

Indicators of Heart Disease

IT326 Data Mining

Group #3



Table of contents

01

Introducing The
Problem

02

Data Mining
Task

03

Dataset
Overview

04

Data
Preprocessing

05

Data Mining
Technique

06

Evaluation and
Comparison

07

Findings



Introducing The Problem

The objective is to identify key risk factors for heart disease and develop predictive models to classify individuals as high or low risk. Heart disease, a leading global cause of death, is often linked to factors like high blood pressure, cholesterol, and smoking. Early detection and prevention can save lives, reduce healthcare costs, and lessen the societal burden. By analyzing health data, we aim to uncover patterns that support personalized treatment, optimize resources, and improve healthcare strategies.

Data Mining Task

Objective

- Analyze health data to identify risk factors for heart disease.
- Develop models to:
 - Classify: Individuals into high-risk or low-risk categories.
 - Cluster: Group individuals with similar health profiles.

Classification Task

- Target: Binary variable:
 - 1: High risk, 0: Low risk.
- Attributes: Age, Gender, Blood Pressure, Cholesterol, Smoking, Activity, Diabetes.
- Evaluation: Accuracy, Confusion Matrix, Train-Test Splits (90-10, 80-20, 70-30)

Data Mining Task

Goal of Classification:

Build predictive models to accurately classify individuals into high or low risk based on attributes **such as:** Age - Gender - Blood Pressure - Cholesterol Levels Smoking.

Evaluate the models using performance metrics such as: Accuracy - Confusion matrix Train-test split variations (e.g., 90%-10%, 80%-20%, 70%-30%).

Clustering Task.

Data Mining Task

Clustering Task

- Attributes: Age, Cholesterol, Blood Pressure, Lifestyle Habits (e.g., smoking).
- Evaluation Metrics: Silhouette Coefficient, Elbow Method.

Importance

- Classification: Early detection and personalized healthcare.
- Clustering: Insights into population subgroups for better resource allocation.

Dataset Overview



Dataset Source

kaggle link:
<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download>



Class Label

HadHeartAttack is the target variable, classified as "Yes" for at risk and "No" for not at risk. It helps identify individuals needing early intervention for heart disease.



Objects and Attributes

Dataset contains 445132 Objects and 40 Attributes

Dataset Overview

Our dataset includes 40 Attributes and the Significant Variables:

- AgeCategory: Age group of individuals (e.g., 18–24, 25–34, etc.).
- PhysicalHealthDays: Number of days in the last 30 days with poor physical health.
- MentalHealthDays: Number of days in the last 30 days with poor mental health.
- SleepHours: Average hours of sleep per day.
- HeightInMeters: Height of individuals in meters.
- WeightInKilograms: Weight of individuals in kilograms.
- BMI: Body Mass Index.
- RemovedTeeth: Number of removed teeth.
- HadHeartAttack (Class Attribute): Whether the individual had a heart attack (Yes/No).

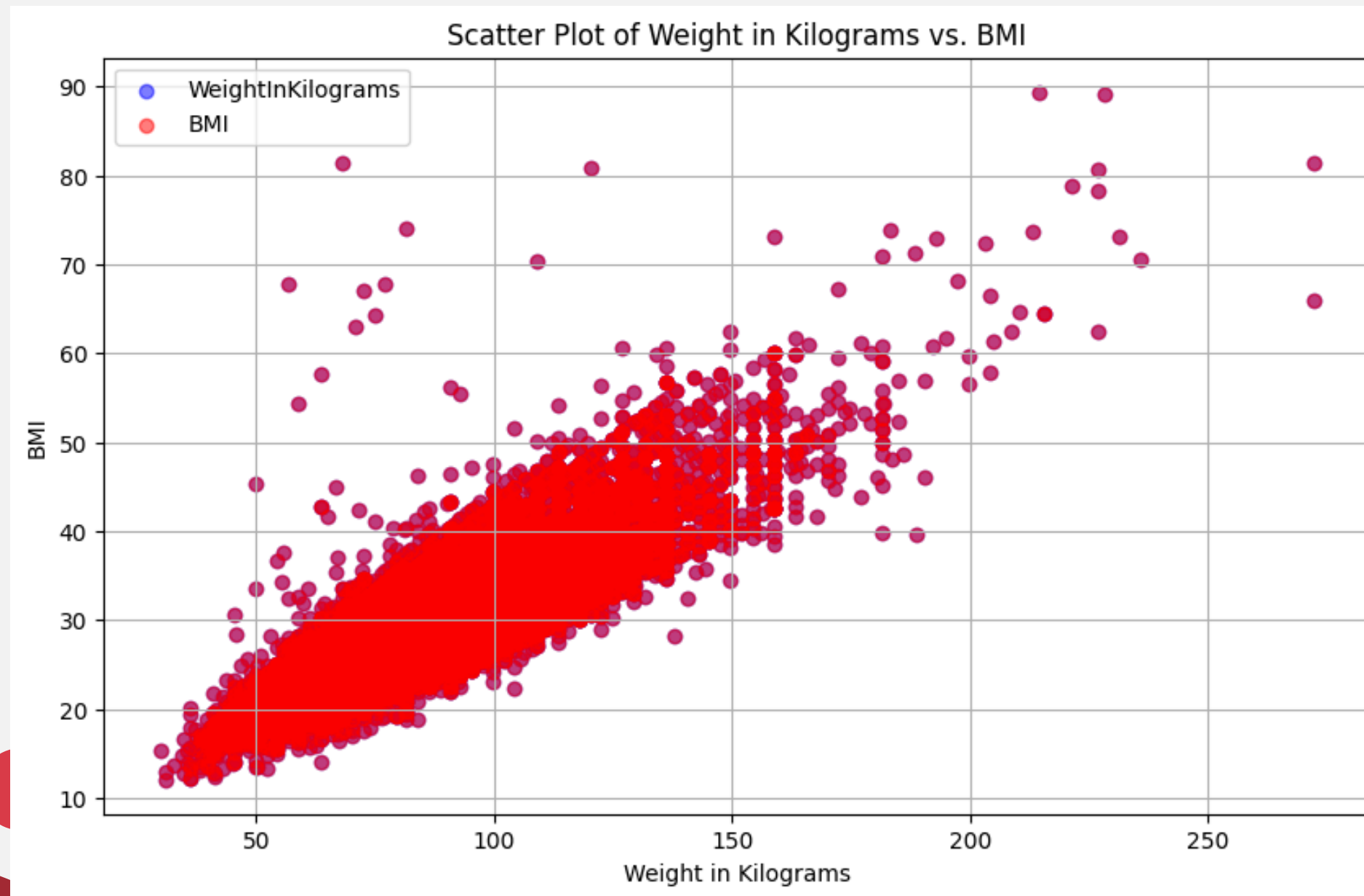
Dataset Overview

<u>Name Of Attributes</u>	<u>Data Type</u>	<u>Possible Values</u>	<u>Missing Values</u>
AgeCategory	Chr (Nominal)	18–24, 25–34, 35–44, ..., 80+	249
Physical Health Days	Num (Numeric)	0–30	652
Mental Health Days	Num (Numeric)	0–30	551
SleepHours	Num (Numeric)	1–24	276
Height In Meters	Num (Numeric)	0.9–2.4	0
Weight In Kilograms	Num (Numeric)	22–292	0
BMI	Num (Numeric)	12–99	0
Removed Teeth	Chr (Nominal)	None, Some, All	651
Had Heart Attack	Chr (Nominal)	Yes, No	0



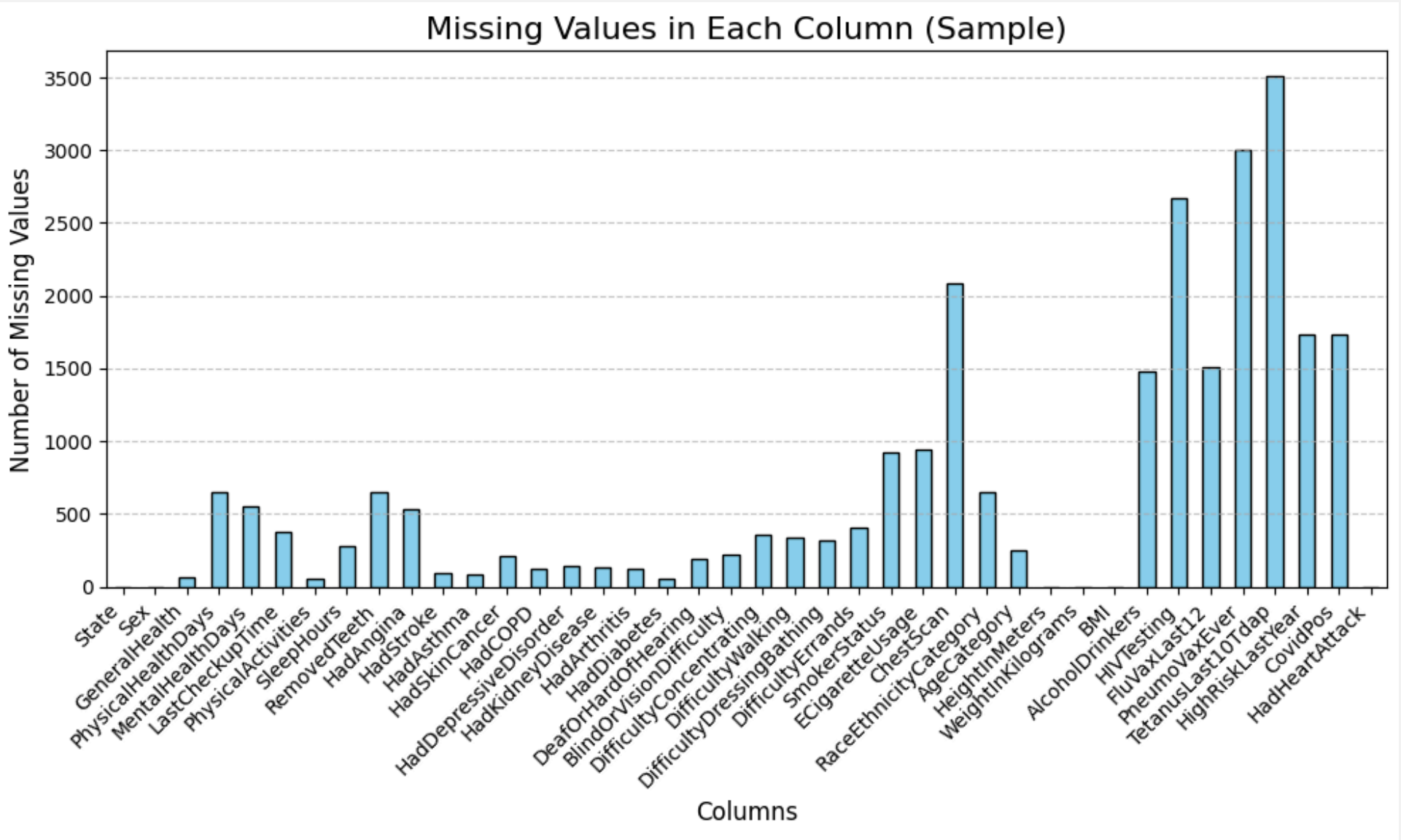
Some Plots Helped
Us Understand
The Dataset

1. Scatter Plot BMI and Weight



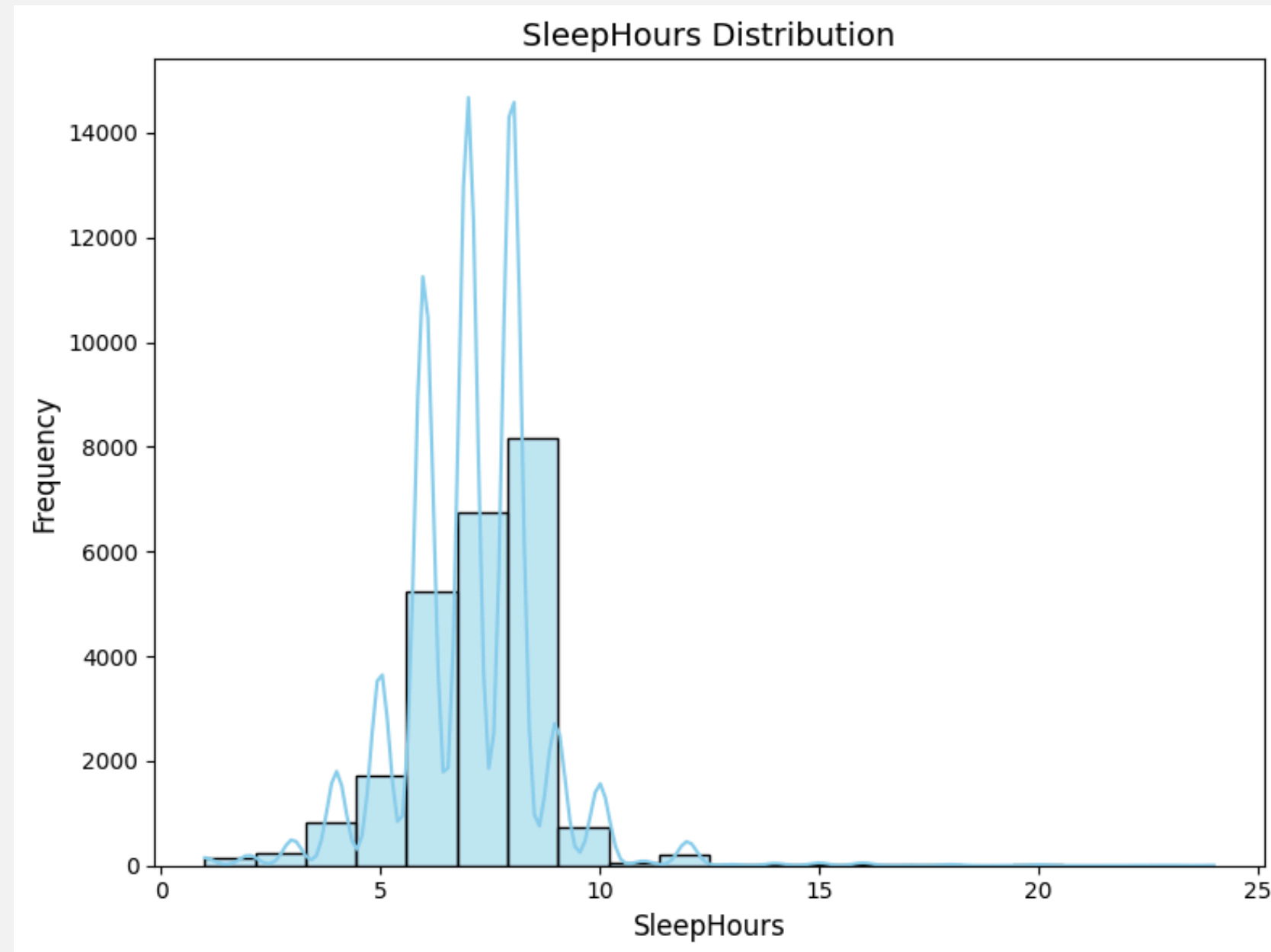
The scatter plot of weight versus BMI shows a strong positive correlation, indicating that as weight increases, BMI also tends to increase proportionally. This high correlation suggests that these two variables provide overlapping information, making one of them redundant for analysis. Therefore, during preprocessing, we can drop one of these variables to reduce multicollinearity and simplify the dataset without losing significant information. This is a key step in preparing the data for efficient and interpretable modeling.

2. Missing Values Barchart



The bar chart highlights the number of missing values across each column in the dataset, showcasing significant variation. Some columns, such as TetanusLast10Tdap, exhibit a high number of missing values, while others like GeneralHealth and PhysicalHealthDays have moderate levels. Columns with few missing values can use simple imputation methods like mean or median, while those with many gaps may need advanced techniques or removal if not critical. Addressing missing values ensures the dataset is clean, reduces bias, and improves the reliability of analysis or modeling.

3. Histogram for SleepHours



The histogram shows the distribution of sleep hours (SleepHours) in the dataset. The observed distribution of SleepHours suggests the need for discretization to group data into logical categories. Without discretization, the continuous nature of the data makes it difficult to draw meaningful comparisons or identify clear trends. Grouping sleep hours into categories such as <5 hours, 5-6 hours, 7-8 hours, and 9+ hours simplifies the analysis, reduces the impact of extreme values, and improves the interpretability of relationships with other attributes, such as health outcomes or demographic factors. This preprocessing step is essential for clearer insights and more reliable modeling.



Data Preprocessing

1. Handling Missing Data

Why:

The dataset contained missing values in several attributes (e.g., TetanusLast10Tdap, PneumoVaxEver). Missing data can introduce bias, reduce the dataset's integrity, and lead to errors in model training. It was essential to handle these missing values to maintain data completeness.

How:

Columns with more than 50% missing values were dropped, as the amount of missing data made them unreliable.

For numerical attributes (e.g., PhysicalHealthDays, MentalHealthDays), missing values were imputed using the median to prevent skewing the data distribution.

For categorical attributes (e.g., TetanusLast10Tdap, PneumoVaxEver), missing values were imputed with the mode, which represents the most frequent category.

Attributes:

all attributes with missing data

1. Handling Missing Data

Raw dataset

I	H	G	F	E	D	C	B	A	
RemovedTeeth	SleepHours	PhysicalActivities	LastCheckupTime	MentalHealthDays	PhysicalHealthDays	GeneralHealth	Sex	State	
	8	No	Within past year (an	0	0	Very good	Female	Alabama	2
	6	No		0	0	Excellent	Female	Alabama	3
	5	Yes	Within past year (an	3	2	Very good	Female	Alabama	4
	7	Yes	Within past year (an	0	0	Excellent	Female	Alabama	5
	9	Yes	Within past year (an	0	2	Fair	Female	Alabama	6
	7	No	Within past year (an	0	1	Poor	Male	Alabama	7
	7	Yes	Within past year (an	0	0	Very good	Female	Alabama	8
	8	No	Within past year (an	0	0	Good	Female	Alabama	9
	6	Yes	Within past year (an	0	0	Good	Female	Alabama	10
	7	Yes	Within past year (an	0	1	Good	Female	Alabama	11
	8	No	Within past year (an	9	8	Fair	Female	Alabama	12
	6	No	Within past year (an	0	0	Good	Female	Alabama	13
	6	No	Within past year (an	0	5	Fair	Male	Alabama	14
	8	Yes	Within past year (an	0	0	Very good	Male	Alabama	15
	8	Yes	Within past year (an	5	30	Good	Female	Alabama	16
	8	Yes	Within past year (an	0	0	Excellent	Female	Alabama	17
	6	Yes	Within past year (an	0	0	Excellent	Female	Alabama	18
	6	Yes	Within past year (an	15	0	Fair	Female	Alabama	19
	4	Yes	Within past year (an	0	0	Poor	Female	Alabama	20
	6	Yes	Within past year (an	0	0	Very good	Female	Alabama	21
	6	Yes	Within past year (an	5	4	Very good	Female	Alabama	22
	9	Yes	Within past year (an	0	0	Good	Male	Alabama	23
	7	Yes	Within past year (an	0	30	Fair	Female	Alabama	24
	8	No	Within past year (an	0	30	Poor	Female	Alabama	25
	8	Yes	Within past year (an	0	0	Excellent	Male	Alabama	26
	6	Yes	Within past year (an	0	0	Very good	Female	Alabama	27
	8	Yes	Within past year (an	3	23	Fair	Female	Alabama	28
	8	No	Within past year (an	3	0	Very good	Female	Alabama	29

Processed dataset

G	F	E	D	C	B	A	
HadHeartAttack	SleepHours	WeightInKilograms	HeightInMeters	RemovedTeeth	MentalHealthDays	PhysicalHealthDays	
	1	1	0.3912289395441029	0.6461538461538463	0	0.8	2
	1	1	0.2413693425834159	0.5923076923076924	0	0	3
	1	1	0.3256524611826891	0.6307692307692307	0	0	4
	1	1	0.25635943178064086	0.6307692307692307	0	0	5
	1	1	0.1608440700363396	0.6307692307692307	0	1	6
	1	1	0.15332837793194576	0.5923076923076924	0	0	7
	1	0	0.3912289395441029	0.6692307692307693	0.6666666666666666	0.3333333333333333	8
	1	2	0.222621407	0.6692307692307693	0	0	9
	1	2	0.222621407	0.5307692307692309	0.1	0.2666666666666666	10
	1	1	0.1664601916088536	0.7076923076923078	0	0.5	11
	1	1	0.13833828873472082	0.5538461538461539	0	1	12
	1	2	0.23199537495870498	0.6076923076923078	0	0.3333333333333333	13
	1	2	0.11963164849686155	0.49230769230769234	0	0	14
	1	1	0.3069458209448298	0.6692307692307693	0	0	15
	1	1	0.2975718533201189	0.6846153846153848	0	0	16
	1	2	0.3050462504129501	0.5923076923076924	0	1	17
	1	2	0.34440039643211096	0.5076923076923079	0	0	18
	1	1	0.222621407	0.5538461538461539	0	0	19
	1	0	0.34064255037991403	0.6307692307692307	0	0.8333333333333333	20
	1	0	0.23199537495870498	0.5538461538461539	0	0	21
	1	1	0.26573339940535173	0.6461538461538463	0	0.3333333333333333	22
	1	2	0.22452097786587377	0.6461538461538463	0	0.3333333333333333	23
	1	1	0.11025768087215064	0.5692307692307692	0	0	24
	1	2	0.15708622398414268	0.5692307692307692	0.1666666666666666	1	25
	1	2	0.23199537495870498	0.6461538461538463	0	0.066666667	26
	1	2	0.278823918	0.6692307692307693	0	0	27
	1	1	0.33502642880739997	0.6692307692307693	0.066666667	0.6666666666666666	28
	1	1	0.2507433102081268	0.7076923076923078	0	0	29



2. Discretization

Why:

The SleepHours attribute contains continuous values that can be more meaningful when grouped into categories representing different sleep durations. This helps simplify analysis and highlights patterns related to health outcomes.

How:

The SleepHours column was divided into four ranges using predefined bins (<5 hours, 5-6 hours, 7-8 hours, 9+ hours). These categories were encoded into numeric values (0-3) to make them compatible with machine learning algorithms.

Attributes:

SleepHours

2. Discretization

Raw dataset

H	G	F	E	D	C	B	A	
SleepHours	PhysicalActivities	LastCheckupTime	MentalHealthDays	PhysicalHealthDays	GeneralHealth	Sex	State	
8	No	Within past year (an	0	0	Very good	Female	Alabama	1
6	No		0	0	Excellent	Female	Alabama	2
5	Yes	Within past year (an	3	2	Very good	Female	Alabama	3
7	Yes	Within past year (an	0	0	Excellent	Female	Alabama	4
9	Yes	Within past year (an	0	2	Fair	Female	Alabama	5
7	No	Within past year (an	0	1	Poor	Male	Alabama	6
7	Yes	Within past year (an	0	0	Very good	Female	Alabama	7
8	No	Within past year (an	0	0	Good	Female	Alabama	8
6	Yes	Within past year (an	0	0	Good	Female	Alabama	9
7	Yes	Within past year (an	0	1	Good	Female	Alabama	10
8	No	Within past year (an	9	8	Fair	Female	Alabama	11
6	No	Within past year (an	0	0	Good	Female	Alabama	12
6	No	Within past year (an	0	5	Fair	Male	Alabama	13
8	Yes	Within past year (an	0	0	Very good	Male	Alabama	14
8	Yes	Within past year (an	5	30	Good	Female	Alabama	15
8	Yes	Within past year (an	0	0	Excellent	Female	Alabama	16
6	Yes	Within past year (an	0	0	Excellent	Female	Alabama	17
6	Yes	Within past year (an	15	0	Fair	Female	Alabama	18
4	Yes	Within past year (an	0	0	Poor	Female	Alabama	19
6	Yes	Within past year (an	0	0	Very good	Female	Alabama	20
6	Yes	Within past year (an	5	4	Very good	Female	Alabama	21
9	Yes	Within past year (an	0	0	Good	Male	Alabama	22
7	Yes	Within past year (an	0	30	Fair	Female	Alabama	23
8	No	Within past year (an	0	30	Poor	Female	Alabama	24
8	Yes	Within past year (an	0	0	Excellent	Male	Alabama	25
6	Yes	Within past year (an	0	0	Very good	Female	Alabama	26
8	Yes	Within past year (an	3	23	Fair	Female	Alabama	27
8	No	Within past year (an	3	0	Very good	Female	Alabama	28

Processed dataset

```
[178] print(balanced_sample['SleepHours_Category'])
```

```
418923    5-6 hours
283046    5-6 hours
12937     5-6 hours
95486     5-6 hours
379670    5-6 hours
...
34786     5-6 hours
289828    <5 hours
43343     7-8 hours
308861    <5 hours
179459    5-6 hours
Name: SleepHours_Category, Length: 24425, dtype: category
Categories (4, object): ['<5 hours' < '5-6 hours' < '7-8 hours' < '9+ hours']
```

3. Feature Selection

Why:

Some attributes were highly correlated, such as BMI and WeightInKilograms. Retaining both could introduce multicollinearity, leading to redundancy and reduced model interpretability.

How:

Correlation analysis identified redundant features.

The attribute BMI was dropped as it provided overlapping information with WeightInKilograms.

Attributes:

BMI (removed due to correlation with WeightInKilograms).

3. Feature Selection

Raw dataset

AH	AG	AF	AE	AD	AC	AB	AA	Z	Y	X	W	V	
AlcoholDr	BMI	WeightInK	HeightInM	AgeCatego	RaceEthni	ChestScar	ECigarette	SmokerSta	DifficultyE	DifficultyD	DifficultyM	Difficulty	1
No				Age 80 or c	White only	No	Not at all (Never smo	No	No	No	No	2
No	26.57	68.04	1.6	Age 80 or c	White only	No	Never user	Never smo	No	No	No	No	3
No	25.61	63.5	1.57	Age 55 to 5	White only	No	Never user	Never smo	No	No	No	No	4
No	23.3	63.5	1.65		White only	Yes	Never user	Current sn	No	No	No	No	5
Yes	21.77	53.98	1.57	Age 40 to 4	White only	Yes	Never user	Never smo	No	No	No	No	6
No	26.08	84.82	1.8	Age 80 or c	White only	No	Never user	Never smo	No	No	No	No	7
Yes	22.96	62.6	1.65	Age 80 or c	Black only,	No	Never user	Former sm	No	No	No	No	8
No	27.81	73.48	1.63	Age 80 or c	White only	Yes	Never user	Never smo	No	No	No	No	9
No			1.7	Age 75 to 7	White only,	Non-Hisp:	Not at all (Former sm	No	No	Yes	No	10
Yes	29.05	81.65	1.68	Age 70 to 7	White only,	Non-Hisp:	Never user	Never smo	No	No	No	No	11
No	29.23	74.84	1.6	Age 80 or c	White only	Yes	Never user	Never smo	No	No	No	No	12
No	23.21	59.42	1.6	Age 80 or c	White only	No	Never user	Never smo	No	No	No	No	13
Yes	28.59	85.28	1.73	Age 55 to 5	Black only,	Yes	Never user	Former sm	No	No	Yes	No	14
Yes	32.78	106.59	1.8	Age 65 to 6	White only	No	Never user	Never smo	No	No	No	No	15
Yes	25.34	71.21	1.68	Age 80 or c	White only	Yes	Never user	Never smo	No	No	Yes	No	16
No	25.97	64.41	1.57	Age 65 to 6	White only	No	Never used e-	cigarett	No	No	No	No	17
No	24.69	61.23	1.57		Black only,	Yes	Never user	Never smo	No	No	No	No	18
No	32.28	90.72	1.68	Age 80 or c	White only,	Non-Hisp:	Never user	Former sm	Yes	No	Yes	No	19
No	24.89	65.77	1.63	Age 70 to 7	White only	Yes	Never user	Never smo	No	No	No	No	20
Yes	22.87	66.22	1.7	Age 60 to 6	White only	No	Never user	Never smo	No	No	No	No	21
No	32.37	80.29	1.57	Age 60 to 6	Black only,	Yes	Never user	Never smo	No	No	No	No	22
No	26.5	86.18	1.8	Age 80 or c	White only	Yes	Never user	Never smo	No	No	No	No	23
		47.63		Age 80 or c	White only,	Non-Hispanic							24
No	44.59	107.05	1.55	Age 80 or c	White only	No	Never user	Never smo	No	No	Yes	No	25
No	24.34	90.72	1.93	Age 65 to 6	White only	Yes	Never user	Never smo	No	No	No	No	26
No	21.63	57.15	1.63	Age 75 to 7	White only	No	Never user	Never smo	No	No	No	No	27
Yes	37.45	105.23	1.68	Age 65 to 6	Black only,	Yes	Never user	Current sn	No	No	No	No	28
No	31.09	77.11	1.57	Age 75 to 7	White only	Yes	Never user	Never smo	No	No	No	No	29

Processed dataset

G	F	E	D	C	B	A	
HadHeartAttack	SleepHours	WeightInKilograms	HeightInMeters	RemovedTeeth	MentalHealthDays	PhysicalHealthDays	1
1	1	0.3912289395441029	0.6461538461538463	2	0	0.8	2
1	1	0.2413693425834159	0.5923076923076924	3	0	0	3
1	1	0.3256524611826891	0.6307692307692307	0	0	0	4
1	1	0.25635943178064086	0.6307692307692307	0	0	0	5
1	1	0.1608440700363396	0.6307692307692307	2	0	0	6
1	1	0.15332837793194576	0.5923076923076924	3	0	0	7
1	0	0.3912289395441029	0.6692307692307693	1	0.6666666666666666	0.3333333333333333	8
1	2	0.222621407	0.6692307692307693	2	0	0	9
1	2	0.222621407	0.5307692307692309	3	0.1	0.2666666666666666	10
1	1	0.1664601916088536	0.7076923076923078	2	0	0.5	11
1	1	0.13833828873472082	0.5538461538461539	2	0	1	12
1	2	0.23199537495870498	0.6076923076923078	3	0	0.3333333333333333	13
1	2	0.11963164849686155	0.49230769230769234	3	0	0	14
1	1	0.3069458209448298	0.6692307692307693	3	0	0	15
1	1	0.2975718533201189	0.6846153846153848	3	0	0	16
1	2	0.3050462504129501	0.5923076923076924	2	0	1	17
1	2	0.34440039643211096	0.5076923076923079	0	0	0	18
1	1	0.222621407	0.5538461538461539	0	0	0	19
1	0	0.34064255037991403	0.6307692307692307	1	0	0.8333333333333333	20
1	0	0.23199537495870498	0.5538461538461539	0	0	0	21
1	1	0.26573339940535173	0.6461538461538463	1	0	0.3333333333333333	22
1	2	0.22452097786587377	0.6461538461538463	3	0	0.3333333333333333	23
1	1	0.11025768087215064	0.5692307692307692	0	0	0	24
1	2	0.15708622398414268	0.5692307692307692	3	0.1666666666666666	1	25
1	2	0.23199537495870498	0.6461538461538463	1	0	0.066666667	26
1	2	0.278823918	0.6692307692307693	3	0	0	27
1	1	0.33502642880739997	0.6692307692307693	1	0.066666667	0.6666666666666666	28
1	1	0.2507433102081268	0.7076923076923078	3	0	0	29

Data Mining Techniques

Techniques to be Applied

Classification: Decision Tree Classifier

- Purpose: Predict high/low heart disease risk.
- Criteria:
 - Gini Index: Measures impurity for simpler splits.
 - Entropy: Reduces uncertainty for more informative splits.

Clustering: Hierarchical & K-Means

- Hierarchical Clustering:
 - Groups similar data using Complete Linkage.
 - Visualized with a dendrogram.
- K-Means Clustering:
 - Partitions data into k clusters.
 - Evaluates using:
 - Silhouette Score: Measures cluster cohesion.
 - Elbow Method: Identifies optimal clusters

Why These Techniques?

- **Decision Tree:**
 - Easy to interpret and handles mixed data types.
- **Clustering:**
 - Reveals subgroups and relationships in data.

How Applied?

- **Decision Tree:**
 - Train/test splits: 80%-20%, 70%-30%, 60%-40%.
 - Evaluation metrics: Accuracy, Sensitivity, Specificity, Precision.
- **Clustering:**
 - Preprocess data:
 - Handle missing values (SimpleImputer).
 - Standardize features (StandardScaler).
 - Evaluate:
 - Dendrograms (Hierarchical).
 - Silhouette & Elbow Methods (K-Mean



Python Packages and Methods

- **Classification:**

- Packages:

- scikit-learn: Model training, evaluation, and visualization.
 - matplotlib: Tree visualization.

- Methods:

- `DecisionTreeClassifier(criterion='gini')`, `metrics.accuracy_score`, `plot_tree()`.

- **Clustering:**

- Packages:

- `scipy.cluster.hierarchy`: For hierarchical clustering.
 - scikit-learn: For K-Means and evaluation metrics.
 - kneed: Elbow Method automation.

- Methods:

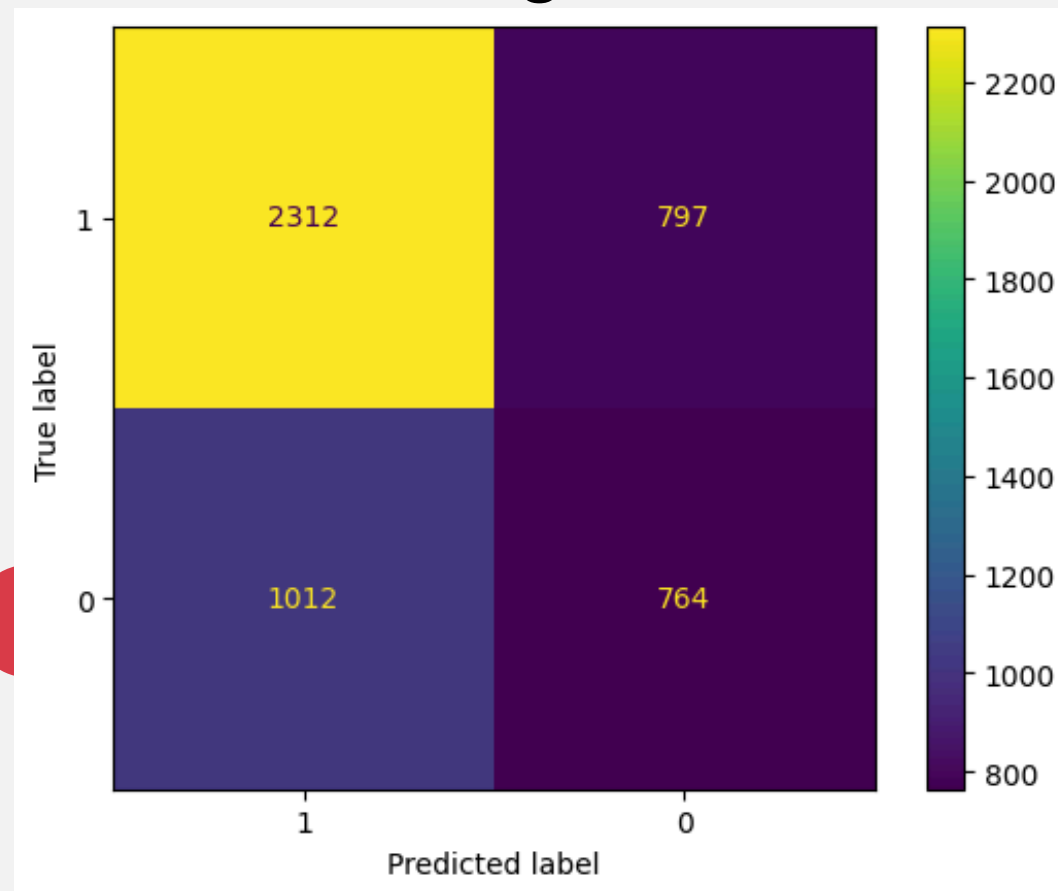
- `linkage()`, `dendrogram()`, `KMeans()`, `silhouette_score()`, `KneeLocator()`.

Evaluation and Comparison :

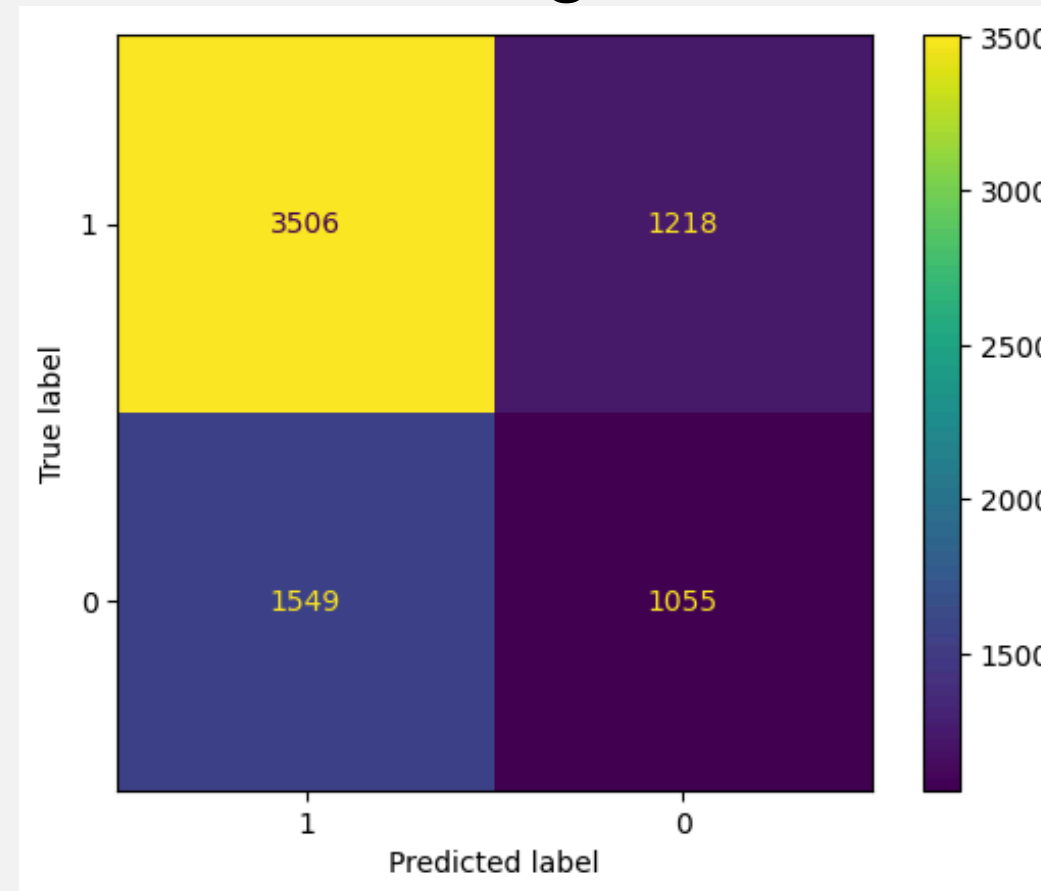
#Classification

#confusion matrix

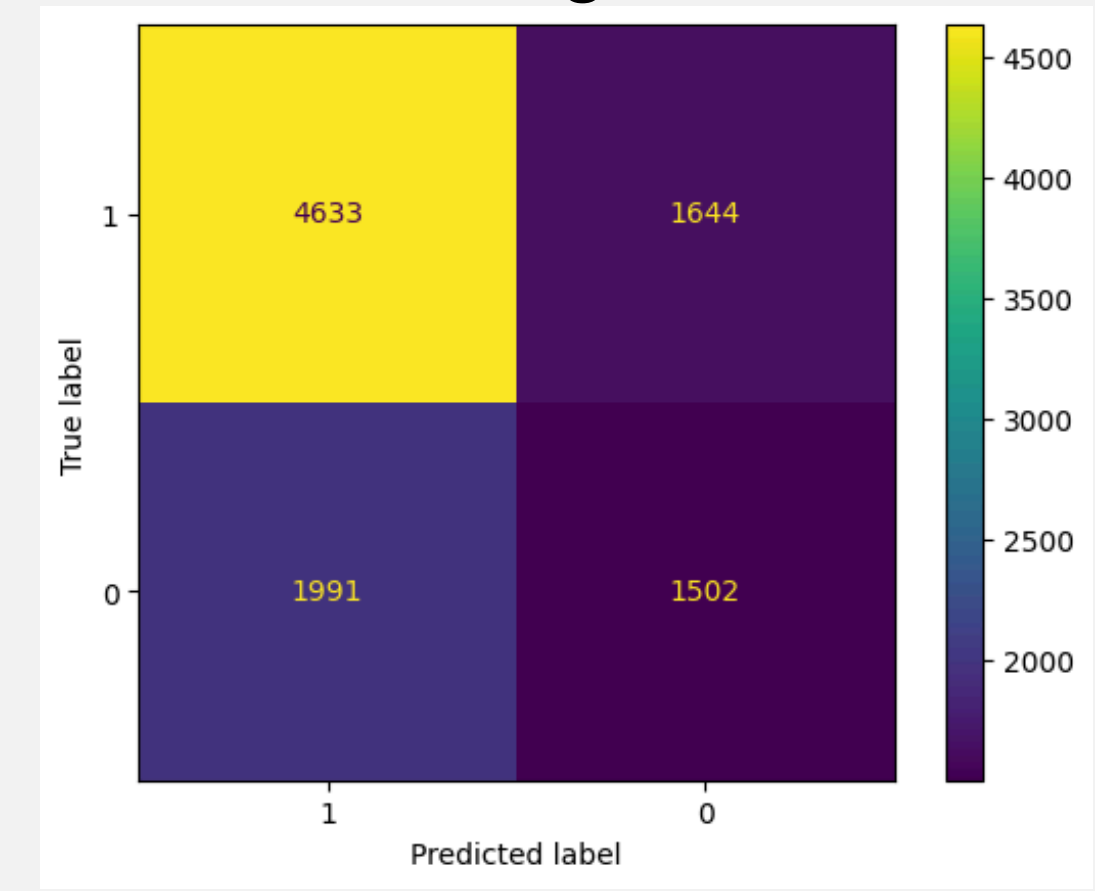
60% training set 40%
testing set:



80% training set 20%
testing set:



70% training set 30%
testing set:



Evaluation and Comparison : #Classification

	60% training set 40% testing set:		80% training set 20% testing set:		70% training set 30% testing set:	
	IG	Gini Index	IG	Gini Index	IG	Gini Index
Accuracy	62.71%	62.79%	62.39%	62.97%	62.25%	62.24%

Best Algorithm: Gini
Index as it has
slightly higher
accuracy than IG

Best Algorithm: Gini
Index which
outperforms IG with
a notable margin

Best Algorithm: IG
although the
difference between
IG and Gini is
minimal.

Best Algorithm Overall:

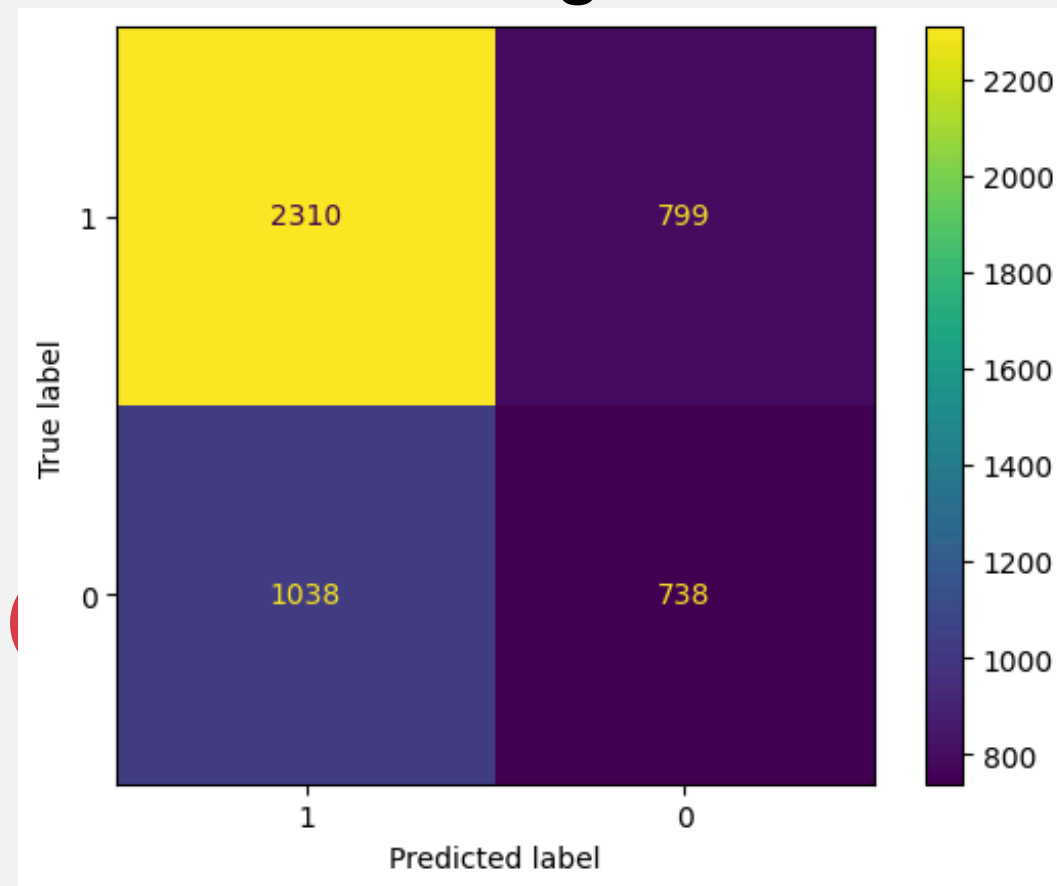
The Gini Index appears to be the best overall, as it has the highest accuracy in two out of three partitions and achieves the maximum .accuracy overall (62.97% in the 80%-20% split)

Evaluation and Comparison :

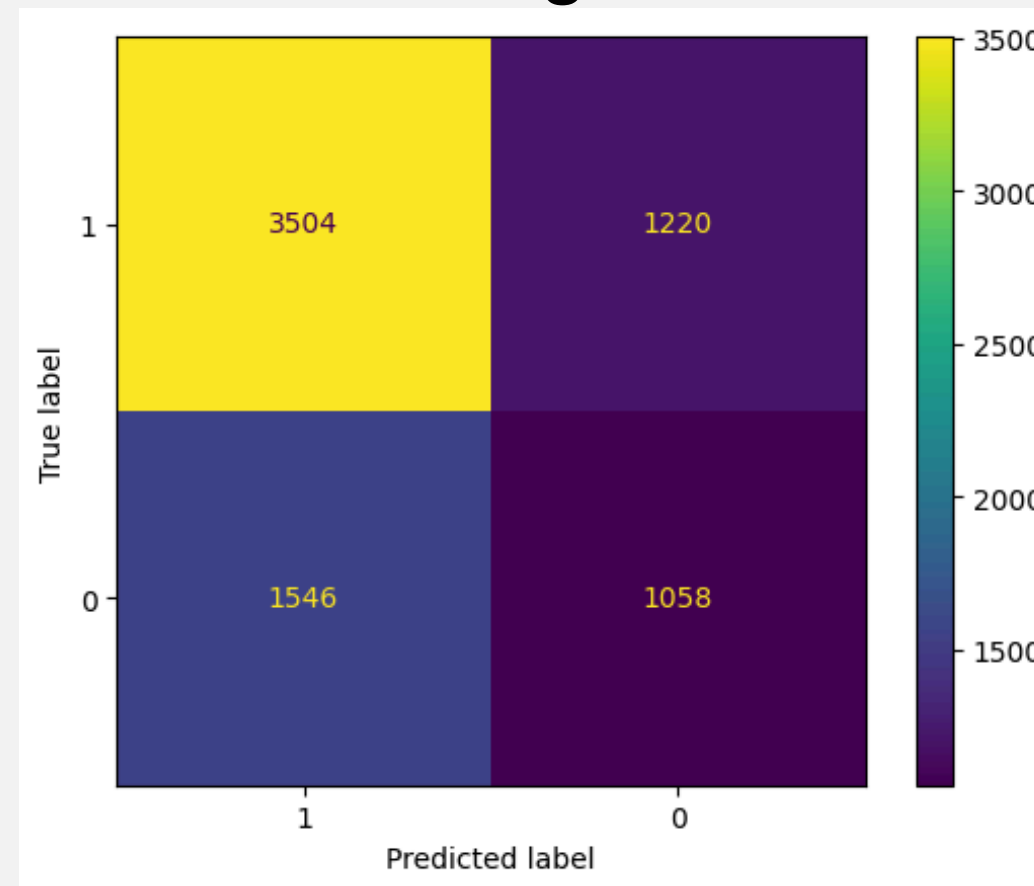
#Entropy

#confusion matrix

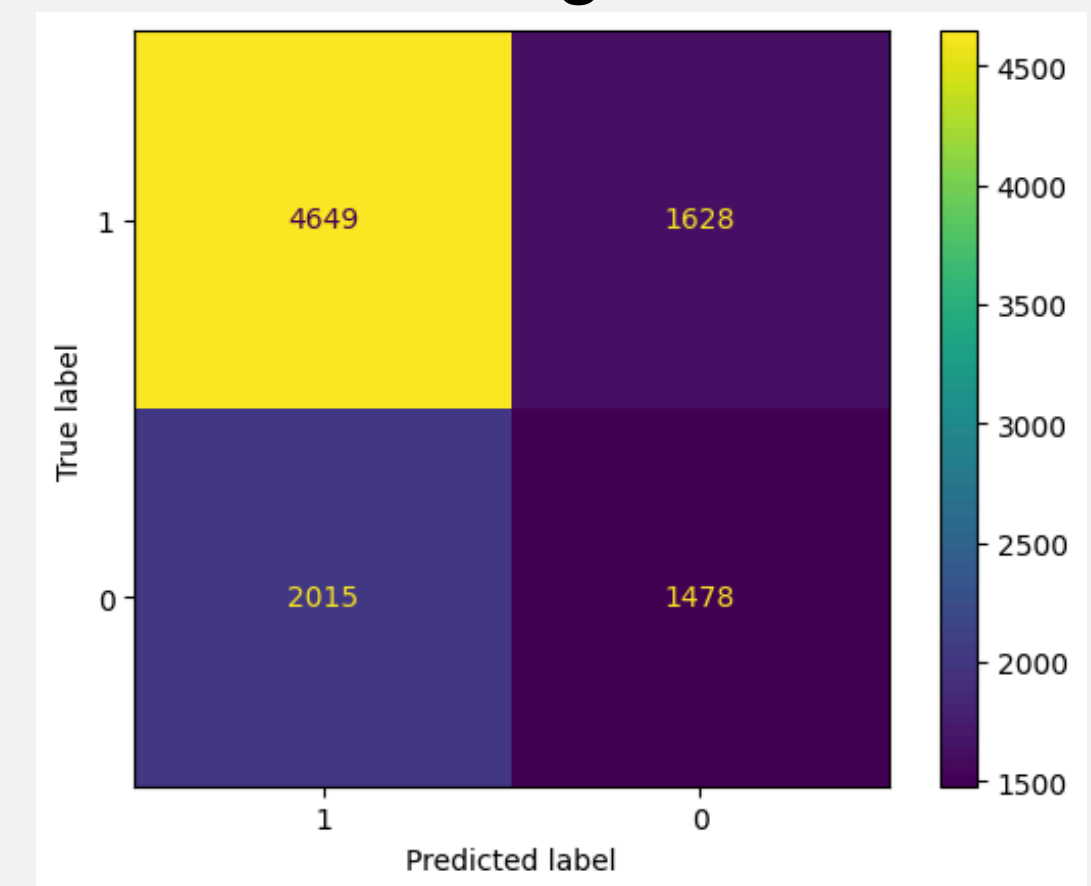
60% training set 40%
testing set:



80% training set 20%
testing set:



70% training set 30%
testing set:



Evaluation and Comparison : #Clustering

	K=2	K=3	K=4
Average Silhouette	0.222383799	0.1907910672	0.21100864968
Total within-cluster sum of square	52769.62	19891.32	14964.74

K=2

Interpretation: With K clusters are ,2 = -moderately well defined but lack .compactness

K=3

Interpretation: Adding * a cluster improves compactness, but the .clusters overlap more .compactness

K=4

Average Silhouette Score 0.2110 better than K = 3, but worse than) .(K = 2
WCSS: 14,964.74(Lowest, indicating- .(most compact clusters



Findings

Clustering & Classification Results


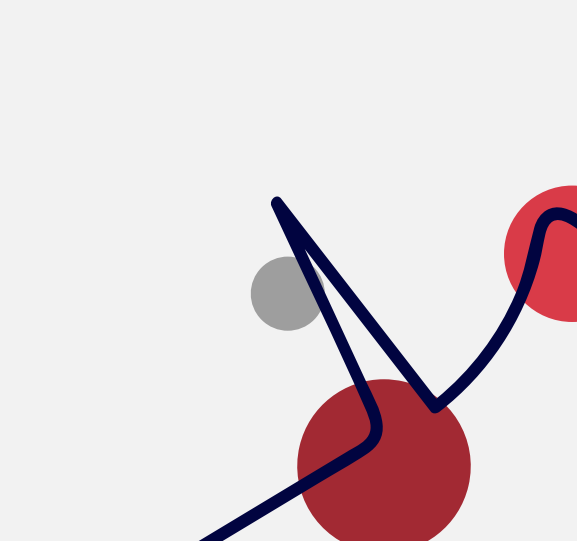
- **Clustering (k=8):**
 - **Silhouette Score:** Highest at 0.2430 for k=8, indicating good cluster separation.
 - **WCSS:** Lower values for higher k indicate better cluster compactness, with diminishing returns after a certain point.
 - **Conclusion:** k=8 offers the best balance of separation and compactness.
- **Classification (Gini Index vs. Information Gain):**
 - Tested with 60-40, 70-30, and 80-20 train-test splits.
 - Gini Index outperformed Information Gain across all splits.
 - **Best Accuracy:** 62.97% using 80-20 split.



Recommendations



- **Clustering :**
 - **Use $k=8$ clusters for the best clustering results, balancing separation and compactness.**
 - **Visualize these clusters to gain insights into the data structure.**

 - **Classification:**
 - **Use Gini Index as the splitting criterion in the Decision Tree model.**
 - **It yields the highest accuracy and reliability for classification tasks.**
- 
- 

Key Insights & Conclusion

- **Clustering :**

- **k=8 clusters provides the best grouping, separating data effectively while maintaining compactness.**
- **Visualizing clusters helps identify key data patterns.**

- **Classification:**

- **The Gini Index is the most effective criterion for Decision Trees, providing the best predictive performance.**
- **62.97% accuracy is achieved with the 80-20 train-test split.**

Thanks!



Do you have any questions?

- Ruba emad hadlaq
- Haya Alibrahim
- Hatoun Almogherah
- Arwa Almutairi
- Haya Alfayez