DS ASSIGNMENT-4

SUBMITTED TO:
PROF. MANDAR KARYAKARTE

VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

COMPUTER ENGINEERING DEPARTMENT

BY:
NAME: ABHISHEK MORE

G.R No.: 21810033

ROLL NO.: 323036

CLASS: T.Y COMP

BATCH: COMP-C2

Assignment-4

Description: Compare different classification models with respect to feature selection and accuracy. Infer the result: which model best suit for the dataset chosen.

Models I built:

1. Naive Bayes(NB)
2. KNeighbors Classifier(KNN)
3. Decision Tree Classifier
4. Linear Support Vector Classifier(LinearSVC)
5. Logistic Regression

1. Naïve Bayes: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

2. KNN: K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors.

3. Decision Tree Classifier: A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

4. LinearSVC: The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is.

5. Logistic Regression: Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by

estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution.

Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y).

Interpretation: Covid 19 Dataset: It tells the no. cases in the specific country , recovered cases, deaths, new cases, new deaths .

Outcomes: Prediction of diabetes depending on the blood pressure and glucose level.

Accuracy Score:

1. Naïve Bayes: 15.56%
2. KNN: 20.58%
3. Decision Tree Classifier: 36.63%
4. Linear SVC: 22.20%
5. Logistic Regression: 24.64%