UT.3
Resumen: Preparación de datos. (I)

IES Abastos

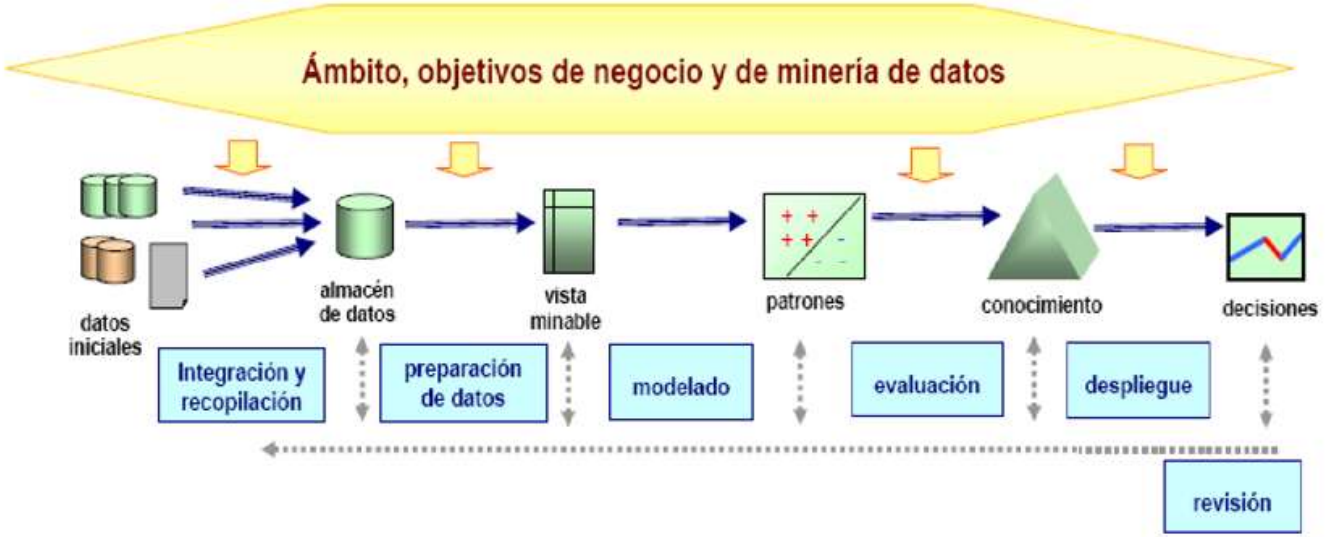
CE Inteligencia Artificial y Big Data

Sistemas de Aprendizaje Automático.

Bloques de la UT2:

- Introducción.
- Práctica 1: Preparación de datos.
- Práctica 2: Red wine.
- Práctica 3: Horse Colic.

Introducción.



Práctica 1. Preparación de datos.

Categóricos: Las variables categóricas contienen un número finito de categorías o grupos distintos. Ejemplo: [1,2,3,4,5]

Como normal general los literales los transformamos en datos categóricos.

```
[ ]  df['Sex'] = df['sex'].map({'Male': 1, 'Female':0})
```

Práctica 1. Preparación de datos.

Categóricos: La moda es un buen valor para reemplazar cada valor nulo.

Se puede utilizar clasificadores para sustituir estos valores, debido a que estos van a predecir clases,

Práctica 1. Preparación de datos.

Numéricos: La media o la mediana son buenos valores para reemplazar cada valor nulo.

La mediana está más orientada para los outliers. Debido a que estos valores anómalos no tienen inferencia en la mediana, 50 % de la posición de los datos, pero si en el cómputo de la media..

Práctica 1. Preparación de datos.

Dummies : Esta técnica sirve para variables numéricas y categóricas. Es una técnica que se suele utilizar con regresiones lineales.

Cuando se usan variables categóricas, según algoritmos de aprendizaje no tiene sentido asignar valores como 1, 2, 3 a valores como "azul", "verde" y "marrón" porque no tiene sentido decir que el verde es dos veces tan colorido como el azul o que el marrón es tres veces más colorido que el azul.

Práctica 1. Preparación de datos.

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

Income	Age	Married	Divorced
$45,000	23	0	0
$48,000	25	0	0
$54,000	24	0	0
$57,000	29	0	0
$65,000	38	1	0
$69,000	36	0	0
$78,000	40	1	0
$83,000	59	0	1
$98,000	56	0	1
$104,000	64	1	0
$107,000	53	1	0

Práctica 2: Red wine.

Discretización: Es el proceso de transformar una variable numérica en una categórica. Un ejemplo muy común es el de convertir una variable como “Edad” en categóricas tales como “20-30”, “60-79” etc.

```
[ ]  # Vamos a considerar que los vinos con valores 1, 2 y 3 son malos (valor 1),
     # con valores 4, 5, y 6 tienen una calidad media (valor 2),
     # con valores 7 y 8 son buenos (valor 3) y aquellos con 9 y 10 son excelentes (valor 4)
     df['calidad'] = df['quality']
     df.loc[df['quality'] <= 3, 'calidad'] = 1
     df.loc[(df['quality'] >= 4) & (df['quality'] <= 6), 'calidad'] = 2
     df.loc[(df['quality'] >= 7) & (df['quality'] <= 8), 'calidad'] = 3
     df.loc[(df['quality'] >= 9), 'calidad'] = 4
```

Práctica 2: Red wine.

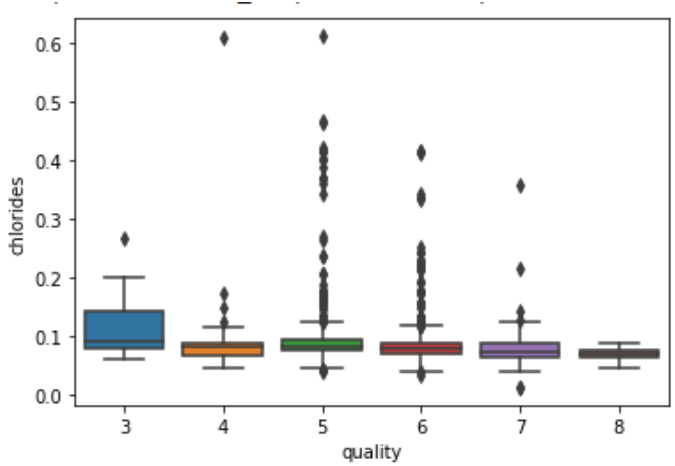
Normalización de los datos: La idea es la siguiente, llevamos todos los valores de la variable entre 0 y 1, y mantenemos las distancias entre los valores.

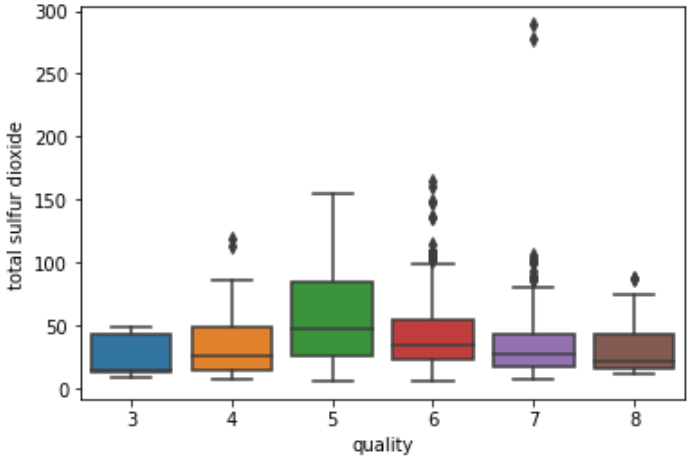
Trabajar con datos de escala variable puede suponer un problema para su análisis, ya que una variable numérica con un rango de valores incluido entre 0 y 10 000 será más pesada en el análisis que una variable cuyos valores estén entre 0 y 1, lo que causaría en consecuencia un problema de sesgo

Práctica 2: Red wine.

Sin embargo, hay que tener cuidado de no considerar la estandarización como un paso obligatorio en el procesamiento de nuestros datos, ¡constituye una pérdida inmediata de información y puede ser perjudicial en ciertos casos! por ejemplo cuando el rango de las variables numéricas no difieran en exceso.

Práctica 2: Red wine.





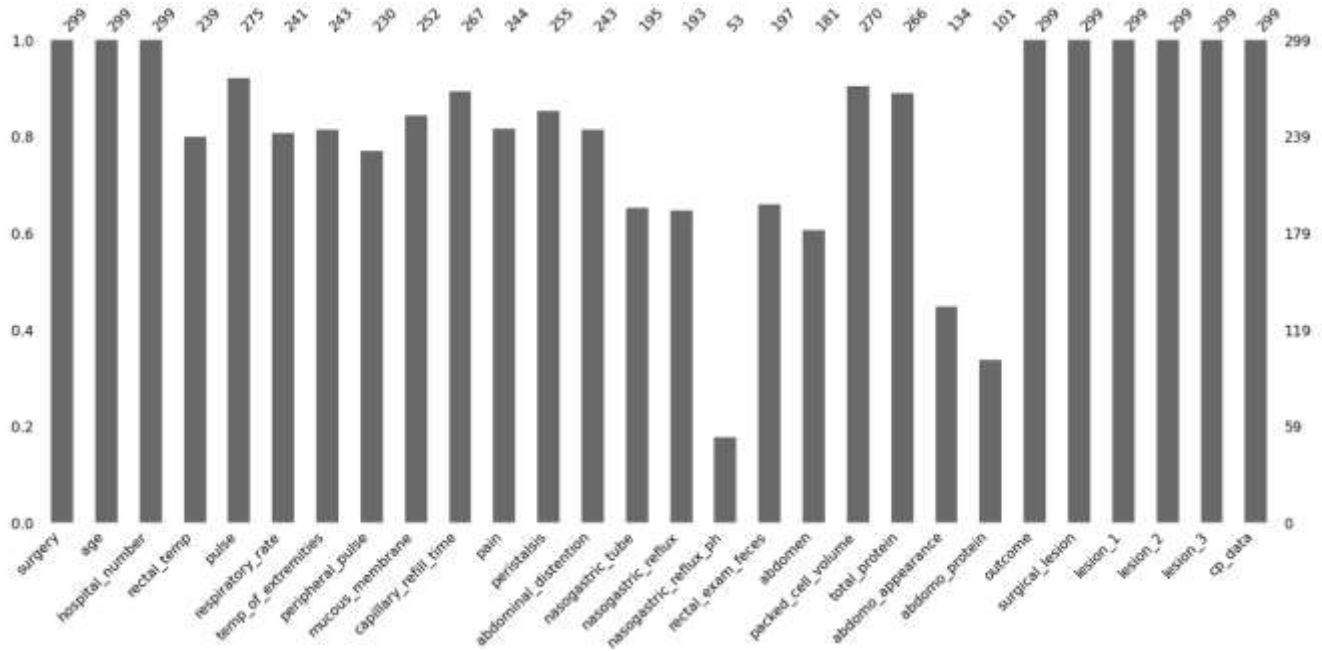
Práctica 2: Red wine.

¿ Después de normalizar cuál será el número más pequeño que podrá ser el mínimo del vector ?

¿ Y el número más grande qué podrá ser el máximo ?

Práctica 3. Horse Colic

Histogramas y mapas de calor para detectar nulos.



Práctica 3. Horse Colic

Reducción de dimensionalidad: La reducción de dimensionalidad se encarga de identificar y eliminar las características que disminuyen el rendimiento del modelo de aprendizaje automático.Conseguir modelos más simples y eficientes.

Un ejemplo puede ser un conjunto de datos de 25 columnas que puede ser representado por únicamente 7 de ellas, capaces de representar el 95% del efecto sobre la variable objetivo. Así pues, se pueden eliminar hasta 18 funciones, simplificando el modelo de aprendizaje automático sin que este pierda eficiencia.

Práctica 3. Horse Colic

Ventajas:

- Menor tiempo de entrenamiento, requiere menos recursos computacionales.
- Solventa el problema del sobreentrenamiento.
- Desaparecen las características redundantes.
- Se facilita la visualización de datos más rápidamente.
- Menos espacio necesario.

Práctica 3. Horse Colic

Desventajas:

- Cuando se produce la reducción de dimensionalidades, se pierde hasta un 15% de la variabilidad en los datos originales.
- Se pueden perder algunos datos debido a la reducción de la dimensionalidad.

Práctica 3. Horse Colic

Desventajas:

- Cuando se produce la reducción de dimensionalidades, se pierde hasta un 15% de la variabilidad en los datos originales.
- Se pueden perder algunos datos debido a la reducción de la dimensionalidad.

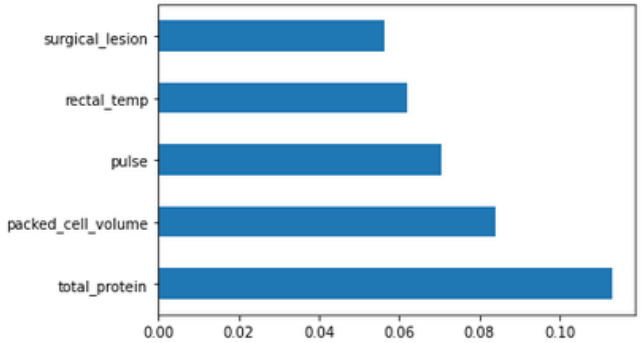
Práctica 3. Horse Colic

Valores faltantes: Es poco probable que las columnas de datos con demasiados valores faltantes contengan mucha información útil. De este modo, se pueden eliminar las columnas de datos con una relación de valores que faltan superior a un umbral determinado. Cuanto más alto sea el umbral, más agresiva será la reducción.

Práctica 3. Horse Colic

Árboles aleatorios: Son muy buenos clasificadores, por lo que tienden a indicarnos qué características son las mejores.

```
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(5).plot(kind='barh')
plt.show()
```



Práctica 3. Horse Colic

Baja varianza: Las columnas de datos con pocos cambios en los datos contienen poca información. De este modo, se pueden eliminar todas las columnas de datos con una desviación inferior a un umbral determinado. Observa que la varianza depende del rango de columnas y, por lo tanto, es necesario normalizar antes de aplicar esta técnica.

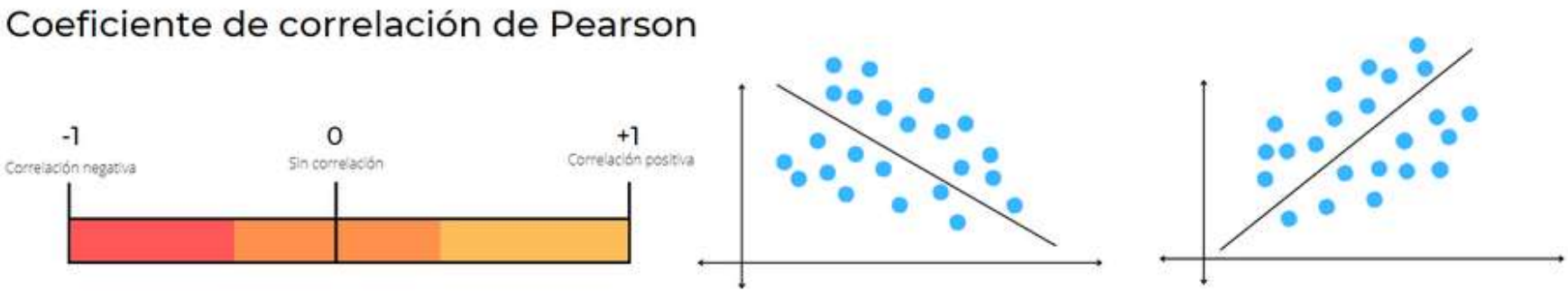
Práctica 3. Horse Colic

Coeficiente de correlación de pearson: Es probable que las columnas de datos con tendencias muy similares también contienen información muy similar.

Se usa como una medida para cuantificar la dependencia lineal entre dos variables numéricas X e Y, su valor varía de -1 a +1.

Se recomienda normalizar previamente.

Práctica 3. Horse Colic



Práctica 3. Horse Colic

Backward elimination:

La eliminación hacia atrás es un método estadístico utilizado para encontrar el modelo más simple que explique los datos.

El valor p más alto se elimina del modelo. Este proceso se repite hasta que todas las variables del modelo tengan un valor p por debajo de un umbral determinado.

Práctica 3. Horse Colic

```
cols = list(X.columns)
pmax = 1
while (len(cols)>0):
    p= []
    X_1 = X[cols]
    X_1 = sm.add_constant(X_1)
    model = sm.OLS(y,X_1).fit()
    p = pd.Series(model.pvalues.values[1:],index = cols)
    pmax = max(p)
    feature_with_p_max = p.idxmax()
    if(pmax>0.05):
        cols.remove(feature_with_p_max)
    else:
        break
selected_features_BE = cols
print(selected_features_BE)
```

```
['pulse', 'total_protein', 'outcome']
```

Práctica 3. Horse Colic

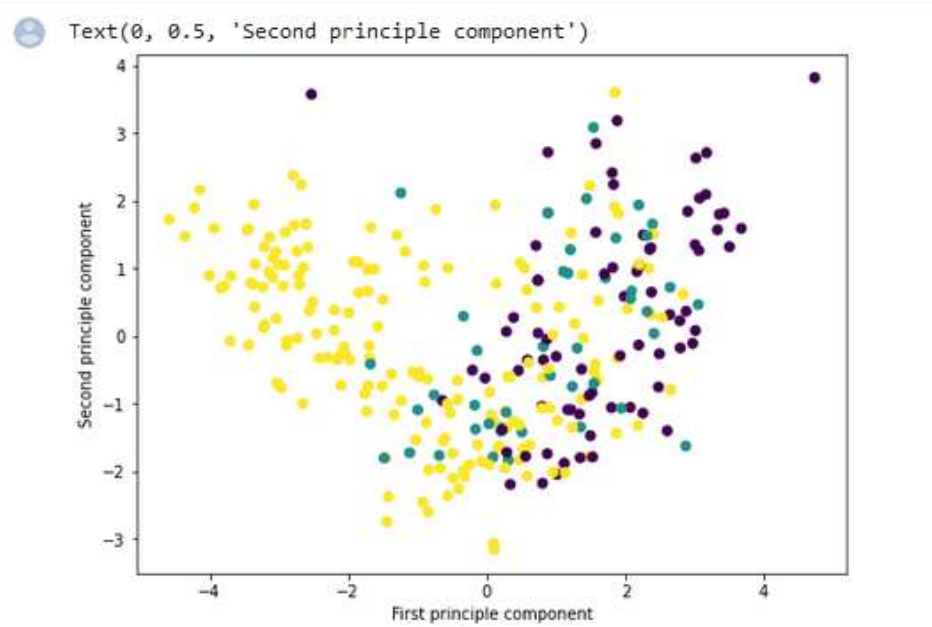
Backward elimination:

La eliminación hacia atrás es un método estadístico utilizado para encontrar el modelo más simple que explique los datos.

El valor p más alto se elimina del modelo. Este proceso se repite hasta que todas las variables del modelo tengan un valor p por debajo de un umbral determinado.

Práctica 3. Horse Colic

PCA: Análisis de componentes principales es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.



Práctica 3. Horse Colic

PCA: Los algoritmos de aprendizaje automático no supervisados no requieren expertos humanos de los que aprender, sino que descubren patrones en los datos de forma autónoma. Algunos ejemplos de problemas resueltos con métodos no supervisados son el clustering (agrupamiento) y la asociación. PCA se considera una técnica de reducción de dimensionalidad no supervisada.