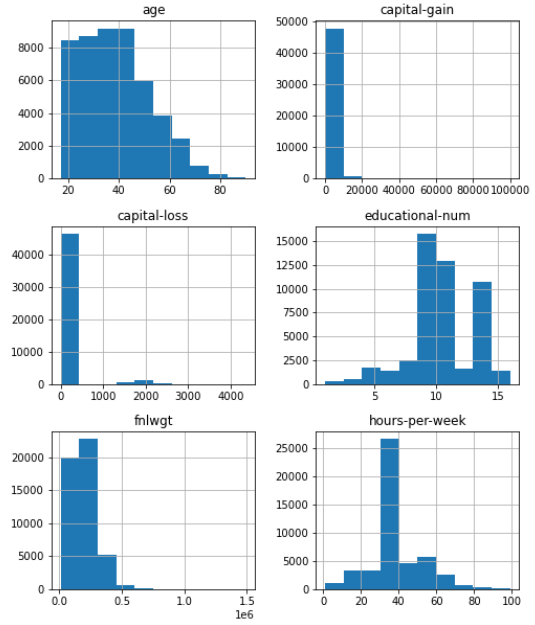

Fig 1. Histograma de variables cuantitativas
Fuente: Elaboración propia.

III. PROBLEMA

¿Qué algoritmo de clasificación se ajusta más para realizar la predicción de ingresos de una persona a partir del conjunto de datos *Census income*?

IV. HIPÓTESIS

Ho: El modelo de clasificación Random Forest obtiene mejores resultados respecto a los demás modelos clásicos.

Ha: El modelo de clasificación Random Forest no obtiene mejores resultados respecto a los demás modelos clásicos.

V. ANTECEDENTES

Explorando en la web, se han realizado diversos esfuerzos por parte de otros investigadores haciendo uso de modelos de aprendizaje automático y así predecir los niveles de ingresos, por ejemplo:

N. Chakrabarty et. al. [1] en su artículo tiene como objetivo identificar los factores clave para la mejora de los ingresos de las personas, además utilizan el método de extra tree clasifier para las variables continuas y categóricas. Además, los autores usan el diagrama de correlación de Pearson para variables continuas, gradient boosting classifier para el ajuste de hiperparametros, obteniendo los siguientes resultados.

- Accuracy: 88.16%
- Área debajo de la curva ROC: 0.93

Por otra parte, C. Lemon et. Al. [2] hace una comparación de diversas técnicas de aprendizaje automático, para ello usa distribuciones de frecuencia para identificar variables que ayuden en el problema de clasificación.

Usaron Baselines, Naive Bayes, Logistic Regression, Decision Tree, con los siguientes resultados (se midieron con el porcentaje de error obtenido de los modelos)

Classifier	Train Error	Test Error
Baselines	24.008%	23.623%
Naive Bayes	19.893%	20.432%
Naive Bayes (Grouped)	22.353%	24.128%
Logistic Regression	39.370%	38.612%
Decision Tree	18.940%	14.778%

Figura 01: Porcentaje de erros en la data de prueba y entrenamiento.

Fuente: C. Lemon, Chris Zelaso, K. Mulakaluri: "Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques"

Por último, S. Mishra [3] utiliza diversos algoritmos de clasificación, al igual que C. Lemon, compara modelos de clasificación, sin embargo, Mishra los compara por el resultado de accuracy y así ve cual es más idóneo para el dataset, obteniedo los siguientes resultados.

- Logistic Regression: 80.18%
- Random Forest Classifier: 85.25%
- Boosting Gradient Descent: 86.99%
- BernoulliNB Classifier: 73.25%
- Support Vector Classifier: 74.98%

Es así como podemos observar que existen otras investigaciones que nos sirven como base y guía para realizar el presente estudio.

VI. EXPLORACIÓN DE DATOS

En este punto realizaremos el análisis exploratorio de datos, esto con el fin de conocer nuestra data y conocer más sobre la dependencia de variables.

Como primer punto realizaremos una descripción mediante gráficos de barras mostrando las tendencias de nuestras variables y la relación entre estas:

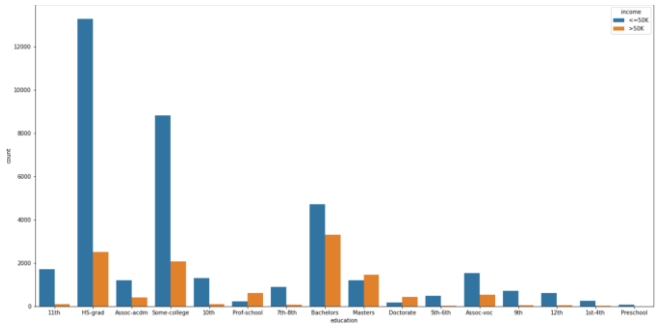

Figura 02: Relación entre el nivel educativo (*education*) y el *income*

Fuente: Elaboración propia.

Según el gráfico anterior se observa que las barras azules son las personas que ganan menor o igual a 50k y se puede apreciar que se dispara en personas que tienen solo estudios secundarios, abandonaron la escuela o que son solo bachilleres.

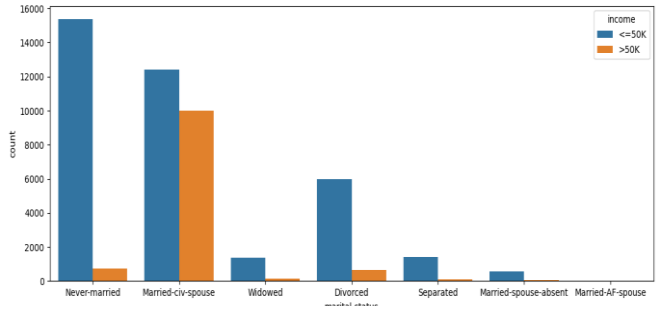

Figura 03: Relación entre el estado civil (*realtionship*) y el *income.*

Fuente: Elaboración propia

Según la figura 3, en relación a estas dos variables se observa que los que ganan más de 50k son los que están en el estado de "Married-civ-spouse", sin embargo, también tienen un gran porcentaje de los que ganan menos o igual a 50k.

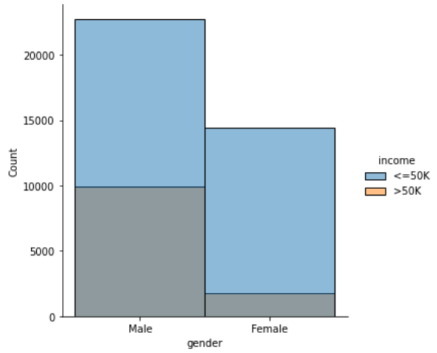

Figura 04: Relación entre el género (*gender)* y el *income.*

Fuente: Elaboración propia

Realizando un análisis más exhaustivo observamos que el porcentaje de personas que gana más de 50 mil dólares al año es mayor en los hombres (casi 50%) mientras que en el caso de las mujeres no alcanza el 25%. Sucede algo parecido cuando diferenciamos la distribución de las edades por income, la gente que gana más de 50 mil dólares al año tiene una media próxima a 45 años, por el otro lado, los que ganan menos o igual a 50 mil dólares una media cercana a 50 años.

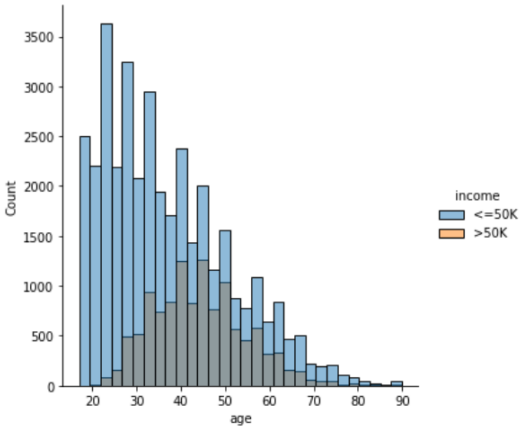

Figura 05: Relación entre la edad (*age)* y el *income.*

Fuente: Elaboración propia

En esta relación observamos que los que empiezan a trabajar desde una temprana edad ganan menos de 50k, sin embargo, también se observa un pequeño porcentaje que ganan más de 50k, lo que es muy interesante y cabe analizar los demás factores que los hacen tener esta gran ganancia anualmente.

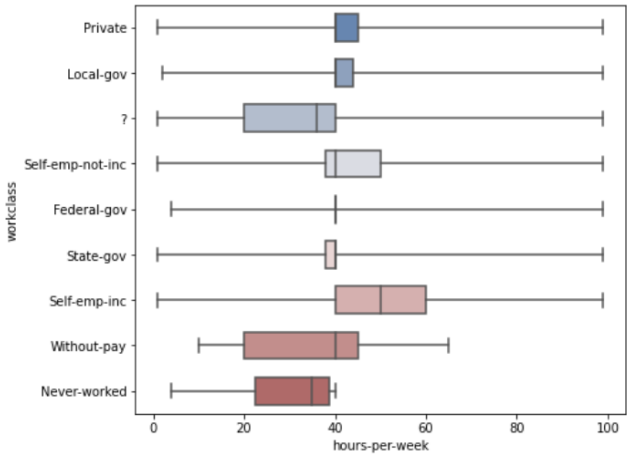

Figura 06: Análisis de las horas trabajadas por semana y el sector de trabajo

Fuente: Elaboración propia

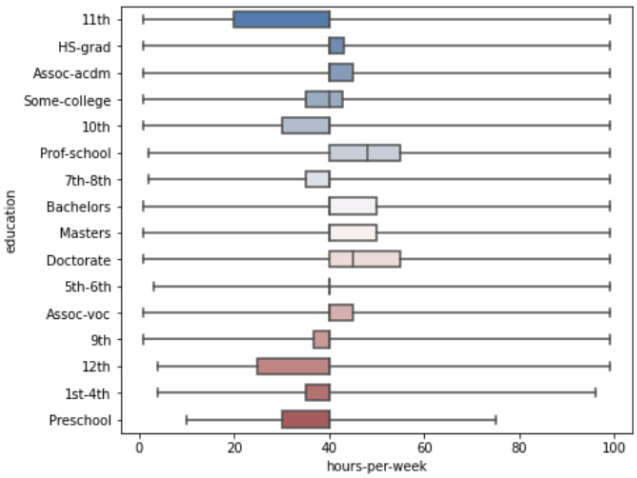

Figura 07: Análisis de las horas trabajadas por semana y el nivel educativo

Fuente: Elaboración propia

Comparando las horas por semana trabajas según el workclass podríamos inferir que los trabajos del sector público están muy próximos a las 40 horas, mientras que los trabajadores independientes trabajan más que eso. En el caso de la gente que se encuentra en un primer empleo y los que no reciben remuneración, trabajan menos de 40 horas a la semana.

VII. PREPROCESAMIENTO

En el preprocesamiento realizaremos la limpieza de datos, eliminando todos los outliers y datos que no aportan a nuestra investigación y que nos puede generar que nuestros resultados no sean acertados. Posteriormente agrupamos las categorías en variables como education y marital-status, facilitando el procesamiento por los algoritmos. De igual manera agrupando variables numéricas age y hours-per-week obtendríamos mejores resultados.

Realizando un barrido se observó que existen intancias con valores "?" los cuales eliminaremos para poder trabajar con las instancias con datos completos. Las variables workclass, occupation y native country contienen estos valores por los cual los reemplazaremos con NaN y procedemos a eliminarlos.

Luego hicimos un label-encoding a todas las variables categóricas y dividimos las variables en explicativas y objetivo (income). A continuación, divimos el dataset en una muestra para train y otra para test, tomando las proporciones 60% y 40 % respectivamente.

Además, se realizó una estandarización de datos en todas las variables, de esta forma podríamos facilitar el procesamiento. Se eliminaron de educación, país de origen, edad, horas de trabajo para evitar sesgos en la data.

VIII. MODELO PROPUESTO

Los modelos propuestos para elegir la mejor técnica de clasificación mediante el accuracy (exactitud) serán los siguientes:

- Decision tree:

```
The accuracy of the Decision tree model is  0.8105478467576981
              precision    recall  f1-score   support

       <=50k       0.87      0.88      0.87     13541
        >50k       0.63      0.59      0.61      4548

    accuracy                           0.81     18089
   macro avg       0.75      0.74      0.74     18089
weighted avg       0.81      0.81      0.81     18089
```

El árbol de decisión será el primer modelo y su accuracy es de 81.05%.

- Gaussian Naive Bayes:

```
The accuracy of Gaussian Naive Bayes model is 0.7969484216927415
              precision    recall  f1-score   support

       <=50k       0.82      0.94      0.87     13541
        >50k       0.67      0.38      0.48      4548

    accuracy                           0.80     18089
   macro avg       0.74      0.66      0.68     18089
weighted avg       0.78      0.80      0.78     18089
```

El Gaussian Naive Bayes obtuvo un accuracy de 80%.

- K-Nearest Neighbors:

```
The accuracy of the KNN Model is 0.8262479960196805
              precision    recall  f1-score   support

       <=50k       0.87      0.91      0.89     13541
        >50k       0.68      0.59      0.63      4548

    accuracy                           0.83     18089
   macro avg       0.77      0.75      0.76     18089
weighted avg       0.82      0.83      0.82     18089
```

El método de K-Nearest Neighbors obtuvo un accuracy de 82.62%.

- Suport Vector Classifier:

```
The accuracy of SVC model is 0.8402343965946155
              precision    recall  f1-score   support

       <=50k       0.86      0.94      0.90     13541
        >50k       0.75      0.55      0.63      4548

    accuracy                           0.84     18089
   macro avg       0.81      0.74      0.77     18089
weighted avg       0.83      0.84      0.83     18089
```

La técnica del Support Vector Classifier obtuvo un accuracy de 84.02%.

- Logistic Regression:

```
The accuracy of the Logistic Regression model is 0.829067389020952
              precision    recall  f1-score   support

       <=50k       0.85      0.93      0.89     13541
        >50k       0.72      0.53      0.61      4548

    accuracy                           0.83     18089
   macro avg       0.79      0.73      0.75     18089
weighted avg       0.82      0.83      0.82     18089
```

La Logistic Regression obtuvo un accuracy de 83%.

- Random Forest Model

```
The accuracy of the Random Forest Model is 0.8428326607330422
              precision    recall  f1-score   support

       <=50k       0.88      0.92      0.90     13541
        >50k       0.72      0.62      0.66      4548

    accuracy                           0.84     18089
   macro avg       0.80      0.77      0.78     18089
weighted avg       0.84      0.84      0.84     18089
```

El Random Forest Model obtuvo un accuracy de 84.28%.

IX. Comparación de métodos

Una vez ejecutados las técnicas de clasificación se procede a realizar la comparación de dichos métodos, por tal motivo presentamos la siguiente tabla.

Tabla 01:
Comparación de métodos

	Classification	Accuracy
1	Logistic Regression	83.00 %
2	Support Vector Classifier	84.02 %
3	Random Forest Classifier	84.28 %
4	Decision Tree	81.05 %
5	Gaussian Naive Bayes	80.00 %
6	K-Nearest Neighbors	82.62 %

Fuente: Elaboración propia.

Con dicha tabla podemos observar que el accuracy es muy alto para todos los métodos, sin embargo, necesitamos saber cuál es el mejor, por lo tanto, observamos que el que tienen un mayor porcentaje de exactitud a la hora de predecir futuros valores en el método de Random Forest Classifier.

Resultados y Conclusiones

- Podemos concluir en que el método de Random Forest Classifier es el mejor para poder predecir si una persona ganara más o menor igual a 50k en base a las variables que la afectan.
- Como observamos en la comparación de métodos, los demás métodos no se quedan atrás y también tienen un alto porcentaje de exactitud a la hora de predecir valores futuros.
- Se recomienda que en futuras investigaciones sobre este dataset se analice las variables en su totalidad para así obtener con más exactitud

Referencias

[1] N. Chakrabarty, S. Biswas: "A Statistical Approach to Adult Census Income Level Prediction", https://arxiv.org/ftp/arxiv/papers/1810/1810.10076.pdf

[2] C. Lemon, Chris Zelaso, K. Mulakaluri: "Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques", http://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf

[3] S. Mishra: "Classification Algorithms for the prediction of Income from Adult Census Income Dataset", https://www.academia.edu/40250231/Classification_Algorithms_for_the_prediction_of_Income_from_Adult_Census_Income_Dataset

[4] Kaggle Adult Census Income Data Set: https://www.kaggle.com/uciml/adult-census-income