Práctica 2 (35% nota final)

Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com).
Algunos ejemplos de dataset con los que podéis trabajar son:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y **justificar**) son las siguientes:

1. Descripción del dataset.

   El dataset fue obtenido a partir del enlace de Kaggle y está constituido por:

   - PassengerId. Identificador del pasajero registrado en la base de datos
   - Survived. Indica si sobrevivió, siendo 1 positivo y 0 negativo
   - Pclass. Clase del tique. Tomando valores de 1, 2 y 3 (1ª, 2ª y 3ª)
   - Name. Nombre del pasajero
   - Sex. Sexo del pasajero.
   - Age. Edad en años
   - Sib. Nº de hermanos o cónyuges embarcados
   - Sp Parch. Nº de hijos o padres embarcados
   - Ticket. Nº de tique
   - Fare. Precio del tique
   - Cabin. Tipo de camarote
   - Embarked. Puerto embarcado. C = Cherbourg, Q = Queenstown, S = Southampton

   ¿Por qué es importante y qué pregunta/problema pretende responder?

   Para esta práctica se ha utilizado el dataset Titanic: Machine Learning from Disater debido a que me parece interesante al estar relacionado con mis estudios anteriores y por su influencia en el mundo marítimo. Debido a él se reúne la OMI (Organización Mundial Internacional) y crea el primer Convenio Internacional para la Seguridad de la Vida Humana en el Mar (SOLAS), el cual fue revisado en varias ocasiones y es de aplicación actualmente.

   El dataset es importante debido porque nos va a proporcionar información acerca del precio del billete según la edad, sexo, la clase del billete de la camarote, familiares a bordo

y puerto de embarque. Nos va a dar información como nos influyen unos factores mas que otros o no influyen en el precio. Esto nos puede servir para como ejercicio para ejercer una campaña de publicidad desde la competencia al conocer los precios de los billetes.
También se va a hacer una regresión lineal para ver qué factores influyeron en la supervivencia del pasaje del buque, lo cual nos puede proporcionar si es más seguro viajar en clases altas (camarotes encima de la flotación) o la edad influye (más joven mayor capacidad de resistencia) o familiares a bordo (se intenta salvar a los familiares)

2. Integración y selección de los datos de interés a analizar.

Los datos que no vamos a usar de la base de datos son los siguientes:
- Idpassanger. No es un dato relevante porque es una correlación de números.
- Name. Los nombres no son interesantes porque los datos para uso de evaluación de datos deben estar anonizados para no vulnerar la legislación
-Ticktet. El identificador del tique no proporciona información relevante en nuestro estudio.
- Cabin. El numero de camarote no proporciona información que nos influya en el estudio.
Los restantes datos se van a evaluar de cómo influye en el precio del billete los parámetros siguientes:

- Pclass.
- Sex.
- Age.
- SibSp.
- Parch
- Embarked.

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Se va a evaluar cada variable para comprobar si hay elementos vacíos o ceros. En cada caso se procederá a cambiar los elementos vacíos por un valor y los valores cero se comprobará que son correctos o no. En caso negativo, se procederá a cambiar el valor por el mismo procedimiento que el caso de elemento vacío.
- Survived. Se comprueba que no hay valores vacíos por lo cual no se realiza ninguna transformación. Los valores ceros son correctos al ser la gente que no sobrevive. Los valores están dentro de los establecidos (0 y 1).
- PClass. No hay valores vacíos ni tampoco nulos. Todos están dentro de los establecidos (1,2,3)
- Sex. No hay valores vacíos ni tampoco nulos. Todos están dentro de los establecidos (Hombre y Mujer)
- Age. En este caso hay valores vacíos y se procede a cambiar por la mediana de los datos, pero no hay igual a 0 pero si cercanos a él siendo correctos porque en el buque hubo bebes de meses embarcados.
- SIbSp. En este caso no hay valores vacíos, pero si hay valores con cero siendo razonable porque puede haber pasajeros que no tengan hermanos o cónyuges embarcados.
- Parch. No hay valores vacíos, no obstante, si hay valores con cero siendo razonable porque puede haber pasajeros que no tengan hermanos o cónyuges embarcados.
- Fare. En este caso no hay valores vacíos Se comprueba que si hay valores cero no siendo razonable porque el precio del billete debe ser mayor de cero. Se procede a cambiarlo por la mediana de los valores.

- Embarked. Se comprueba que hay valores vacíos y se cambian por el primer puerto S = Southampton. Se ve que no hay valores 0.

3.2. Identificación y tratamiento de valores extremos.

Se va a comprobar los valores extremos para las variables age y fare debido a que las demás variables son factores o se comprobó que no hay valores extraños en el anterior apartado.
- Age. En el diagrama de cajas se comprueba que hay valores extremos pero sus valores pueden ser acordes a edades de pasajeros.
- Fare. Se comprueban que hay valores extremos siendo llamativos los cercanos a 500 porque desde cerca de 270 no hay otro valor hasta llegar a este. Procedemos a cambiar este valor por la media más 3 veces la desviación Se vuelve comprobar los valores extremos, pero son válidos porque se corresponden a precios elevados pero una cierta continuidad sin puntos singulares como anteriormente.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Se va a analizar el precio del billete según sea mujer o hombre, también por la clase del billete o por el puerto de salida para ello se procede a agrupar los datos en las siguientes variables:

- precio.mujer
- precio.hombre
- precio.primera
- precio.segunda
- precio.tercera
- precio.cherbourg
- precio.queenstrown
- precio.southampton

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Para la comprobación de la normalidad de los datos se va a utilizar el test de Shapiro-Wilk con un valor de confianza del 95%. Se comprueba que las variables siguientes no siguen una distribución normal:

- Age
- SibSp
- Parch,
- Fare

Se va a comprobar que los datos de las variables anteriores siguen una distribución normal:

- precio.mujer
- precio.hombre
- precio.primera
- precio.segunda
- precio.tercera
- precio.cherbourg
- precio.queenstrown
- precio.southampton

y se comprueba que no sigue una distribución normal.

Se va a comprobar la igualdad de las varianzas de precio con Age, para ello se van a exponer las hipótesis:

- H0 (Hipótesis nula): Varianza Precio = Varianza Age
- H1 (Hipótesis alternativa): Varianza Precio ≠ Varianza Age

Para comprobar la validez de la hipótesis nula se va a usar el test de Fligner-Killeng con un valor de confianza del 95%.
No se acepta la hipótesis nula (igualdad de varianzas) porque el p-value (0.00005634) es menor que alpha (1-0.95)

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

- Variables que influyen en el precio y su proporción
  Para ello se va a realizar un análisis de correlación entre las distintas variables para determinar las más influyentes. Se procede a calcular el factor de correlación con el Test de Spearman y su p-valor para comprobar que cumple el nivel de confianza.
  Se obtiene:

	Estimado	P-valor
Age	0.1353788	5.034672e-05
SibSp	0.4427160	4.636445e-44
Parch	0.4071481	6.711774e-37

  Se comprueba que la variable que más influye en el precio es Parch frente a Age y SibSp con p-valor menos que alfa (1-0.95) estando dentro del nivel de confianza establecido.

- ¿El precio del billete fue superior en caso de ser mujer o hombre?
  Se va a realizar un contraste de hipótesis sobre dos muestras para determinar si el precio del billete depende si es mujer o hombre. Se establecen dos variables con los precios si es mujer o hombre.
  Para poder realizar un análisis de hipótesis se supone que las muestras siguen una distribución normal (n>30 en ambas variables) y para ello es establece las hipótesis:
  Así, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa:
  - $H0 : \mu1 - \mu2 = 0$
  - $H1 : \mu1 - \mu2 < 0$

  donde μ1 es la media de la población de la que se extrae la primera muestra y μ2 es la media de la población de la que extrae la segunda. Así, tomaremos $\alpha = 0, 05$.
  Como el valor de p-valor (1) es mayor que el valor de significación fijado (0.05), se da por válida la hipótesis nula. Lo que nos indica que el valor del precio del billete no influye si es mujer o hombre.

- Modelo de regresión lineal.

Se calcula una regresión lineal de la variable fare (precio) como explicada y las variables Survived, Pclass, Sex, Age, SibSp, Parch y Embarked como explicativas. Obteniendo:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  87.89225    5.01541  17.524  < 2e-16 ***
Survived1     0.35172    2.53138   0.139  0.88953
Pclass2     -57.25792    3.06090 -18.706  < 2e-16 ***
Pclass3     -66.27064    2.80628 -23.615  < 2e-16 ***
Sexmale      -5.77190    2.47966  -2.328  0.02015 *
Age          -0.15275    0.08247  -1.852  0.06433 .
SibSp         5.79461    0.97809   5.924 4.49e-09 ***
Parch         9.40180    1.34746   6.977 5.91e-12 ***
EmbarkedQ    -7.27062    4.14613  -1.754  0.07985 .
EmbarkedS    -8.43866    2.60489  -3.240  0.00124 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.53 on 881 degrees of freedom
Multiple R-squared:  0.5408,    Adjusted R-squared:  0.5361
F-statistic: 115.3 on 9 and 881 DF,  p-value: < 2.2e-16
```

Se comprueba que las variables Survived, Pclass, SibSp y Parch tienen un nivel de significación cercano al cero y con un nivel de significación del 1% la variable Embarked.
El coeficiente de determinación ajustado de 0,5361 de este modelo por lo cual se puede decir que no se ajusta a la perfección a una recta.
El p-value del modelo es significativo (2.2e-16) por lo que se puede aceptar que el modelo no es por azar, al menos uno de los coeficientes parciales de regresión es distinto de 0.
Se procede a representar los valores de los errores de la regresión por las diferentes variables y del total.
Se comprueba que los errores están cercanos al eje de abscisas de los valores estimados (Cercanos a cero) y no sigue una distribución conocida. También se comprueba que los valores de los errores frente a las diferentes variables no siguen un modelo muy definido como se puede comprobar en los diferentes gráficos.
Se procede a predecir el valor de precio con los siguientes valores:

- Age = 55
- Survived = 1
- Pclass = 3
- Sex = "male"
- SibSp = 0
- Parch = 0
- Embarked = "C"

Dándonos un precio estimado de 7.800433

- Modelo de regresión logística

  Se calcula una regresión lineal de la variable Survived como explicada y las variables Fare, Pclass, Sex, Age, SibSp, Parch y Embarked como explicativas. Obteniendo:

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.2202774  0.4969709   8.492  < 2e-16 ***
Fare              0.0005802  0.0031858   0.182  0.85549
factor(Pclass)2  -1.0117026  0.3116950  -3.246  0.00117 **
factor(Pclass)3  -2.2638986  0.3209783  -7.053 1.75e-12 ***
factor(Sex)male  -2.7230968  0.2009880 -13.549  < 2e-16 ***
Age              -0.0388469  0.0078615  -4.941 7.76e-07 ***
SibSp            -0.3118112  0.1099087  -2.837  0.00455 **
Parch            -0.0743435  0.1201445  -0.619  0.53606
factor(Embarked)Q -0.0762307  0.3811902  -0.200  0.84150
factor(Embarked)S -0.4644293  0.2383889  -1.948  0.05139 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  785.98  on 881  degrees of freedom
AIC: 805.98
```

Se comprueba que las variables Pclass, y Age tienen un nivel de significación cercano al cero y con un nivel de significación del 1% la variable SibSp.

Se procede a predecir la posibilidad de sobrevivir con los siguientes valores:

- Age = 55
- Fare = 7
- Pclass = 3
- Sex = "male"
- SibSp = 0
- Parch = 0
- Embarked = "C"

Se comprueba que no se va a sobrevivir.

5. Representación de los resultados a partir de tablas y gráficas.

Se va a tabular los valores siguientes y sus datos estadísticos:

- precio.mujer
- precio.hombre
- precio.primera
- precio.segunda
- precio.tercera
- precio.cherbourg
- precio.queenstrown
- precio.southampton

Se procede a realizar una gráfica de edad frente precio y gráficos de frecuencia de pasajeros por:

- Sexo
- Clase
- Puerto de embarque

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Se puede estimar el precio no depende del sexo (según el estudio de contraste de hipótesis realizados) y que las varianzas de edad y precio no son iguales por el otro estudio realizado. Se llega a obtener una regresión lineal con un coeficiente de fiabilidad del 95% y con un ajuste cercano al 0,5; pudiéndose considerar para estimar valores de precio. Las variables que influyen en el precio son:

- Survived. No tengo explicación
- Pclass. La clase influye en el precio del billete
- SibSp y Parch. Se hacen descuentos cuando se compran billetes en conjunto por lo cual si se tienen familiares embarcados se supone que se han comprado en conjunto los tiques.
- Embarked. El puerto de embarque influye en el precio porque cuanto más trayecto aumenta el precio.

También se consigue estimar si un pasajero iba a sobrevivir por las variables evaluadas, siendo las de mayor importancia:

- Pclass. Las clases mejores tienen camarotes y zonas comunes sobre la floración siendo más fácil poder escapar del buque. Además, en aquella época tenían derecho a los botes de rescate solo la primera clase.
- Age. Las personas jóvenes tienen más resistencia la frio como al ejercicio para mantenerse a flote.
- SibSp. Tener familiares a bordo influye porque se intenta buscar y salvar a los familiares desviando esfuerzos para ello.

Los resultados permiten resolver el problema planteado.

Contribuciones	Firma
Investigación previa	María Pérez Ameneiro
Redacción de las respuestas	Integrante 1, Integrante 2, ...
Desarrollo código	Integrante 1, Integrante 2, ...