TRABAJO PRACTICO 4-\_BIG DATA\_GRUPO 16

**Grupo** 16: Compuesto por: Maximiliano Petrini, Augusto Fasciani, Lautaro Solís.

**Motivo del trabajo de investigación:** El siguiente trabajo se realizó gracias a las encuestas realizadas, las EPH (Encuesta Permanente de Hogares), por el INDEC (Instituto Nacional de Estadísticas y Censos) en los años 2004 y 2024

Gracias a lo elaborado en los trabajos anteriores, no fue de necesidad elaborar nuevas variables, o continuar limpiando la base de datos ya que previamente habían sido limpiado en los anteriores de trabajos.

Las variables con las que tuvimos que trabajar fueron: Edad, Edad2, Educación, Salario Semanal, Mujer, Ubicación Laboral y Categoría Laboral.

***Edad:***

Son las edades de los encuestados

***Edad2:***

Serían las edades de los encuestados elevado al cuadrado.

***Educación (Edcu):***

Esta variable transforma el grado o progreso de nivel de educación del encuestado en cantidad de años educativos***.***

***Salario Semanal:***

Es lo que gana la persona por semana, para ello utilizamos la data de salario mensual y la dividimos según las semanas trabajadas.

***Mujer:***

Esta variable tomaba el valor de 1 si el sexo del encuestado es de mujer o 2 si el sexo encuestado es de un varón.

***Ubicación Laboral:***

Nos muestra en donde queda el trabajo de cada persona, si en su misma localidad, en otra localidad o en otra provincia.

***Categoría de relación laboral***

Nos indica que tipo de relación laboral tiene el encuestado, siendo así: Patrón, Cuenta propista o Empleado

**Desarrollo del trabajo:**

El objetivo de este trabajo práctico es intentar predecir si una persona está desocupada o no utilizando distintas variables de características individuales y los distintos clasificadores vistosa lo largo de nuestra enseñanza

**Parte A.1**

Por ende, empezamos utilizando nuestra base de datos de las personas que habían respondido la encuesta exitosamente (Dicha base de datos previamente calculada en el TP2). Luego la dividimos en distintas bases de datos: una base de datos para el entrenamiento de nuestros modelos (. La base de entrenamiento comprende el 70% de los datos) y otra para el testeo. A su vez separamos la base de datos según el año, 2004 o 2024.

Como variables a utilizar para la categorización y formación de la base de datos utilizamos a los desocupados como nuestra variable dependiente en la base de entrenamiento.

A continuación, presentaremos una tabla la cual diferencia las medidas entre la base de entrenamiento de ambos años (2004 y 2024).

|  |  |  |
| --- | --- | --- |
|  | 2004 | 2024 |
| Variables | 102 | 106 |
| Mayor diferencia | 0.0138 | 0.0078 |
| Promedio de diferencias | 0.019 | 0.0018 |
| CH04\_Varón | 0.0138 | 0.0078 |
| CH08\_No paga ni le desc | 0.0115 | 0.0063 |
| CH08\_Obra Social | -0.0995 | -0.0061 |

Las ultimas 3 filas muestran el Top 3 de mayores diferencias encontradas.

**Parte B.2**

Para esta parte del trabajo utilizamos las bases de entrenamiento creadas previamente en la parte A y estimamos con la utilización de algunos modelos a Salario Semanal como variable dependiente y el resto de las variables que se nombraran a continuación como variables dependientes (predictores).

* + 1. salario\_semanal en edad
    2. salario\_semanal en edad y edad2
    3. salario\_semanal en edad, edad2 y educ
    4. salario\_semanal en edad, edad2, educ y mujer
    5. salario\_semanal en edad, edad2, educ, mujer y ubi\_trab
    6. salario\_semanal en edad, edad2, educ, mujer y cat\_trab

A continuación, se presentará una tabla la cual índica los coeficientes y desvió estándares (se encontrará en paréntesis) obtenido de modelar la siguiente función con sus variables respectivas utilizando la base de entrenamiento 2004

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 | Modelo 6 |
| Edad | 813.463  (161.598) | 1.701  (664.24) | 1.445  (680.29) | 1.423  (659.91) | 1.423  (660.93) | 5734.108  (698.07) |
| Edad2 | - | -185.047  (7.39) | -153.814  (7.63) | -150.274  (7.40) | -150.322  (7.41) | -56.855  (7.79) |
| Educ | - | - | 9527.331  (742.28) | 1.002  (720.58) | 1.001  (723.99) | 8545.553  (669.39) |
| Mujer | - | - | - | 8.968  (5808.78) | 8.968  (5809.53) | 5.438  (5556.147) |
| Ubi\_tra | - | - | - | - | 1662.305  (8866.74) | - |
| Cat\_ocu | - | - | - | - | - | 1.078  (4293.72) |
| No. Observations: | 3758 | 3758 | 3758 | 3758 | 3758 | 3758 |
| R2 | 0.007 | 0.149 | 0.185 | 0.233 | 0.233 | 0.344 |

A continuación, se presentará una tabla la cual índica los coeficientes y desvió estándares (se encontrará en paréntesis) obtenido de modelar la siguiente función con sus variables respectivas utilizando la base de entrenamiento 2024

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 | Modelo 6 |
| Edad\_24 | 8.331  (4.46 ) | 428.135  (18.47) | 372.332  (19.26) | 368.282  (18.90) | 370.557  (18.89) | 201.429  (20.89) |
| Edad2\_24 | - | -4.683  (0.20) | -4.048  (0.21) | -3.967  (0.20) | -3.982  (0.20) | -2.143  (0.22) |
| Educ\_24 | - | - | 195.415  (21.35) | 215.316  (21.01) | 210.740  (21.03) | 179.073  (20.42) |
| Mujer\_24 | - | - | - | 2036.34  (167.32) | 2028.545  (167.12) | 1507.132  (164.85) |
| Ubi\_tra\_24 | - | - | - | - | 1226.541  (372.53) | - |
| Cat\_ocu\_24 | - | - | - | - | - | 2148.544  (130.61) |
| No. Observations: | 3799 | 3799 | 3799 | 3799 | 3799 | 3799 |
| R2 | 0.001 | 0.126 | 0.145 | 0.177 | 0.179 | 0.232 |

**Parte B.3**

A continuación, se presentará una tabla la cual índica el MSE (Mean square error) o en español el error cuadrático medio (ECM) hecho en la base de testeo de 2004 y hecha con la base de 2024

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 | Modelo 6 |
| MSE TEST | 48484095317.438 | 42741656969.674 | 40760593924.494 | 38488271273.379 | 38493254015.937 | 34355591418.8811 |
| MSE  TESTE  2024 | 30724665.996 | 26794174.935 | 26292470.224 | 25770595.000 | 25823131.437 | 24418517.270 |

**Parte B.4**

En esta parte presentaremos el modelo de mejor performance que ilustre la predicción de salarios\_semanal. El modelo mejor performance fue el de 2024 donde El error cuadrático medio paso de 31 a 30. Mientras que el de 2004 pasó de 40 a 48.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

En ambas ocasiones podemos observar que en el trade off de Sesgo-varianza sale ganando el Sesgo ya que hay una varianza baja y u Sesgo Alto. A su vez es correcto que tenga su máximo entre los 40 y 50 años la función ya que ahí mayormente se determina la especialidad laboral o los puestos más avanzado en la carrera laboral de cada uno.

también está correcto que la función decrezca a los largo de dicha, ya que luego la gente empieza a obtener las jubilaciones las cuales son menores que los salarios laborales semnales.

A continuación, también le mostramos el de 2004.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

**Parte 5**

En esta parte implementamos los métodos de regresión logística (logit) y de vecinos cercanos(KNN) utilizando la base de entrenamiento. Luego utilizando la base de testeo, creamos la matriz de confusión para p>0.5, la curva ROC, los valores de AUC y de Accuaracy de testeo de cada uno.

Para este caso tomaremos que los valores de la matriz de confusión:

predicción

real tn fp

fn tp

Implican que:

* 1 = desocupado
* 0 = ocupado

Como resultado tuvimos:

**Año 2004**

| **Modelo Logit** | **4 falsos positivos y 561 falsos negativos -> Acerto el 93.5% de los casos. AUC = 0.88** |
| --- | --- |
| Modelo KNN | 149 falsos positivos y 645 falsos negativos -> Acerto el 91% de los casos. AUC = 0.72 |

**Año 2024**

| **Modelo Logit** | **2 falsos positivos y 292 falsos negativos -> Acerto el 96.7% de los casos. AUC = 0.88** |
| --- | --- |
| Modelo KNN | 38 falsos positivos y 304 falsos negativos -> Acerto el 96% de los casos. AUC = 0.68 |

Deducimos entonces que para todas las métricas que se pidieron, el modelo Logit fue superior en su aprendizaje respecto a la base de entrenamiento. Su accuracy, AUC y curva ROC fueron ampliamente superior a las del modelo KNN con K=5 en ambos periodos.

Podemos decir, entonces:

* Gráfico, Gráfico de líneas

  El contenido generado por IA puede ser incorrecto.**Gráfico, Gráfico de líneas

  El contenido generado por IA puede ser incorrecto.**Se entreno con una base de entrenamiento de variables independientes a los modelos Logit y KNN para predecir la tasa de desocupación respecto del año 2004 y el 2024. El modelo más preciso a la hora de predecir fue el Logit, prediciendo aproximadamente el 93.5% de los casos para el 2004 y el 96.7% de los casos para el 2024. Es por este motivo que se utilizará para el siguiente inciso.

**Año: 2004**

**Año: 2024**

**Modelo Logit**

**Matriz de Confusión:**

[[8546 2]

[ 292 25]]

AUC: 0.8800

Accuracy: 0.9668

**Modelo KNN**

**Matriz de Confusión:**

[[8510 38]

[ 304 13]]

AUC: 0.6849

Accuracy: 0.9614

**Modelo Logit**

**Matriz de Confusión:**

[[8042 4]

[ 561 164]]

AUC: 0.8837

Accuracy: 0.9356

**Modelo KNN**

**Matriz de Confusión:**

[[7897 149]

[ 645 80]]

AUC: 0.7255

Accuracy: 0.9095

**Parte 6**

En este punto debimos utilizar con el método que seleccionamos que personas son desocupadas dentro de la base de datos no respondieron. Utilizando el modelo Logit, se estima que alrededor del 21.52% de la población dentro de la región "no contestaron" podría estar desocupada. Aproximadamente, la proporción seria 43 de cada 200 personas.