Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 153-165
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 05/06/2023 | Aceptado: 01/09/2023 | Publicado: 30/09/2023

Identificadores persistentes:
DOI: 10.48168/innosoft.s12.a112
ARK: ark:/42411/s12/a112
PURL: 42411/s12/a112

Aplicación de árboles de decisión para un pronóstico salarial

Application of decision trees for wage forecasting

Corina Hancco Vargas [1][0000-0001-6879-882X], **David Flores Silva** [2][0000-0002-3414-1077], **Jessica Hancco Velásquez** [3][0000-0003-3664-6121], **Alejandra Fernandez Ninahuaman** [4][0000-0003-3605-7554]*

[1] Universidad Nacional de San Agustín. Perú. chanccov@unsa.edu.pe
[2] Universidad Nacional de San Agustín. Perú. dfloressi@unsa.edu.pe
[3] Universidad Nacional de San Agustín. Perú. jhanccove@unsa.edu.pe
[4] Universidad Nacional de San Agustín. Perú. mfernandezn@unsa.edu.pe

* Autor para correspondencia: mfernandezn@unsa.edu.pe

Resumen
Este artículo detalla el proceso que se realizó para el pronóstico salarial en una base de datos de un censo dado en 1996, donde se utilizó el lenguaje de programación Python, para el análisis de los datos del dataset se utilizó el servidor Google Colab para ejecutar los algoritmos en la nube, ya que el equipo considero que la velocidad de análisis de datos en Google Colab es más rápido. También se hizo uso de una de las técnicas de minería de datos para clasificar las variables usando árboles de decisión que tienen la capacidad de representar gráficamente varias soluciones alternativas con el fin de determinar los cursos/rutas de acción más efectivos para la clasificación de la obtención del sueldo de una persona.

Palabras clave: Árboles de decisión, Minería de datos, Pronóstico salarial.

Abstract
This article details the process that was carried out for the salary forecast in a database of a census given in 1996, where the Python programming language was used, for the analysis of the data of the dataset the Google Colab server was used to execute the algorithms in the cloud, since the team considered that the speed of data analysis in Google Colab is faster. One of the data mining techniques was also used to classify the variables using decision trees that have the ability to graphically represent several alternative solutions in order to determine the most effective courses/routes of action for the classification of the obtainment. of a person's salary.

Keywords: *Decision trees, Data mining, Salary forecast.*

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 153-165
https://revistas.ulasalle.edu.pe/innosoft

Introducción

La situación laboral ha cambiado en todo el mundo a causa de la pandemia por el Covid-19, según noticias dadas por las Naciones Unidas [1], se predecía que para finales del 2021 se habrían perdido 125 millones de empleos y la recuperación desigual atribuiría al crecimiento de la brecha entre los mercados laborales entre países industrializados y las naciones en desarrollo, diferencias entre los países pobres y ricos. Sin embargo, desde antes de la pandemia ya había conocimiento de los problemas en el mercado laboral, tanto en la búsqueda como también en el tipo de selección que se daba para los empleados. Como ejemplo de ello, un informe de la Organización Internacional del Trabajo (OIT), analizó que el principal problema de los mercados laborales en el mundo era el empleo de mala calidad donde millones de personas, a pesar de tener un trabajo fijo, se veían obligadas a aceptar condiciones de trabajo deficientes. Y se predecía que para 2018, la mayoría de los 3300 millones de personas empleadas en el mundo no gozarían de un nivel suficiente de seguridad económica, bienestar material o igualdad de oportunidades. Es más, el avance de la reducción del desempleo a nivel mundial no se veía reflejado en la mejora de la calidad del trabajo. Lo que en general se preveía por diversos déficits de trabajo decente, y se advertía de que, a ese ritmo, la consecución del objetivo de trabajo decente para todos, establecido entre los Objetivos de Desarrollo Sostenible (ODES), sería inalcanzable para muchos países [2].

En el artículo de Madero [3] titulado “Factores relevantes del desarrollo profesional y de compensaciones en la carrera laboral del trabajador” se pudo concluir que destacaban diferencias en el apoyo, desarrollo y compensación profesional, por en un estudio realizado con población entre profesionales de México y Estados Unidos. Las diferencias de género influyen en lo que se espera de cada uno en su labor en el trabajo, sin embargo, las diferencias de ubicación influyen en los factores de apoyo y compensación profesional esperados entre mexicanos y estadounidenses, destacando como factores influyentes las diferencias de género como también las nacionalidades.

Además, en otro estudio por Heras et. al [4] que, aparte de resaltar las diferencias de salarios por géneros, define como determinantes económicos del salario: el nivel de capital humano y un indicador de éxito. Siendo algunas de sus variables *edad*, *antigüedad* y *nivel de estudios*, como parte del capital humano y haciendo comparaciones salariales para el indicador de éxito, ya que este es definido como “la consecución salarial asociada al hecho de recibir más salario que aquellos con los que se comparte la misma dotación de capital humano”.

Para una mayor comprensión de las condiciones en las que está el mercado laboral en el mundo, la OIT (Organización Internacional del Trabajo) [5] ofrece un informe referencial del asunto, en el cuál se indican 4 principales preocupaciones del mundo laboral actualmente: Primero, las proyecciones de crecimiento económico no se verán reflejados en la disminución de pobreza en países de menores ingresos; segundo, la edad vendría a ser un filtro más en la contratación de empleadores, por la fuerza que se dedica en un trabajo; tercero, aún existen deficiencias en la calidad del empleo, que consiste en: un trabajo seguro, saludable, con el acceso a la protección social, para expresar opiniones propias, y que defienda los derechos fundamentales, como la no discriminación; esto también influenciado por el informalismo; y cuarto, se da la segmentación entre trabajadores por su ubicación geográfica, sexo y edad.

Se habla entonces de ciertos factores que determinan el conseguir un trabajo decente y por lo tanto un buen salario. Con el avance de la tecnología se ha desarrollado un campo que ayudaría a analizar y trabajar con datos recolectados para determinar predicciones, la inteligencia artificial, que cuenta con modelos de aprendizaje automático, machine learning, que servirían para la clasificación de grandes cantidades de datos, según Arana [6] los árboles de decisión son el mejor ejemplo de ello no sólo por su interpretabilidad sino también por ser la base de los modelos más potentes utilizados en la actualidad.

En el trabajo realizado por Ponce J. et. al [7] se aplicó la utilización de árboles de decisión para un análisis de la competitividad empresarial, tomando como indicadores recursos financieros, la mercadotecnia, recursos humanos y tecnología. Este se realizó con el fin de determinar el cambio de crecimiento o disminución de esos indicadores con respecto al tiempo, finalmente se concluyó en la mejora de técnicas en la toma de decisiones más convenientes para una empresa.

Por lo tanto se propone realizar un análisis de una base de datos obtenida de Becker [8], quién tomó como variables la edad, género, nivel de educación, país de natalidad y datos adicionales, para determinar si una persona podría obtener un sueldo mayor o menor de 50 000 dólares al año. Se utilizará una de las técnicas de minería de datos para la clasificación de estos, usando árboles de decisión a fin de determinar qué variables si son las determinantes para la estimación de la obtención del sueldo de una persona pudiendo utilizarse esto como ayuda más precisa de estudios de factores determinantes en el mercado laboral.

Fundamentación teórica

Árboles de decisión

Representación gráfica de varias soluciones alternativas que están disponibles para resolver un problema determinado, con el fin de determinar los cursos/rutas de acción más efectivos, [9]. La elección de esta misma se debió a la guía realizada por el docente en un implementación.

KDD

Técnica de minería de datos para mejorar la toma de decisiones con grandes cantidades de datos. Su procedimiento suele constar de los siguientes pasos, según la escuela ESAN BUSINESS [10]:

- Comprensión del área de estudio y fijación de objetivos.
- Implementación de un dataset objetivo.
- Limpieza y procesamiento de información.
- Minería de datos.
- Interpretación y análisis de patrones encontrados.
- Utilización del conocimiento obtenido para la toma de decisiones.

Materiales y métodos o Metodología computacional

Este trabajo utilizó el método predictivo de Árboles de decisión KDD de Machine Learning para el pronóstico salarial en una base de datos de un censo dado en 1996. Este dataset consta originalmente de 15 atributos (columnas) y 32560 instancias (filas).

N°	Variable	Descripción
1	age	continuous
2	workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
3	fnlwgt	continuous
4	education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
5	education-num	continuous

6	marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
7	occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
8	relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9	race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Blac
10	sex	Female, Male
11	capital-gain	continuous
12	capital-loss	continuous
13	hours-per-week	continuous
14	native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
15	salary	<=50K or >50K

Resultados y discusión

Análisis del problema

Mediante un árbol de decisiones se predice si el salario de una persona es <= 50K o > 50K, teniendo en cuenta 11 variables predictivas.

Pre Procesamiento de data

- **Data collecting**

  Del dataset descargado de la página Kaggle, la tabla cuenta con 32560 filas conocidas como registros y 15 columnas, las primeras 14 variables predictoras y la última la variable a predecir. De todas las

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 153-165
https://revistas.ulasalle.edu.pe/innosoft

variables 6 son del tipo int64 y 9 de ellas de tipo object como se muestra en la Figura 1, esta misma la cual será tratada para el estudio en procesamiento de datos antes de entrenar nuestro modelo predictivo.

```
sueldo = pd.read_csv('/content/drive/My Drive/salary.csv', engine='python')
sueldo.head()
sueldo.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   age             32561 non-null  int64 
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64 
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64 
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64 
 11  capital-loss    32561 non-null  int64 
 12  hours-per-week  32561 non-null  int64 
 13  native-country  32561 non-null  object
 14  salary          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

Figura 1. Información descriptiva del dataset descargado de internet

- **Data transformation**

  Debido a que se va a realizar el entrenamiento de datos para un árbol de decisión se vio necesario transformar las variables de tipo object a variables numéricas, para eso haciendo uso de la librería LabelEncoder, Figura 2.

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 153-165
https://revistas.ulasalle.edu.pe/innosoft

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
salary.workclass = le.fit_transform(salary.workclass)
salary.education = le.fit_transform(salary.education)
salary.maritalStatus = le.fit_transform(salary.maritalStatus)
salary.occupation = le.fit_transform(salary.occupation)
salary.relationship = le.fit_transform(salary.relationship)
salary.race = le.fit_transform(salary.race)
salary.sex = le.fit_transform(salary.sex)
salary.nativeCountry = le.fit_transform(salary.nativeCountry)
#salary.salary = le.fit_transform(salary.salary)
salary.head(5)
```

Figura 2. Codigo de transformación de variables object a int64

	age	workclass	education	educationNum	maritalStatus	occupation	relationship	race	sex	hoursPerWeek	nativeCountry	salary
0	39	7	9	13	4	1	1	4	1	40	39	<=50K
1	50	6	9	13	2	4	0	4	1	13	39	<=50K
2	38	4	11	9	0	6	1	4	1	40	39	<=50K
3	53	4	1	7	2	6	0	2	1	40	39	<=50K
4	28	4	9	13	2	10	5	2	0	40	5	<=50K

Figura 3. Resultado del proceso de data transformation

- **Data cleaning**

  Eliminación de columnas

  Para el tratado de la data set se realizó la eliminación de las columnas fnlwgt, capital-gain, y capital-loss razones justificadas por información innecesaria para el caso se estudió, gran cantidad de valores desconocidos, valores desconocidos.

```
[71] salary = sueldo.drop(['fnlwgt','capital-gain', 'capital-loss'],1)
     salary.head()
```

Figura 4. Código de eliminación de columnas del dataset

- **Data reduction**

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 153-165
https://revistas.ulasalle.edu.pe/innosoft

Reconocimiento y eliminación de ruido

Se realizó el reconocimiento de valores ruidosos en el dataset en las columnas de cómo se puede ver en las Figura 5 y Figura 6.

```
[27] salary.age.hist()
     plt.show()          # que sea hasta 75
```

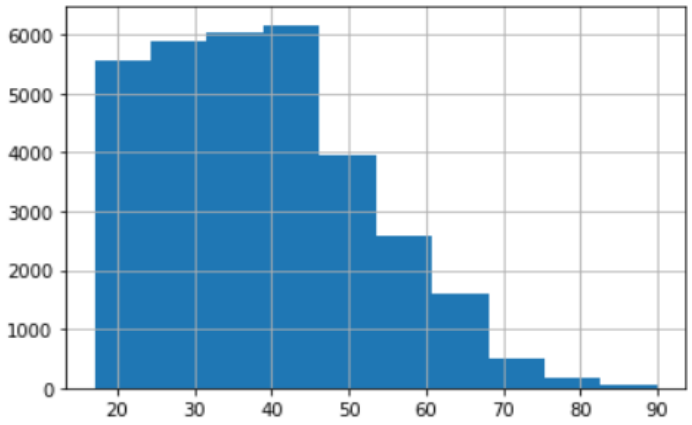

Figura 5. Reconocimiento de ruido mediante un histograma de la columna age

```
[36] salary.nativeCountry.hist()
     plt.show()                          # mayor o igual a 39
```

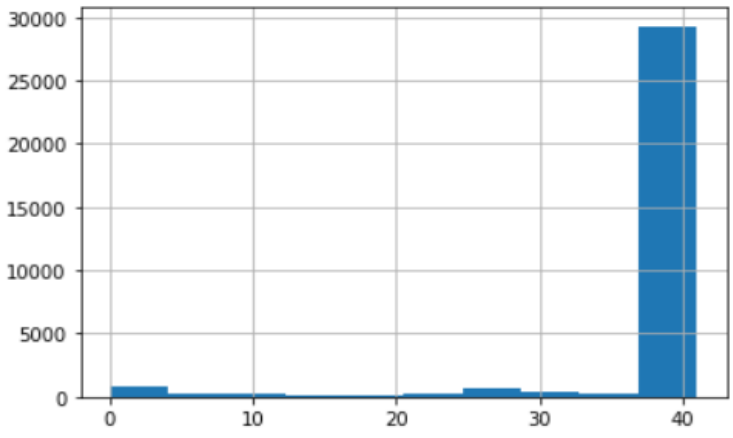

Figura 6. Reconocimiento de ruido mediante un histograma de la columna nativeCountry

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 153-165
https://revistas.ulasalle.edu.pe/innosoft

Eliminación de ruido

Los valores ruidos reconocidos en el dataset fueron eliminados como se muestra en la Figura 7 estableciendo valores limitados a las variables.

```
[74] filtered_salary = salary[(salary['age'] <= 75) & (salary['workclass'] >= 4) & (salary['education'] >= 6)& (salary['educationNum'] >= 9)&
                              (salary['maritalStatus'] <= 4) & (salary['race']>=2) & (salary['hoursPerWeek']<=60) & (salary['nativeCountry']>=39)
                              ]
```

Figura 7. Código de la eliminación de datos que generan ruido

Preparación de Datos

En esta sección se definieron las variables predictoras las cuales son age, workclass, education, educationNum, maritalStatus, occupation, relationship, race, sex, hoursPerWeek y la definicon de la variable a predecir que salary.

```
#Variables Predictoras
x = filtered_salary .iloc[:,0:10] #que considere todas las filas, las variables predictoras

#Variable a Predecir
y = filtered_salary .iloc[:,11] #Se encuentra en la ultima columna
```

Figura 8. Definición de variables

Se realizó también la división de datos para testing & training, asignándo un 80% de los datos para el entrenamiento del modelo y un 20% para pruebas.

```
#x_train y y_train para entrenamiento
#x_test y y_test para prueba
x_train, x_test, y_train, y_test = train_test_split(x,y,train_size=0.80)
```

Figura 9. Asignación de datos para testing y training

Creación del modelo

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 153-165
https://revistas.ulasalle.edu.pe/innosoft

Haciendo uso de la libreria DecisionTreeClassiifier se hace la construcción del árbol indicando que la evaluación será la entropía, además se aclara que se realizará un árbol con una profundidad de 4

```
[80] #llamamos al constructor del arbol de decision
     #arbol = DecisionTreeClassifier() el arbol completo
     arbol = DecisionTreeClassifier(criterion="entropy",max_depth=4)

     #Entrenamos el modelo
     arbol.fit(x_train, y_train)

     #realizo una prediccion
     y_pred = arbol.predict(x_test)
```

Figura 10. Código de creación del modelo de arbol de desicion

Validación del entrenamiento

En la matriz de confusión se muestra que 2688 datos el algoritmo los clasificó como salario <=50K y si eran <=50K, 430 datos fueron clasificados por el algoritmo como >50K y si eran >50K, 672 datos el algoritmo los clasificó como <=50K y no eran <=50K, 107 datos fueron clasificados por el algoritmo como >50K y no eran >50K.

```
[81] matriz = confusion_matrix(y_test, y_pred)
     print('Matriz de confucion')
     print(matriz)

     Matriz de confucion
     [[2688  107]
      [ 672  435]]
```

Figura 11. Matriz de confusión

El modelo trabaja con una exactitud de 0.80 y tiene un error de 0.20

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 153-165
https://revistas.ulasalle.edu.pe/innosoft

```
[83] accuracy_score = accuracy_score(y_test, y_pred)
     print('Exactitud del modelo')
     print(accuracy_score)

     Exactitud del modelo
     0.800358790363916
```

Figura 12. Código de medición de exactitud del modelo

Resultado del árbol de decisión

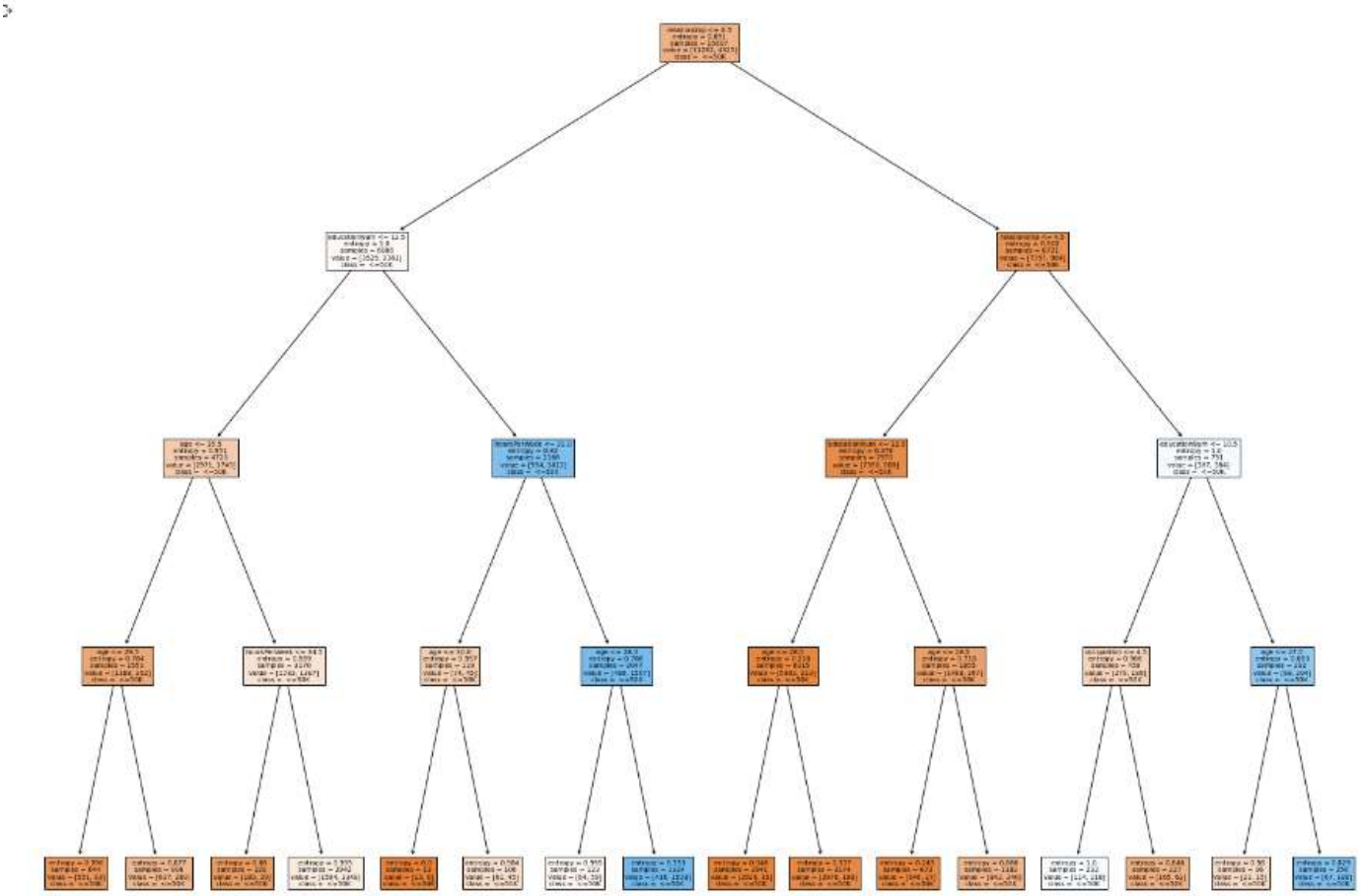

Figura 13. Resultado de árbol de decisiones con un profundidad de 4

Conclusiones

Con la base de datos inicial se tenían en general 14 variables que definirían si una persona podría obtener un sueldo mayor a 50 000 dólares al año, que vendría a ser lo mínimo necesario para la satisfacción de las necesidades básicas en EEUU. Quedando como últimas variables predictoras, después de su procesamiento, en el árbol de decisión: "estado marital", "edad", "nivel de educación", "ocupación" y "horas trabajadas por semana".

Como parte de los trabajos futuros que se podrían implementar a raíz de este, pueden ser:

- Comparativas de distintos métodos predictivos aplicados a esta base de datos, aunque en este caso sólo se vió el algoritmo KDD.
- Para predicciones más exactas es necesaria una actualización de una base de datos.
- Consideración de las variables que sí determinan la obtención de un salario mayor a 50 000 dólares anuales por parte de quienes estén en búsqueda de empleo.

Contribución de Autoría

Corina Hancco Vargas: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **David Flores Silva:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Jessica Hancco Velásquez:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Alejandra Fernandez Ninahuaman:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición.

Referencias

[1] "El panorama laboral después de la pandemia es peor de lo que se preveía | Noticias ONU." https://news.un.org/es/story/2021/10/1499052.

[2] "El gran problema del empleo en el mundo: las malas condiciones de trabajo | Agora: Inteligencia Colectiva para la Sostenibilidad," Feb. 19, 2019. https://www.agorarsc.org/el-gran-problema-del-empleo-en-el-mundo-las-malas-condiciones-de-trabajo/.

[3] S. Madero, “Factores relevantes del desarrollo profesional y de compensaciones en la carrera laboral del trabajador,” p. 22, 2010.

[4] R. L. Heras, A. Maroto Sánchez, Á. Martín-Román, and A. Moral De Blas, “Éxito salarial: indicadores por género en la distribución salarial.” [Online]. Available: www.iaes.es

[5] Organización Internacional del Trabajo, “RESUMEN EJECUTIVO,” 2020.

[6] C. Arana, “Modelos de Aprendizaje Automático Mediante Árboles,” Buenos Aires, Feb. 2021. [Online]. Available: http://hdl.handle.net/10419/238403

[7] J. Ponce, E. Vicente, R. Rodríguez, and S. Muñoz, “Análisis de la competitividad empresarial aplicando árboles de decisión,” pp. 66–80, Aug. 2020. [Online]. Available: https://orcid.org/0000-0002-2965-7263

[8] B. Becker and R. Kohavi, “UCI Machine Learning Repository: Census Income Data Set.” https://archive.ics.uci.edu/ml/datasets/Census+Income.

[9] "Árbol de decisión en Machine Learning (Parte 1) - sitiobigdata.com", sitiobigdata.com, 2019. [Online]. Available: https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1/#.

[10] "Minería de datos: ¿en qué consiste el knowledge discovery in databases?", ESAN Graduate School of Business, 2018. [Online]. Available: https://www.esan.edu.pe/conexion-esan/mineria-de-datos-en-que-consiste-el-knowledge-discovery-in-databases#:~:text=El%20KDD%20es%20un%20proceso,recursos%20%C3%BAtiles%20para%20una%20compa%C3%B1%C3%ADa.