Faculty of Engineering & Technology

Computer Science Department

Machine Learning Algorithms COMP 4388

Report #2

Smoker Prediction using ML and DL

(predicts if a given person has smoker or not)

Prepared by:

Tareq Khanfar 1200265

Instructor: DR.Radi Jarrar

Section: 1

Date:25/1/2024

Introduction:

This project aims to classify if a person smoker or not smoker using their health feature and personal details. We will explore at the data and make some visualization , and test different machine learning and deep learning algorithms to see which one works best for telling if someone is a smoker or not.

Exploratory Data Analysis (EDA)

In this project, I visualized data before using machine learning. So it helps us understand the data better. Here's what I did:

Reading the Data : I used the ReaderData.py code to load the data and see what's inside its .
this is a first 5 rows from the data set :

Age	Gender	BMI	Region	No. Children	Insurance Charges	Smoker
27	male	30.5	north	0	2494.022	no
41	female	31.02	north	0	6185.3208	no
28.47004946	male	26.66785034	north	4	5549.324781	no
50.30353686	female	30.8816603	south	4	11366.35084	no
50.77674293	male	26.84404232	south	3	25729.18463	yes

These columns represent the following :

Age : The age of a person

Gender : the gender of a person

BMI : body mass index
Region : north or south
No. children : number of children
insurance Charges : the amount paid for the insurance company by the person

Smoker : yes for positive (smoker) and no otherwise .

I notice that there are two categorical columns, which means they contain non-numeric, string values . These columns are:

Gender: This column classifies into categories based on their gender such that male or female.

Region: This column classifies into categories based on their gender such that north or south.

Smoker : This column classifies into categories based on their smoker such that yes or no.

So the algorithm that will be used in this project its need the numerical data to work successfully .
so must be convert these categories values to numerical values .
there is a many algorithm that using to convert to numeric values but since the data is a binary categories I am used a label encoding approach to handle a non numeric values .

Therefore, I adopted the following for non-numeric columns as this :
for Gender Column : male = 1 and female = 0 .
for Region Column : south = 1 and north = 0 .

For Smoker : yes = 1 and no = 0 ;

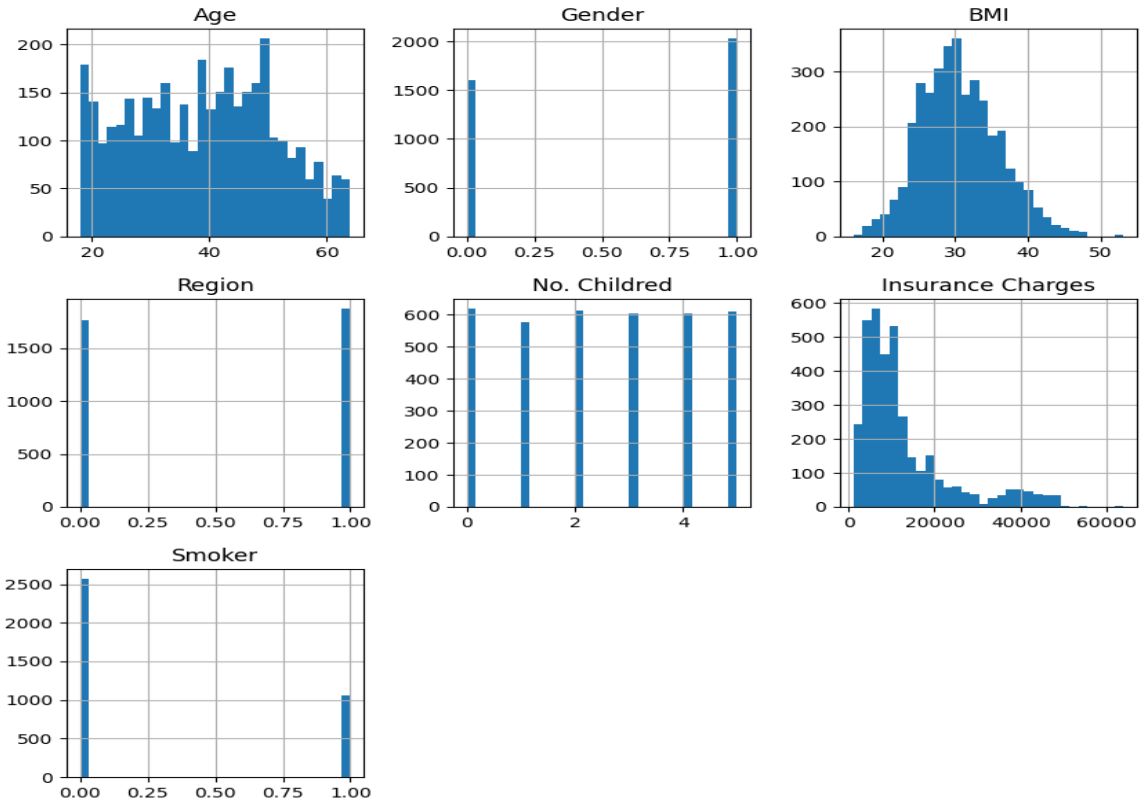
Now after encoding the non numeric values to numeric values . the data become as this :

Age	Gender	BMI	Region	No. Children	Insurance Charges	Smoker
27.000000	1	30.50000	1	0	2494.022000	0
41.000000	0	31.02000	1	0	6185.320800	0
28.470049	1	26.66785	1	4	5549.324781	0
50.303537	0	30.88166	0	4	11366.350840	0

Here is a summery statistics about data set :

Column	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Age	3630.000	38.887	12.151	18.000	29.000	39.171	48.343	64.000
Gender	3630.000	0.559	0.497	0.000	0.000	1.000	1.000	1.000
BMI	3630.000	30.630	5.441	15.960	26.695	30.200	34.100	53.130
Region	3630.000	0.515	0.500	0.000	0.000	1.000	1.000	1.000
No. Children	3630.000	2.504	1.713	0.000	1.000	3.000	4.000	5.000
Insurance Charges	3630.000	12784.809	10746.167	1121.874	5654.818	9443.807	14680.408	63770.428
Smoker	3630.000	0.293	0.455	0.000	0.000	0.000	1.000	1.000

In the below image represent the visualization for dataset :



Data Cleaning :

1 - For my smoker dataset, we did not find any missing but there is a 668 Raws are completely duplicated . I think you took a subset of the Dataset and then copied and pasted it .

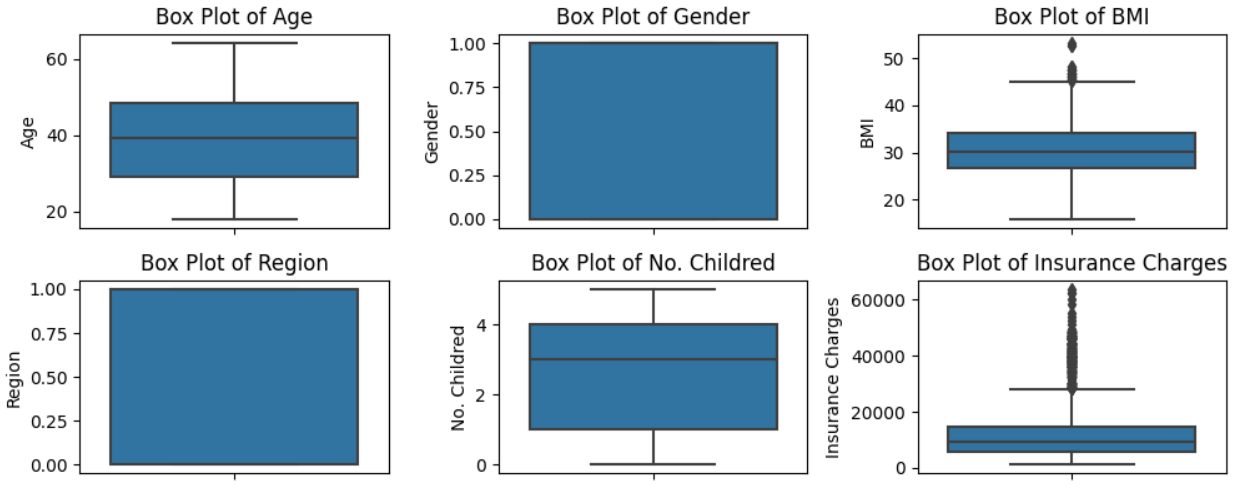
Therefore, I removed these duplicate rows from the data set. the size of data set after that = 2963 raw .

2- Conversion of Age Data from Double to Integer Values

Next, I notice that the "Age" column in the data set contain a decimal values i.e. float values , such as 48.44, but this is not correct . To fix this problem, I rounded all age values and converted them to integers .

3- Check the Outliers :

In the this steps I am visualize the data using box plot to determine if there is outliers or not .



We conclude from the image above that there are outliers in the BMI and insurence charge columns .

Therefore This data must be cleaned by replacing it with appropriate values to increase the strength of the models later on .

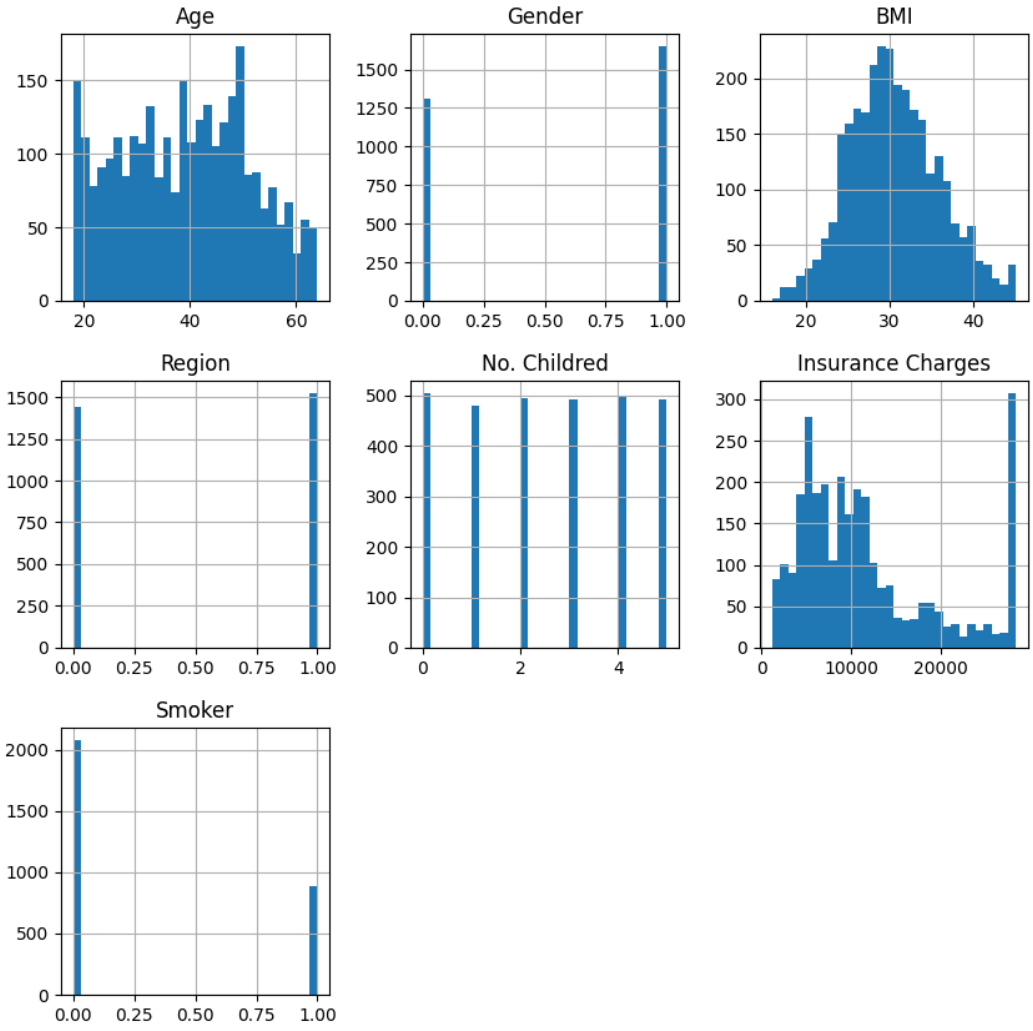
Handling Outliers :

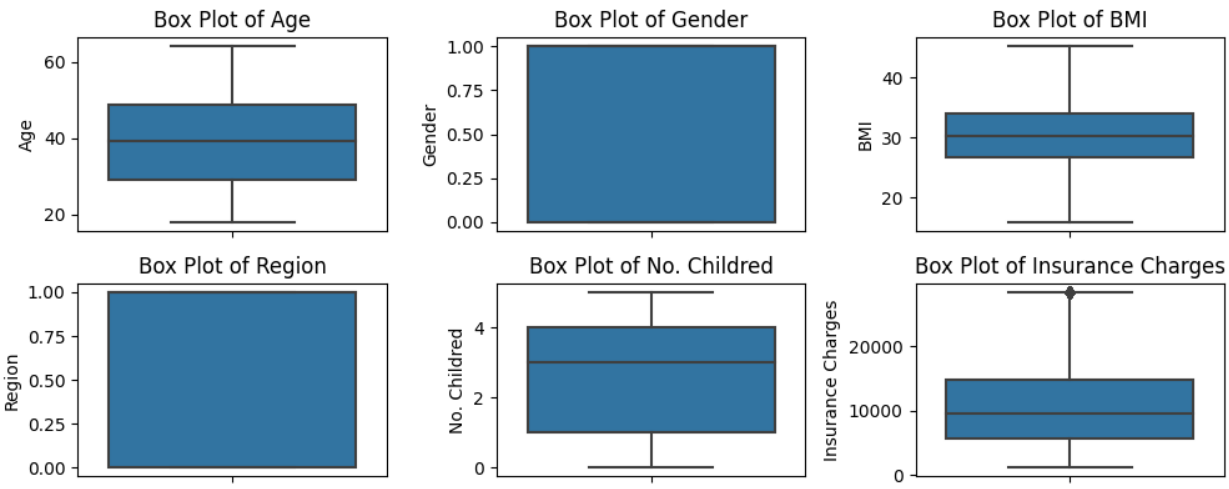
cap_outliers, designed to reduce outliers. It calculates the acceptable data range using the interquartile range (IQR) and sets extreme values accordingly. Values below Q1 - 1.5IQR or above Q3 + 1.5IQR are set to these limits. This method maintains the integrity of the data while minimizing the impact of outliers

Data After Cleaning

The description of dataset after cleaning noise :

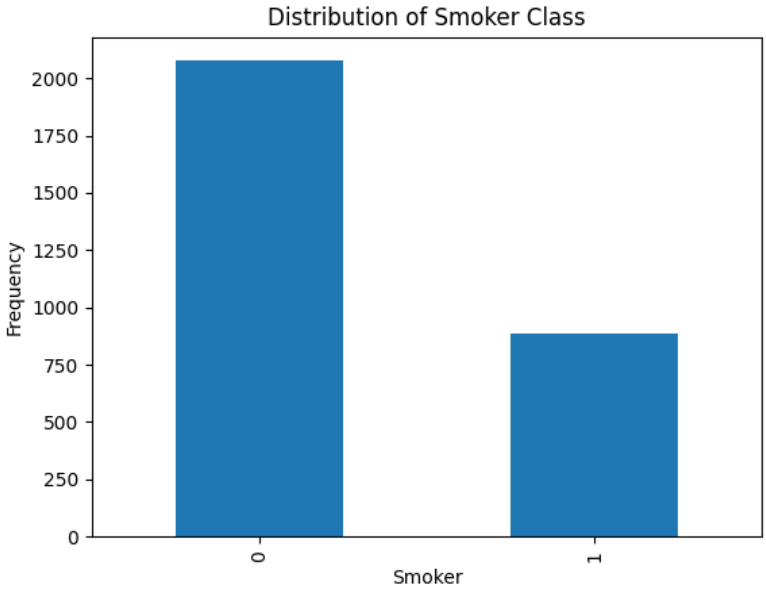
Column	Count	Mean	Std	Min	25%	50%	75%	Max
Age	**2962.0**	**39.0**	**12.20**	**18.0**	**29.000**	**39.0**	**48.0**	**64.000**
Gender	**2962.000**	**0.556**	**0.49**	**0.000**	**0.000**	**1.000**	**1.000**	**1.000**
BMI	**2962.000**	**30.608**	**5.36**	**15.0**	**26.730**	**30.210**	**34.099324**	**45.153310**
Region	**2962.000**	**0.51**	**0.499902**	**0.000**	**0.000**	**1.000**	**1.000**	**1.000**
No. Children	**2962.000**	**2.5**	**1.7118**	**0.000**	**1.000**	**3.000**	**4.000**	**5.000**
Insurance Charges	**2962.000**	**11705.78**	**7838.632999**	**1121.873900**	**5709.962155**	**9563.616073**	**14791.278335**	**28413.252605**
Smoker	**2962.000**	**0.298**	**0.457802**	**0.000**	**0.000**	**0.000**	**1.000**	**1.000**


Age
Gender
BMI
Region
No. Childred
Insurance Charges
Smoker

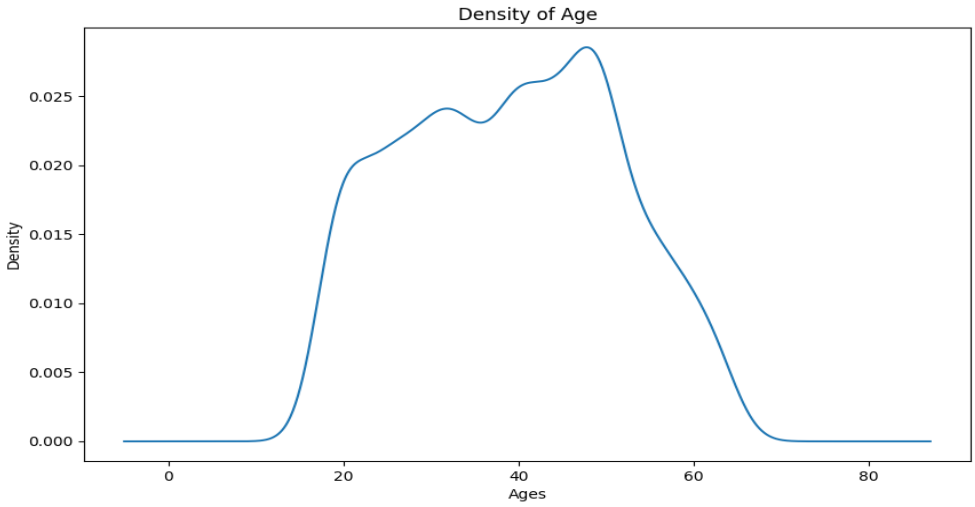

Box Plot of Age
Age
60
40
20
Box Plot of Gender
Gender
1.00
0.75
0.50
0.25
0.00
Box Plot of BMI
BMI
40
30
20
Box Plot of Region
Region
1.00
0.75
0.50
0.25
0.00
Box Plot of No. Childred
No. Childred
4
2
0
Box Plot of Insurance Charges
Insurance Charges
20000
10000
0

Data And Analysis :

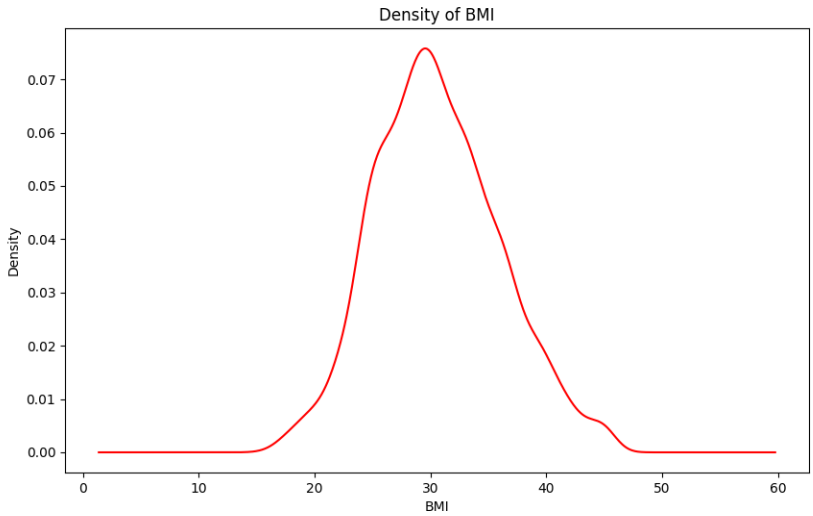
The chart shows the number of smoker and not smoker people.



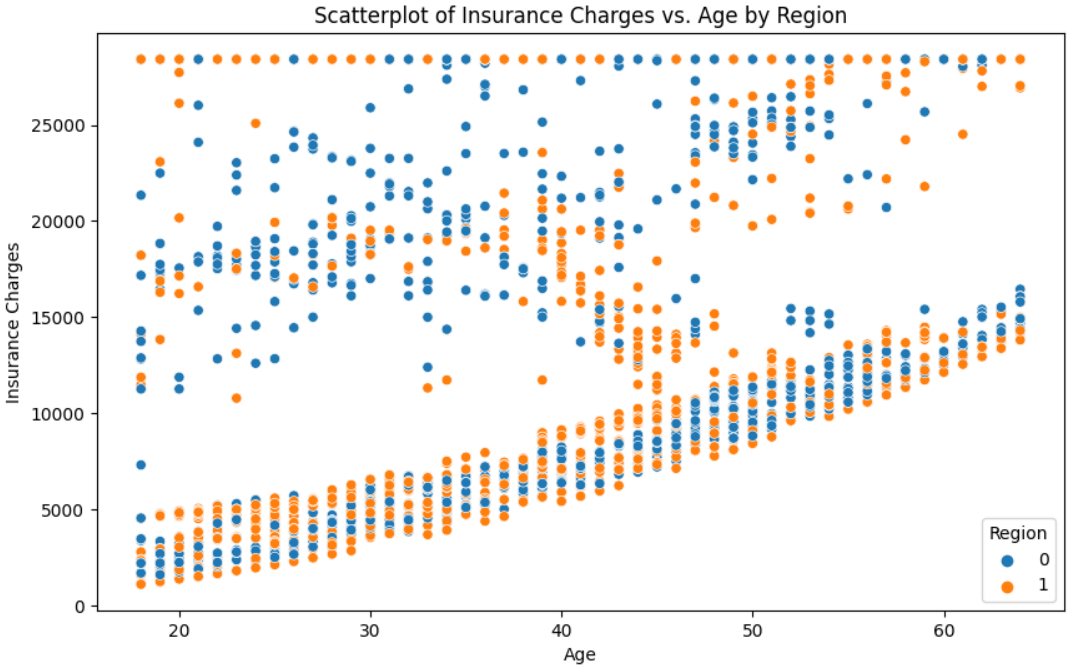
The density chart for age shows an increase in the age group between approximately 20 and 55 years of age and then a gradual decrease with increasing age.



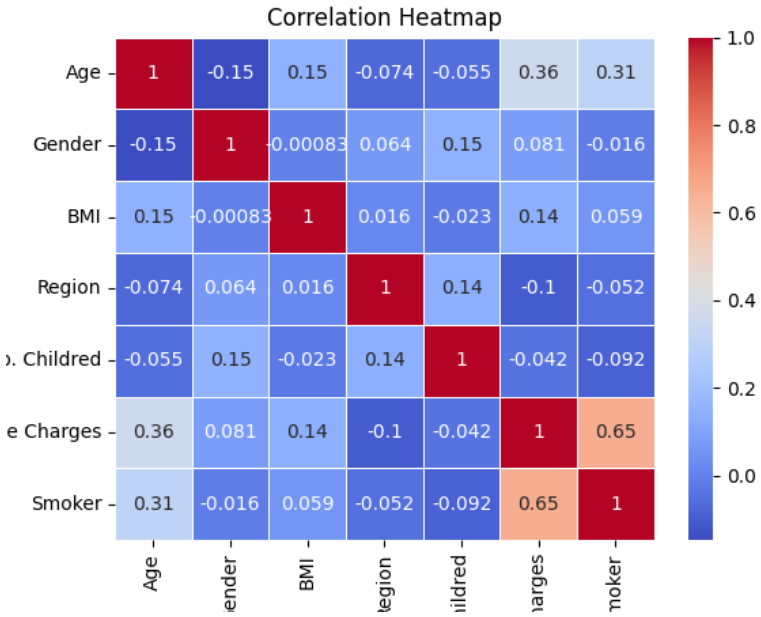
density plot for BMI, with a peak around the middle range at 30 age and tails off towards the lower and higher BMI values



Result from the visualization that as people get older, they usually have to pay more for insurance, regardless of where they live.



Visualize the correlation between all features



Insurance Charges: has 0.65, which is a strong positive with Smoker coulmn . This means that people who are smokers tend to have higher insurance charges.

Age has 0.31, which is a middle positive link. This suggests that older people might be more likely to be smokers.

Region : has -0.052, which is a very weak negative link. This means there isn't much of a pattern to where people live and if they smoke.

No . Children : this means is that there is a slight tendency for individuals with more children to be is not smokers

BMI : This indicates that there is a very slight increase in BMI among smokers compared to not smokers

Gender : This indicates that there is hardly any relationship between gender and smoking status in this dataset

Normalization of Data (Feature Scalling) :

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. [1]

In this project I applied the Min-Max scale, which is to normalize the data. This method transforms features by scaling each one to a specific range, from 0 to 1. The formula used is (value - min) / (max - min).

Split the Data Set :

In this project, I divided the data into two parts: 80% for training, and 20% for testing. This means training the models and ensuring their validity and generalizability. So how to detect smokers and non-smokers.

I am use these feature as an inputs : Age , Insurance Charges , No.children .

note : I am select a NO.children after trying several testing on model . without use it’s the result was not good . but when use it’s the result it become very good . so I am select its .

Machine Learning Algorithms

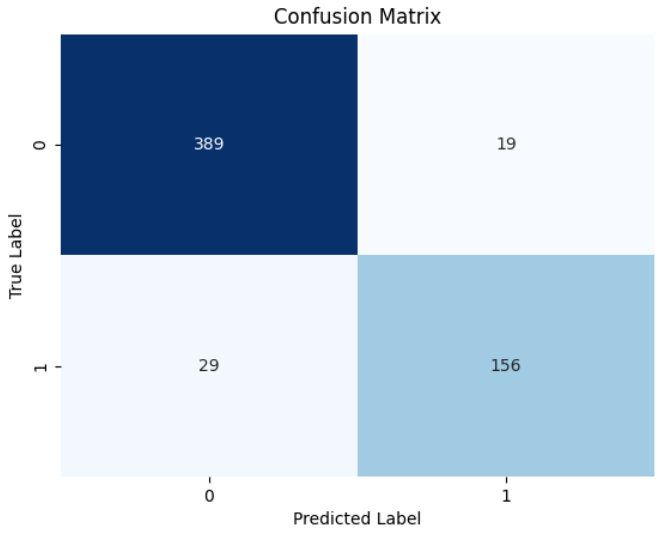
KNN Models:

in this algorithms chose to focus on 3 main features Age , Insurance charges , NO. children

I am train a 3 Models with different value of K .

in this case confusion Matrix is :

when k = 3 . the confusion matrix as this :



True Negative : 389 : The model correctly predicted

False Positive : 19 : The model incorrectly predicted

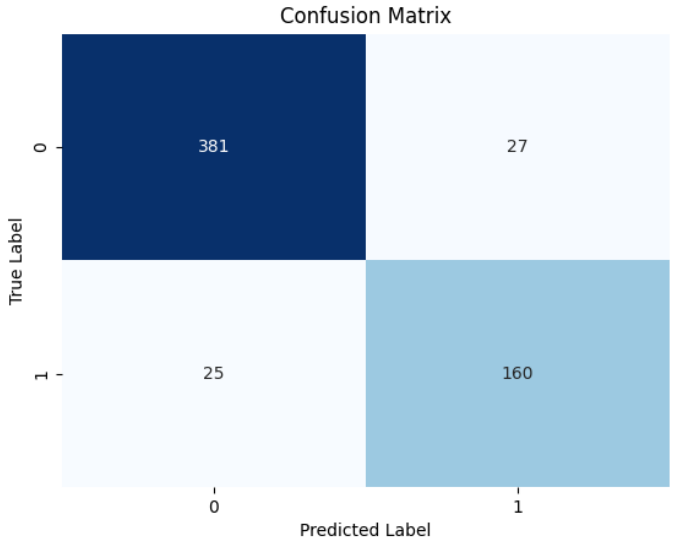
False Negative: 29 : The model incorrectly predicted

True Positive : 156 : The model correctly predicted

Performance :

Metric	Value
Accuracy	92%
Precision	89.6%
Recall	84%
F1 Score	87%
ROC AUC Score	95%

when k = 5 . the confusion matrix as this :



True Negative : 381 : The model correctly predicted

False Positive : 27 : The model incorrectly predicted

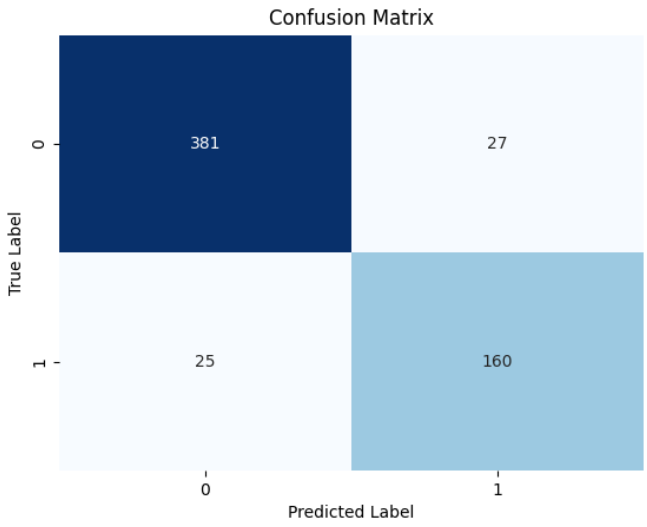
False Negative: 25 : The model incorrectly predicted

True Positive : 160 : The model correctly predicted

Performance :

Metric	Value
Accuracy	91%
Precision	86%
Recall	86%
F1 Score	86%
ROC AUC Score	96%

when k = 7 . the confusion matrix as this :



True Negative : 381 : The model correctly predicted

False Positive : 27 : The model incorrectly predicted

False Negative: 25 : The model incorrectly predicted

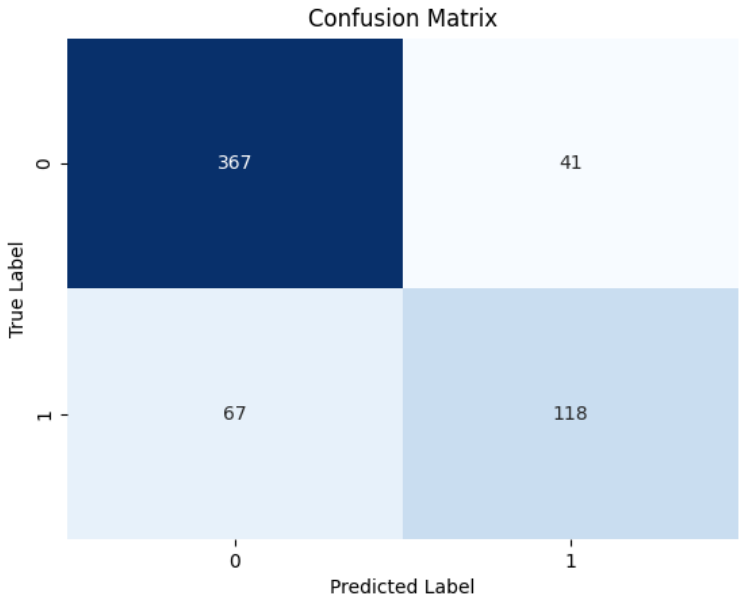
True Positive : 160 : The model correctly predicted

Performance :

Metric	Value
Accuracy	91%
Precision	85%
Recall	86%
F1 Score	86%
ROC AUC Score	96.6%

For Naive Byes algorithm :

The confusion matrix as this :



True Negative : 367 : The model correctly predicted

False Positive : 41 : The model incorrectly predicted

False Negative: 67 : The model incorrectly predicted

True Positive : 118 : The model correctly predicted

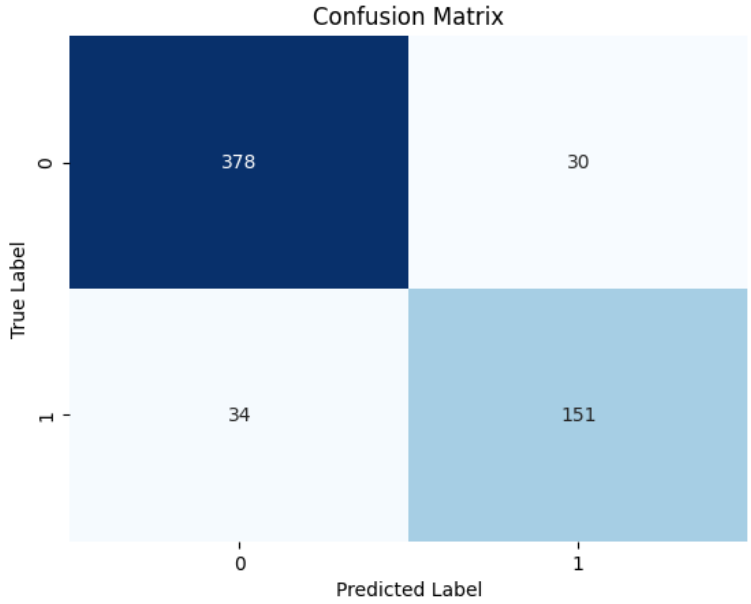
Performance :

Metric	Value
Accuracy	82%
Precision	74%
Recall	64%
F1 Score	69%
ROC AUC Score	92%

Decision Tree (C4 .5) Algorithm :

in this algorithm I am use a chefboost in python . and I am use a weka.exe GUI . but regarding for weka.jar in java not work . I am trying several attempts. But not work . I don't know what is the problem .

the Confusion matrix that generated by python :



True Negative : 378 : The model correctly predicted

False Positive : 30 : The model incorrectly predicted

False Negative: 34 : The model incorrectly predicted

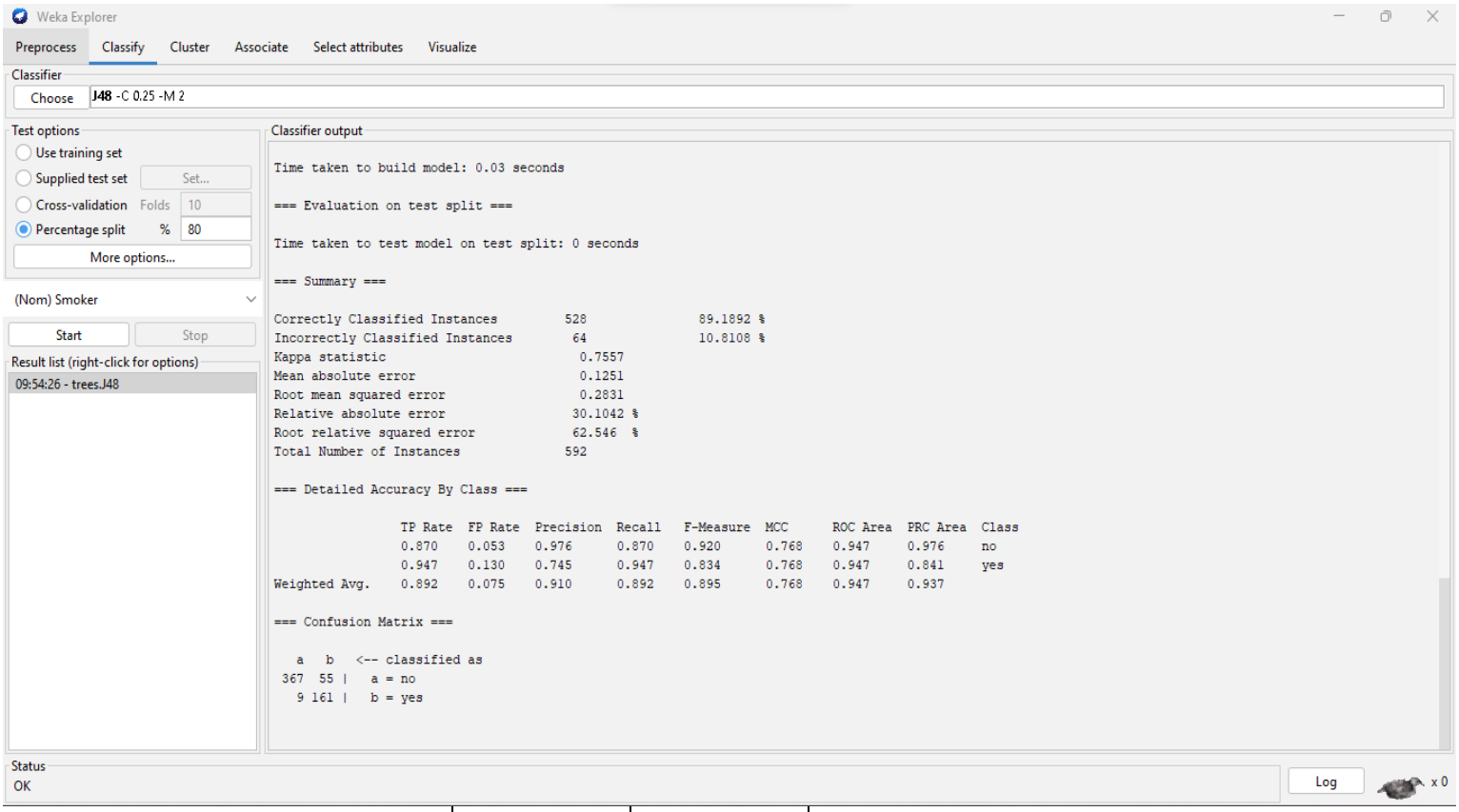
True Positive : 151 : The model correctly predicted

Performance :

Metric	Value
Accuracy	89.2%
Precision	89.1%
Recall	89.2%
F1 Score	89.1%
ROC AUC Score	87%

This is a screen shot for c4.5(J48 is called in weka) algorithm :

The Confusion matrix :



367	**55**
9	**161**

True Negative : **367**: The model correctly predicted

False Positive : **55**: The model incorrectly predicted

False Negative: **9**: The model incorrectly predicted

True Positive : **161**: The model correctly predicted

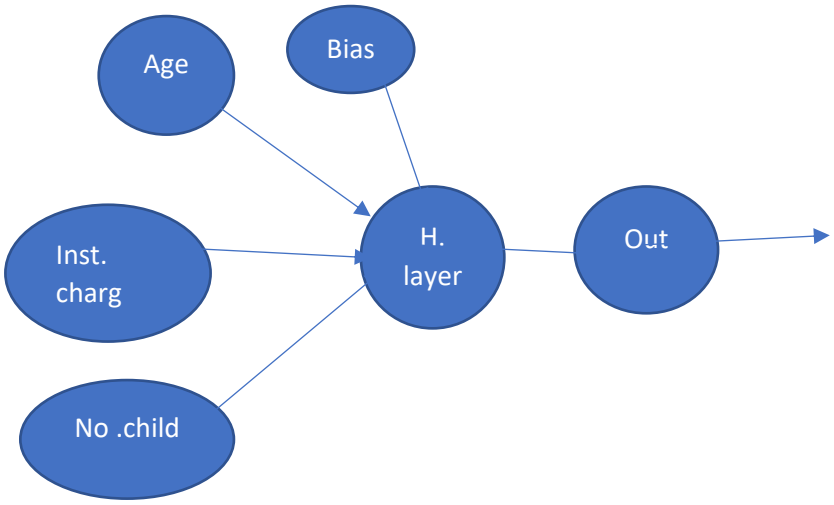
Performance :

Metric	Value
Accuracy	89.19%
Precision	91%
Recall	89%
F1 Score	89.1%
ROC AUC Score	94.7%

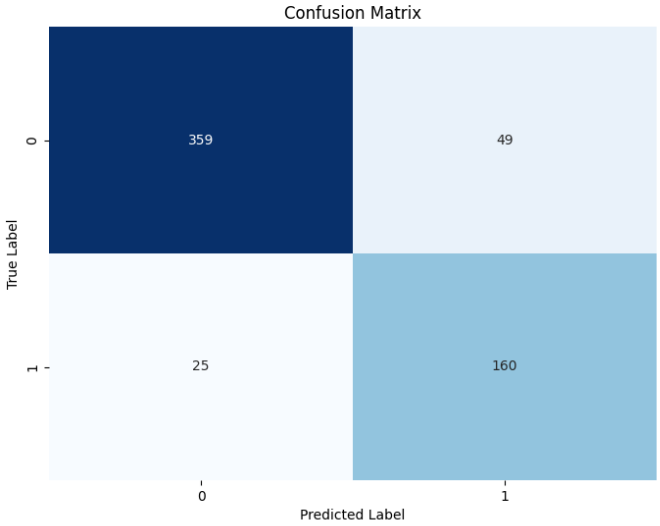
Deep learning Algorithm : ANN

This is a neural network that consist of a 3 layers . as a following : input layer , single hidden layer and single out put layer .

so in this section I will to train these network by the following feature [age , insuranceCharge , No .children] and use 500 epoch for forward and backward propagation .
the splitting
the activation function that used is sigmoid function .



The Confusion Matrix for ANN :



True Negative : 359 : The model correctly predicted

False Positive : 49 : The model incorrectly predicted

False Negative: 25 : The model incorrectly predicted

True Positive : 160 : The model correctly predicted

Performance :

Metric	Value
Accuracy	87.3%
Precision	76.1%
Recall	86.4%
F1 Score	81.1%
ROC AUC Score	87.1%

Comparative Analysis of Machine Learning Models

Model	K-Value	TN	FP	FN	TP	Accuracy	Precision	Recall	F1 Score	ROC
K-Nearest Neighbors	**3**	**389**	**19**	**29**	**156**	**92%**	**89.6%**	**84%**	**87%**	**95%**
K-Nearest Neighbors	**5**	**381**	**27**	**25**	**160**	**91%**	**86%**	**86%**	**86%**	**96%**
K-Nearest Neighbors	**7**	**381**	**27**	**25**	**160**	**91%**	**85%**	**86%**	**86%**	**96.6%**
Naive Bayes	**N/A**	**367**	**41**	**67**	**118**	**82%**	**74%**	**64%**	**69%**	**92%**
Decision Tree	**N/A**	**378**	**30**	**34**	**151**	**89.2%**	**89.1%**	**89.2%**	**89.1%**	**87%**
Artificial Neural Network (ANN)	**N/A**	**359**	**49**	**25**	**160**	**87.3%**	**76.1%**	**86.4%**	**81.1%**	**87.1%**

Explanation of the Table:

Model: This is the type of algorithm used to predict if the person is a smoker or not smoker .

K-Value: For KNN models, this is the number of nearest neighbors considered. It doesn't apply to Naive Bayes or Decision Tree, so it's marked as N/A (not applicable) for those.

Accuracy: How often the model is right.

Precision: When it says someone has diabetes, how often it's correct. (TP / (TP + FP)).

Recall: How many of the actual diabetes cases it catches. (TP / (TP + FN)).

F1 Score: A mix of precision and recall into one number.

ROC AUC Score: How well the model distinguishes between having diabetes and

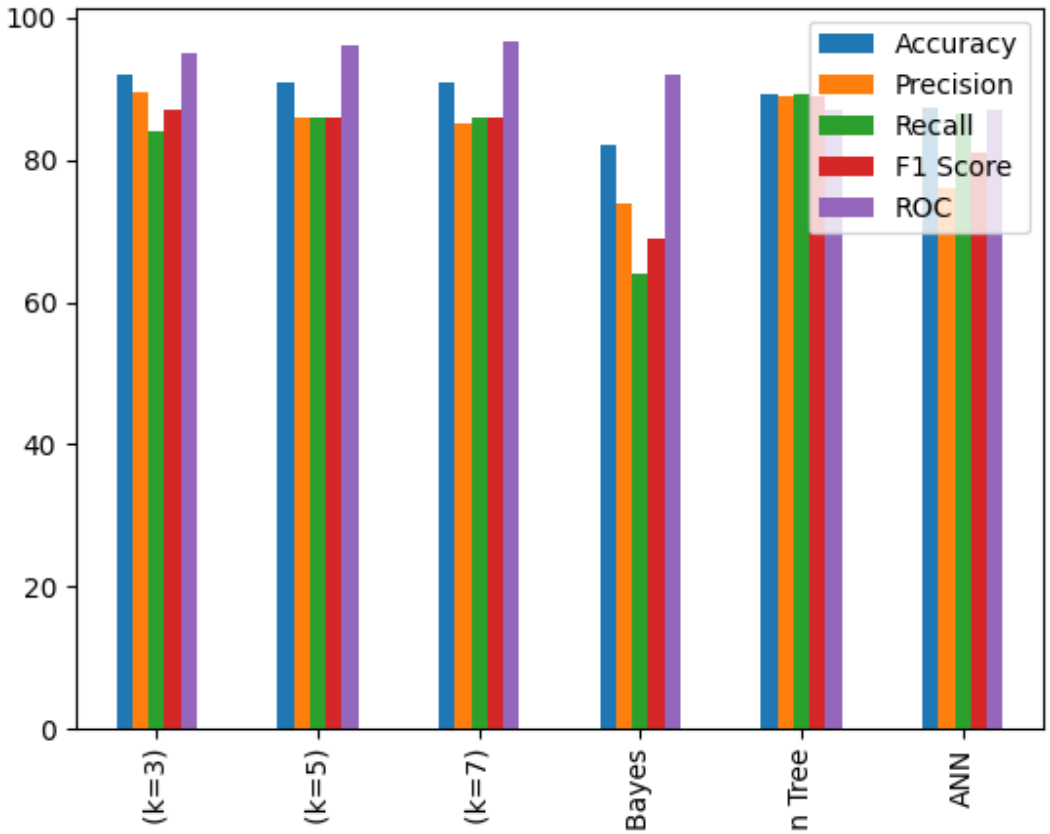
not.

Confusion Matrix: Shows the counts of true negatives, false positives, false

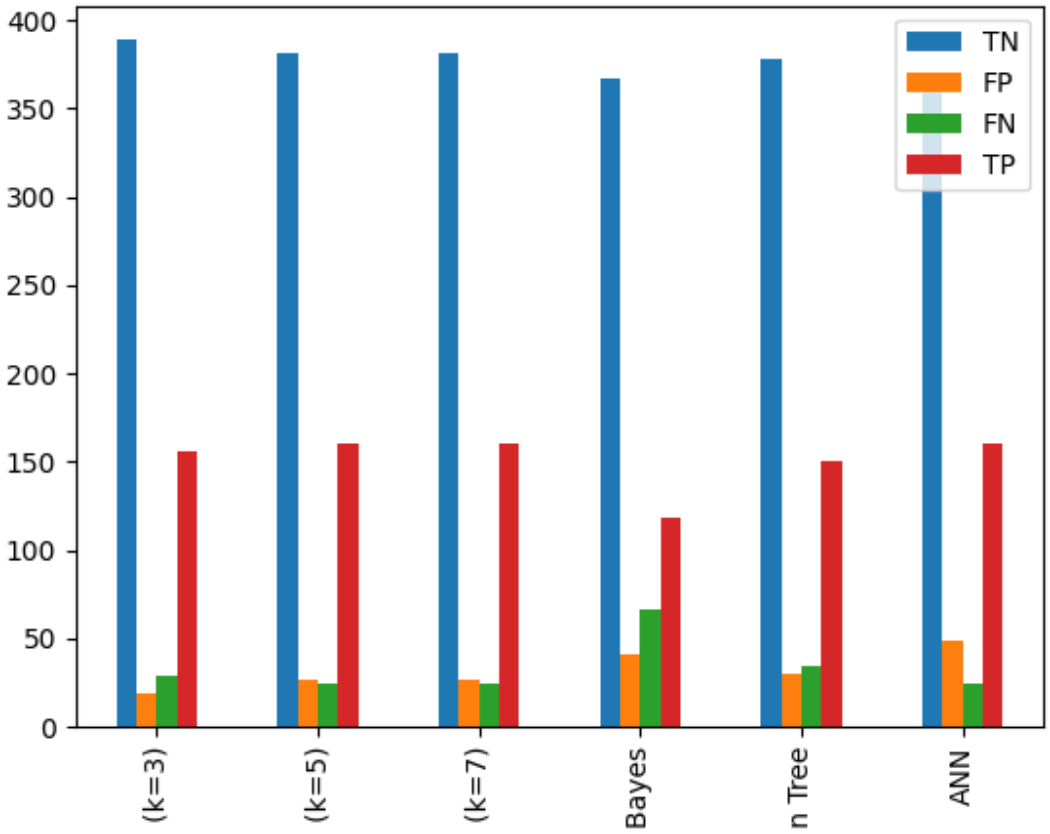
negatives, and true positives.

For visualization, I will create a set of charts to compare these models across different metrics. We'll have:

A bar chart for Accuracy, Precision, Recall, and F1 Score for each model.



A bar chart for TN, FP, FN, and TP for each model



Explain the Our Results :

(KNN):

when K=3 the model was correct 92% in generalization . It was highly reliable with percentage 89.6% such that predicted that person was a smoker , and was detected in 84% of all actual smokers (I mean about recall) T and the F1 score is 87%, which means it balances precision and recall well. It is very good at classification between smokers and not smokers with an ROC of 95%.

Naive byes : This model uses the probability of events to predict whether someone is a smoker or not, without considering the relationship between features. The accuracy for it , is a minimum values with another models 82% . but has an very good precision 89.1% when its predict is the person is smoker or not . but however the KNN models it's better than the current model .

Desecion tree :

This model makes decisions based on the rules it learns from features It is completely accurate 89.2% and reliable with percentage 89.1% when it predicts that someone is a smoker. It is also good at detecting actual cases with percentage 89.2% . It was also concluded that this model is better than the previous model based on TP and FN values is larger than the naive bayes . but the KNN its better than .

Artificial Neural Network (ANN):

The accuracy is lower with percentage 87.3% compared to KNN , and Decision Tree. It is less reliable when it predicts that person is a smoker 76.1% precision . but the native byes is better than this model based on recall only . i.e. the proportion of actual positive cases that are correctly detected by the model . but the KNN better than the all previous models .

There fore , The best model is KNN with k=3. It has the highest accuracy , meaning it best balances smokers and not smokers. It also has the second highest F1 score and a very high ROC AUC score, which means it is good at classifying between smokers and non-smokers.

The second best model is the KNN model with k = 5 and k = 7. Both have the same accuracy, F1 score, and recall. But k=5 has a small higher ROC AUC score . so it is a little better at classifying between smokers and not smoker.

References :

Feature Engineering: Scaling, Normalization, and Standardization - GeeksforGeeks