Assignment 2

In this assignment, you will be working with the stroke dataset that you have created. Your task is to build and evaluate four classification models:

Logistic Regression ,Naive Bayes, K Nearest Neighbor Classifier, Support Vector Machine . You will also be evaluating the performance of these models using accuracy, f-score, and the AUC-ROC curve.

Here are the detailed steps that you need to follow:
Step 1: Load the stroke dataset into your preferred programming language.
Split the dataset into training and testing sets. This dataset is imbalanced, solve this problem by using SMOTE.

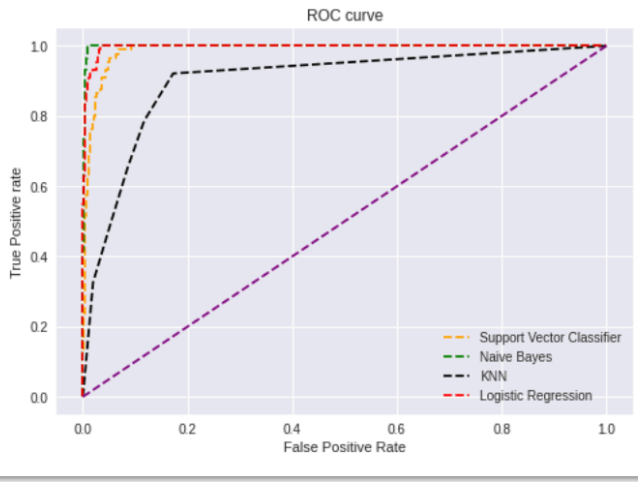
Step 2: Model Building
If you have already standardize your data. (REMEMBER you cannot feed in data to any algorithm that has different scales.)

Build a Logistic Regression model using the training set.
Build a Naive Bayes model using the training set.
Build a K Nearest Neighbor Classifier model using the training set.
Build a Support Vector Machine Classifier model using the training set.

Step 3: Model Evaluation
Evaluate the performance of both models on the testing set using accuracy, f-score, and the AUC-ROC curve.

Interpret the results of the evaluation and compare the performance of the four different models.

Step 3.5 : AUC-ROC Curve.
The evaluation of multiple models against each other is a critical stage. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease. Compute the Receiver operating characteristic for the models. Determine which model is curve is higher. This can help us determine which model did a better job in classification as well.



In summary, this assignment will provide you with an opportunity to apply your knowledge of Logistic Regression , Naive Bayes , K-Nearest Neighbor Classifier , and Support Vector Machine models to a real-world dataset. You will also gain experience in evaluating the performance of these models using various metrics such as accuracy, f-score, and the AUC-ROC curve. This will help you to develop a deeper understanding of the strengths and limitations of different classification models, and how to choose the most appropriate one for a given problem.

Sources that will help you will this project.
https://realpython.com/logistic-regression-python/
https://scikit-learn.org/stable/modules/model_evaluation.html
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html
https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/
https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html