Verzeo Internship - Major Project - ML June Batch

NAME: Gautam Kumar
SEM: 6th Sem
Dataset: Pima Indian Diabetes

Project Explanation:

To perform classification analysis on Pima Indian Diabetes dataset. Pima Indian Diabetes dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
Classification algorithms used are Naive Bayes classifier and Support Vector Machine classifier.

Tools Used:

- **Jupyter Notebook:** The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.
- **Pandas:** Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.
- **Sklearn:** Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Algorithms:

- **Naive Bayes Classifier:** Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
  The problem statement is to classify patients as diabetic or non-diabetic. The datasets had several different medical predictor features and a target that is 'Outcome'. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

```
from sklearn.naive_bayes import GaussianNB
from sklearn import metrics

colnames = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
pima_df = pd.read_csv("pima-indians-diabetes.data", names= colnames)

X = data.drop(Outcome", axis = 1)
Y = data[ ["outcome"] ]
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size= 0.2, random_state = 1)

model = GaussianNB()
model.fit(X_train, Y_train)
Y_pred = model.predict(X_test)
```

```
In [10]: from sklearn.naive_bayes import GaussianNB

In [11]: nb_model = GaussianNB()

In [12]: nb_model.fit(X_train,y_train)
Out[12]: GaussianNB()

In [13]: predicted_nb_model = nb_model.predict(X_test)
```

- **Support Vector Machine Classifier:** Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.
  The first model we fit on the training data is the Support Vector Machine (SVM). SVM uses many kernels to classify the data. We use rbf/Gaussian kernel to fit the first model.

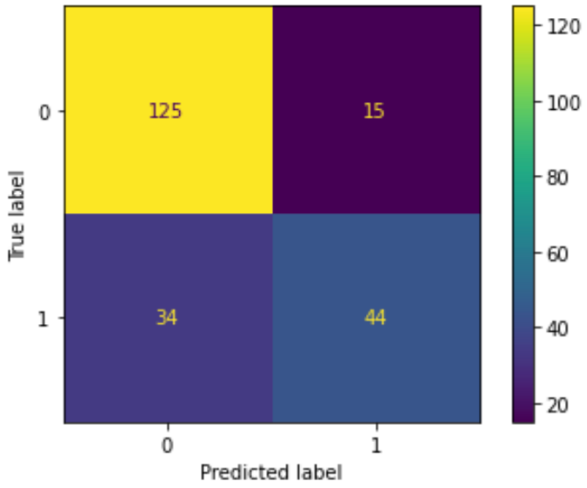
```
from sklearn.svm import SVC

classifier_rbf = SVC(kernel = 'rbf')
classifier_rbf.fit(x_train, y_train)

y_pred = classifier_rbf.predict(x_test)

print('Accuracy of SVC (RBF) classifier on test set:
{:.2f}'.format(classifier_rbf.score(x_test, y_test)))
Out[76]: Accuracy of SVC (RBF) classifier on test set: 0.75

print(f1_score(y_test, y_pred, average="macro"))
print(precision_score(y_test, y_pred, average="macro"))
print(recall_score(y_test, y_pred, average="macro"))
0.7431080565101182
0.7410256410256411
0.7481366459627329
```

```
In [19]: from sklearn.svm import SVC

In [20]: svc = SVC(random_state=0, kernel='rbf')

In [21]: svc.fit(X_train, y_train)
Out[21]: SVC(random_state=0)

In [22]: svm_y_pred = svc.predict(X_test)
```

Conclusion:

- **Naive Bayes Classification:**

```
Confusion Matrix:
 [[125  15]
 [ 34  44]]
Accuracy:
 0.7752293577981652
Average Precision =  0.5766499607270872
```

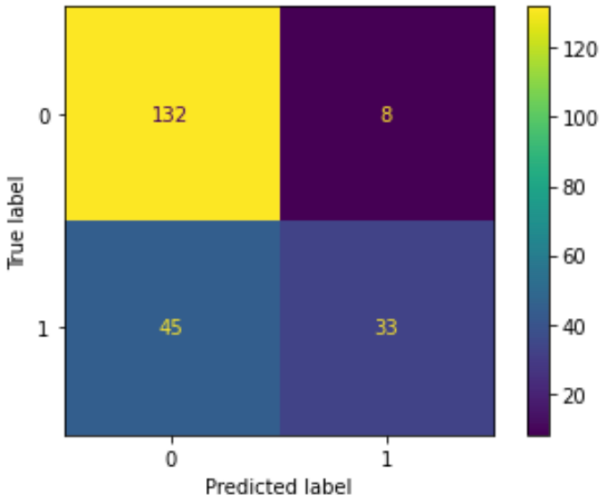


	precision	recall	f1-score	support
0	0.79	0.89	0.84	140
1	0.75	0.56	0.64	78
accuracy			0.78	218
macro avg	0.77	0.73	0.74	218
weighted avg	0.77	0.78	0.77	218

According to confusion matrix of Naive Bayes classifier, there are 44 who has Diabetes, 125 who does not have Diabetes and 49 which are misclassified.

- **Support Vector Machine Classification:**

```
Confusion Matrix:
 [[132   8]
 [ 45  33]]
Accuracy:
 0.7568807339449541
Average Precision =  0.5469473466788302
```



	precision	recall	f1-score	support
0	0.75	0.94	0.83	140
1	0.80	0.42	0.55	78
accuracy			0.76	218
macro avg	0.78	0.68	0.69	218
weighted avg	0.77	0.76	0.73	218

According to confusion matrix of Support Vector Machine classifier, there are 33 who has Diabetes, 132 who does not have Diabetes and 53 which are misclassified.