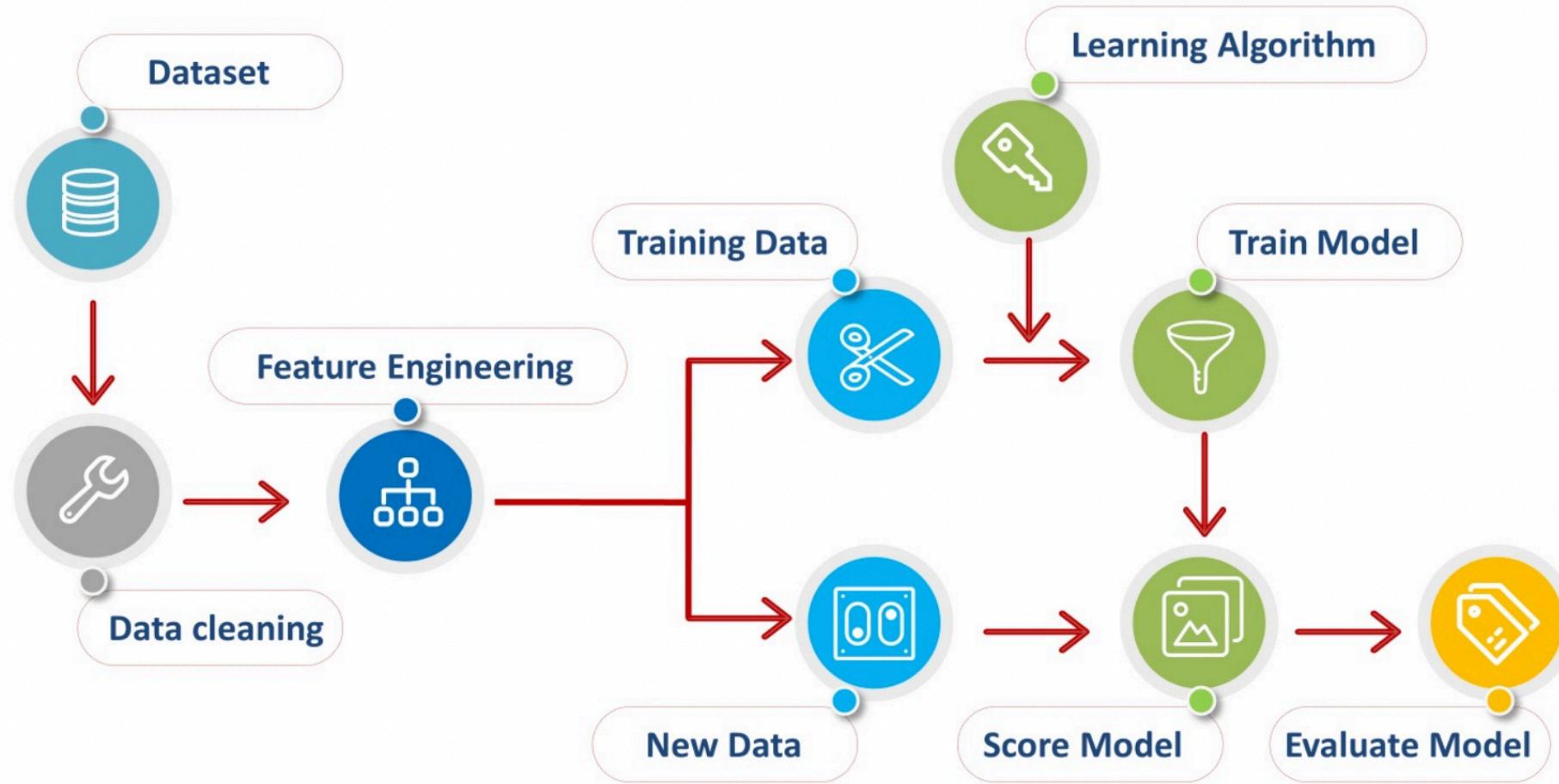


FEATURE ENGINEERING

JAIRO SALAZAR

Feature Engineering Pipeline



CARACTERISTICAS Y VARIABLES OBJETIVO

REGULARMENTE EN UN DATA SET TENDREMOS UNA VARIABLE OBJETIVO (VARIABLE DEPENDIENTE) Y VARIAS CARACTERISTICAS “ASOCIADAS” A DICHA VARIABLE OBJETIVO (VARIABLES INDEPENDIENTES).

LA INGENIERIA DE CARACTERISTICAS (FEATURE ENGINEERING) LO QUE INTENTA ES APROVECHAR AL MAXIMO EL CONOCIMIENTO DE LOS DATOS QUE POSEEMOS PARA GENERAR CARACTERISTICAS NUEVAS O MODIFICAR O TRANSFORMAR DATOS EXISTENTES.

MUCHOS DE LOS ALGORITMOS DE MACHINE LEARNING FUNCIONAN MEJOR O REQUIEREN ESTRUCTURAS ESPECIFICAS DE LOS DATOS.

1. DATOS FALTANTES

- LOS DATOS FALTANTES PUEDEN SER UN VERDADERO DOLOR DE CABEZA PARA CUALQUIER ALGORITMO DE ML, CONSIDERANDO PRINCIPALMENTE QUE TODOS LOS ALGORITMOS DE ML NECESITAN VALORES NUMERICOS Y NO DATOS NULOS PARA PODER FUNCIONAR ADECUADAMENTE.

ACCIONES A TOMAR

- **REMOVER DATOS FALTANTES**
- **COMPLETAR DATOS FALTANTES (DATA IMPUTATION)**

1.1 REMOVER DATOS FALTANTES

- SE REFIERE A REMOVER TODAS LAS OBSERVACIONES QUE POSEAN FALTANTES EN CUALQUIER VARIABLE DEL DATASET.
- DE ESTE MODO SE UTILIZA LA INFORMACIÓN QUE SE CONSIDERA “COMPLETA” DENTRO DEL DATASET.
- APLICABLE A DATOS NUMÉRICOS, CATEGÓRICOS Y MIXTOS. •
- SE RECOMIENDA USAR CUANDO LA CANTIDAD DE DATOS FALTANTES ES MENOR AL 5%

1.1 REMOVER DATOS FALTANTES

VENTAJAS

- Es un enfoque simple,
- No requiere ninguna manipulación interna de los datos,
- Prevalen las propiedades probabilísticas de los datos, no modificamos la distribución de los datos.

DESVENTAJAS

- Puede excluir a una gran cantidad de datos del dataset original.
- Omite observaciones que podrían ser particularmente importantes para la construcción del modelo.
- Podría generarse un dataset sesgado debido a que se pueden omitir observaciones que contengan categorías específicas.
- En producción el modelo podría producir errores ya que si aparece una observación con alguna categoría eliminada, el predictor no sabrá como tratarla.

1.2 IMPUTACION DE DATOS

- LA IMPUTACIÓN DE DATOS SE REFIERE A LA ACCIÓN DE REEMPLAZAR LOS VALORES FALTANTES DE UN CONJUNTO DE DATOS, CON UNA ESTIMACIÓN DEL POSIBLE VALOR REAL.
- LA IDEA PRINCIPAL ES PROVEERLE UN DATASET CON LA MAYOR CANTIDAD DE INFORMACIÓN POSIBLE A UN ALGORITMO DE MACHINE LEARNING.

Datos Numéricas

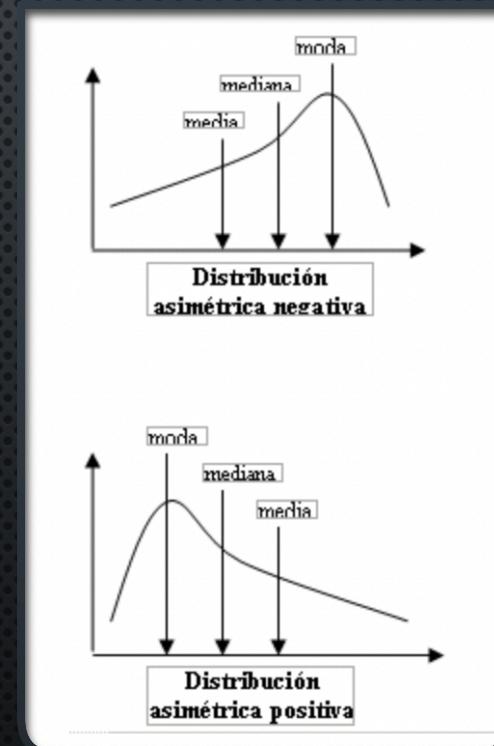
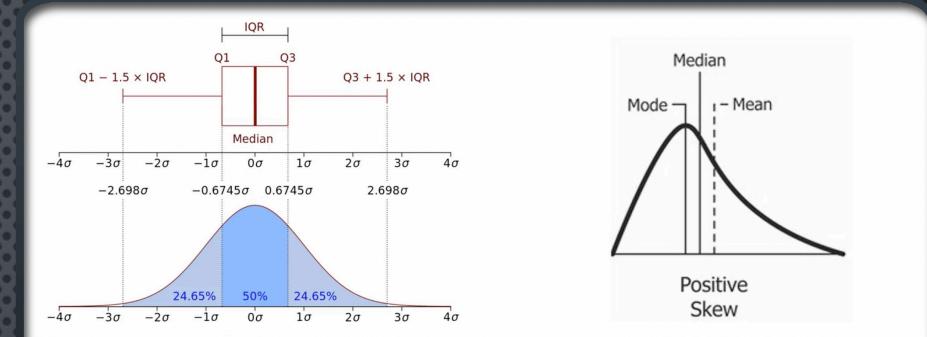
- Imputación de media y mediana.
- Imputación de valores arbitrarios.
- Imputación probabilística.

Variables Categóricas

- Imputación por frecuencia.
- Agregar categoría de faltante.

1.2 IMPUTACION DE DATOS(IMPUTACION MEDIANA O MEDIA).

CONSISTE EN IMPUTAR LA MEDIA O LA MEDIANA EN UNA DISTRIBUCIÓN DE DATOS DONDE EXISTEN FALTANTES, POR SU NATURALEZA NUMÉRICA NO ES POSIBLE APLICAR ESTE TIPO DE IMPUTACIÓN A VARIABLES CATEGÓRICAS.



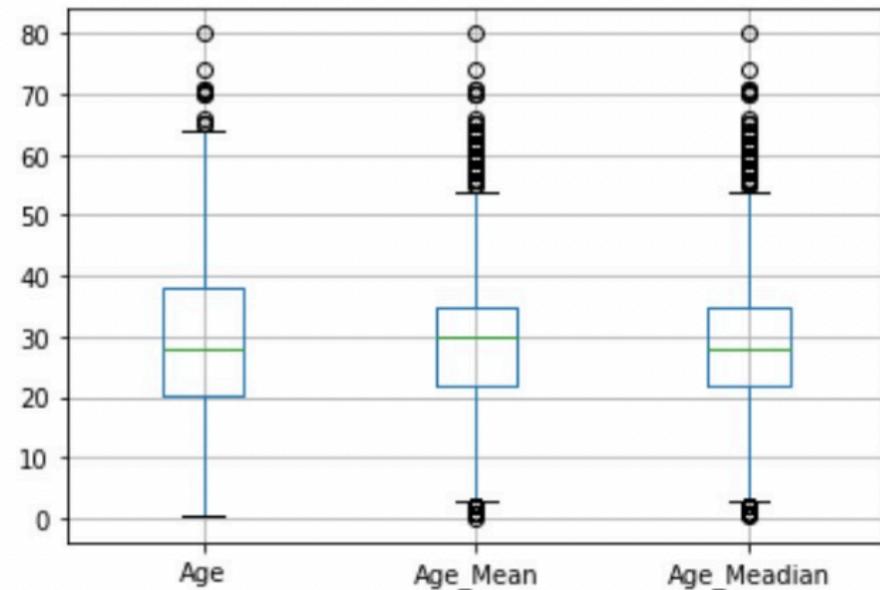
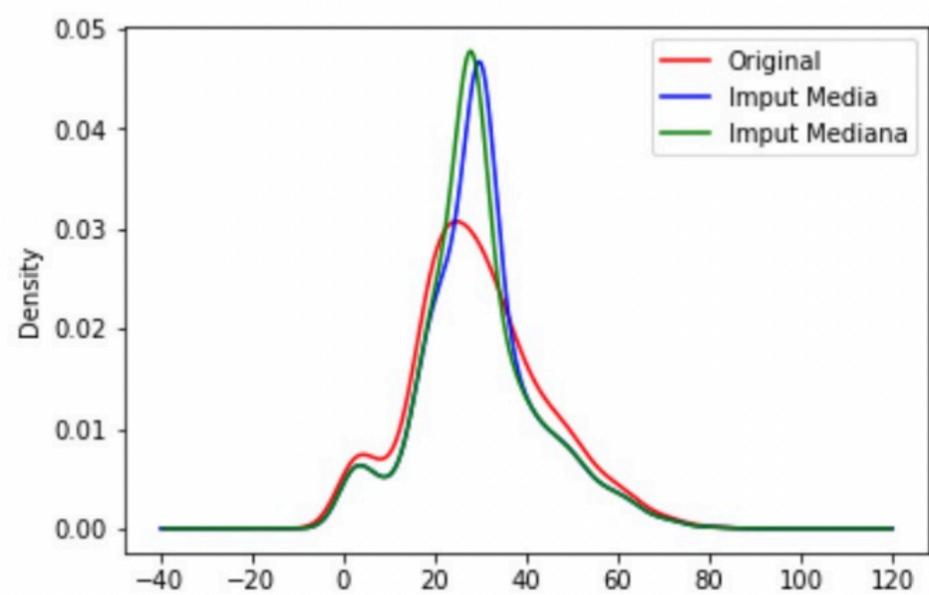
1.2 IMPUTACION DE MEDIA O MEDIANA

VENTAJAS

- ES UN ENFOQUE FÁCIL Y RÁPIDO DE IMPLEMENTAR.
- PERMITE OBTENER DATASETS COMPLETOS DE FORMA EFICIENTE.
- PUEDE INTEGRARSE EN PRODUCCIÓN, SI UN VALOR ES FALTANTE AL MOMENTO DE REALIZAR UNA PREDICCIÓN.

DESVENTAJAS

- DISTORSIONA LA DISTRIBUCIÓN DE LA VARIABLE ORIGINAL.
- DISTORSIONA LA VARIANZA DE LA VARIABLE ORIGINAL.
- DISTORSIONA LA CORRELACIÓN Y COVARIANZA DE LA VARIABLE IMPUTADA RESPECTO A LAS DEMÁS EN EL DATASET.
- ENTRE MÁS NAs, MAYOR SERÁ LA DISTORSIÓN.



2. CODIFICACION DE VARIABLES CATEGORICAS

SE REFIERE AL PROCESO DE PRODUCIR VALORES NUMÉRICOS A PARTIR DE UNA VARIABLE CATEGÓRICA, ESTO CON LA FINALIDAD DE:

- PRODUCIR VARIABLES NUMÉRICAS QUE PUEDEN SER UTILIZADAS POR EL ALGORITMO DE ML.
- PRODUCIR NUEVAS CARACTERÍSTICAS A PARTIR DE LAS CATEGORÍAS DISPONIBLES EN EL DATASET.

Técnicas Tradicionales

- One hot Encoding.
- Frequency Encoding.
- Ordinal Encoding.

Relaciones Monotonicas

- Ordered Label Encoding.
- Mean Encoding.
- Weight of Evidence.

Técnicas Modernas

- Binary Encoding.
- Feature Hashing.

2.1 ONE HOT ENCODING

- CONSISTE EN CODIFICAR LOS VALORES DE UNA VARIABLE CATEGÓRICA CON UN CONJUNTO DE VALORES BOOLEANOS (0 o 1), ESTO NOS PERMITE INDICARLE AL ALGORITMO DE ML SI EL VALOR DE LA CATEGORÍA ESTÁ PRESENTE O NO EN LA OBSERVACIÓN.

Valor
Texas
Florida
California
Texas
Florida
California

	Texas	California	Florida
Texas	1	0	0
Florida	0	0	1
California	0	1	0
Texas	1	0	0
Florida	0	0	1
California	0	1	0

2.2 LABEL ENCODING

Valor
Texas
Florida
California
Texas
Florida
California

Valor
1
2
3
1
2
3

- ESTE ENFOQUE CONSISTE EN CODIFICAR EL VALOR DE LAS CATEGORÍAS UTILIZANDO UN VALOR NUMÉRICO ARBITRARIO QUE VA DESDE 1 A N (O DE 0 A $N - 1$). LA IDEA ES QUE CADA CATEGORÍA TENGA UN VALOR NUMÉRICO DISTINTO.
 - SE PUEDE UTILIZAR PARA CODIFICAR VARIABLES CATEGÓRICAS CON ESCALA NOMINAL, ES DECIR, NO ES NECESARIO QUE EXISTA UN NIVEL DE IMPORTANCIA INTRÍNSECO EN LAS CATEGORÍAS DE LA VARIABLE.

Frecuency Encoding

Valor
Texas
Texas
California
Texas
Florida
California

Valor
1
1
3
1
2
3

Valor
0.5
0.5
0.333
0.5
0.166
0.333

2.3 FREQUENCY ENCODING

- ESTE ENFOQUE CONSISTE EN SUSTITUIR LAS CATEGORÍAS DE UNA VARIABLE EN FUNCIÓN DE LA FRECUENCIA O DENSIDAD DE CADA CATEGORÍA EN LAS VARIABLES EL DATASET.
- ESTE ENFOQUE LA "FUERZA" QUE CADA CATEGORÍA TIENE DENTRO DEL DATASET.
- ES UN ENFOQUE MUY POPULAR EN LAS COMPETENCIAS DE KAGGLE.
- PARA QUE FUNCIONE ADECUADAMENTE, SE ASUME QUE LAS CATEGORÍAS PRESENTES EN UNA VARIABLE TIENE DE ALGUNA FORMA UNA RELACIÓN CON LA VARIABLE A PREDICIR.

3. FEATURE SCALING

- LOS ALGORITMOS QUE ESTÁN BASADOS EN LINEALIDAD SON SENSIBLES A LA ESCALA DE LAS VARIABLES.
- LAS VARIABLES CON RANGOS DE MAGNITUDES MÁS GRANDES SUELEN DOMINAR SOBRE LAS QUE TIENEN RANGOS DE MAGNITUDES MÁS PEQUEÑAS.
- GRADIENT DESCENT CONVERGE MEJOR SI LAS VARIABLES TIENE LA MISMA ESCALA.
- EN EL CASO DE LOS SVM, ES MÁS FÁCIL TRABAJAR CON VARIABLES DE ESCALA SIMILAR.
- EN ALGORITMOS BASADOS EN DISTANCIA (KNN, KMEANS, PCA) LA MAGNITUD DE LAS VARIABLES PUEDE GENERAR DIFERENCIAS MUY GRANDES DIFÍCILES DE MANIPULAR E INTERPRETAR

3. FEATURE SCALING

ALGORITMOS SENSIBLES A LA ESCALA:

- REGRESIONES LINEAL Y LOGÍSTICA,
- REDES NEURALES,
- SVM,
- KNN,
- LDA,
- QDA,
- PCA,
- K-MEANS.

ALGORITMOS INSENSIBLES A LA ESCALA: • TODOS LOS ALGORITMOS BASADOS EN ARBOLES:
ARBOLES DE DECISIÓN, RANDOM FOREST, ADABoost, XGBoost, ETC...

FEATURE SCALING

- SE REFIERE A UN MECANISMO EL CUAL CONSISTE EN NORMALIZAR LOS VALORES DEL CONJUNTO DE VARIABLES DE UN DATASET.
- LA IDEA ES IGUALAR LA ESCALA DE TODAS LAS VARIABLES DESCritAS EN EL DATASET.
- TíPICAMENTE, EL PROCEDIMIENTO DE FS ES EL ULTIMO PASO ANTES DE ENTRENAR EL MODELO DE ML.

Feature Scaling

Standardization (*).

Normalización de Media.

MinMax Scaling(*)

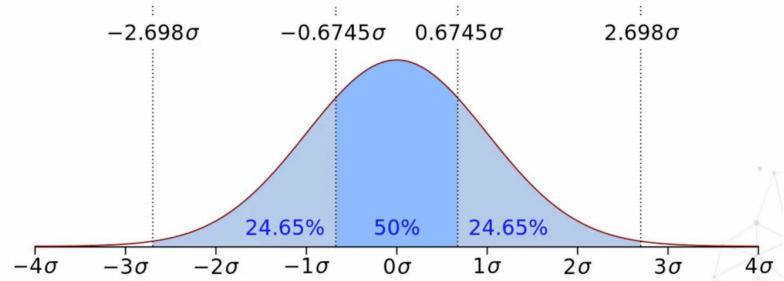
Max Absolute Scaling

Robust Scaling.

STANDARDIZATION

- CONSISTE EN REALIZAR LA ESCALA DE LOS VALORES UTILIZANDO UNA CONVERSIÓN AL VALOR Z DE LA VARIABLE, BASADA EN LA MEDIA Y DESVIACIÓN ESTÁNDAR.

$$Z = \frac{x_i - \bar{\mu}}{\sigma}$$



Standardization

Centra la media en 0.

Escala la varianza a 1.

Conserva la forma de la distribución original.

Conserva los valores máximos y mínimos.

Conserva los outliers.

MEAN NORMALIZATION

- CONSISTE EN UTILIZAR INFORMACIÓN SOBRE LOS VALORES EXTREMOS PARA REALIZAR LA NORMALIZACIÓN, SEGÚN SE MUESTRA EN LA SIGUIENTE ECUACIÓN:

$$X_{scaled} = \frac{x_i - \bar{\mu}}{\max(x) - \min(x)}$$

Mean Normalization

Centra la distribución en 0,

Cambia la varianza de la distribución si la variable tiene mucho sesgo.

Podría modificar la forma de la distribución.

Los valores varían entre -1 y 1.

Conserva los outliers.

MEAN MAX SCALING

CONSISTE EN UTILIZAR INFORMACIÓN SOBRE LOS VALORES EXTREMOS PARA REALIZAR LA ESTANDARIZACIÓN, SEGÚN SE MUESTRA EN LA SIGUIENTE ECUACIÓN:

$$X_{scaled} = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

MinMax Scaling

Devuelve valores entre 0 y 1,

Modifica la media y la varianza.

Podría modificar la forma de la distribución.

Los valores varían entre 0 y 1.

Conserva los outliers.