IN TIME TEC

INTERNSHIP REPORT

PREDICTING COVID TEST RESULTS

Under the guidance of: Mr. Rakesh P

Submitted by: Ronit Jain

Date of Submission: 1 July 2021

Abstract:

DATA ANALYSIS: -

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

ABOUT DATASET: -

The dataset contains covid test results for many patients with several other tests results like Hemoglobin, Platelets, Red blood cells, etc.

Since a large number of tests are required to identify the positivity of COVID-19 in a person, so this the place where machine learning and data mining techniques comes in to identify the best way to predict the presence of COVID-19 in a person.

DATA MODELLING AND MACHINE LEARNING ALGORITHMS: -

Many machine learning algorithms like logistic regression, k-Nearest Neighbour, Decision tree, and Support Vector Machine classification algorithms are used to prepare the models and then conduct the predictions.

Then confusion matrix is built and test accuracy of each model is calculated.

The model with best accuracy is the best model to predict covid test results.

Problem:

For the given dataset of covid test come up with the following analysis:

1) Analyze the given dataset and create a subset of accurate and valid data points.
2) Create a model using the same data with the following:
    a) KNN classifier to classify the presence of covid patients and create a confusion matrix to validate the model accuracy.
    b) SVM classifier to classify the presence of covid patients and create a confusion matrix to validate the model accuracy.
    c) Decision Tree classifier to classify the presence of covid patients and create a confusion matrix to validate the model accuracy.
3) Choose the best model based on the accuracy.

Solution:

DATA CLEANING: -

Firstly, unnecessary data columns are removed, which are of no use.

Then by taking an overview of the data, it is determined that more than 85% of data is missing. So, data cleaning is done by dropping out some rows and columns according to missing data percentages.

Then data transformation is done by converting categorical data into numerical data like “positive, negative” is converted into “1,0”.

HANDLING MISSING DATA: -

After reducing the number of rows and columns, there is still some missing data in the dataset. It is taken care of by data imputation, which includes mean, median and mode method, regression method, etc.

Missing discrete data is filled using the mode, and missing continuous data is filled using linear regression.

FINDING RELATION B/W VARIABLES: -

Relation between target variable and independent variables is found using Correlation. With the help of correlation matrix, relevant variables are selected.

Then using correlation heatmap, independent variables with high correlation are taken care of.

CREATING MODEL: -

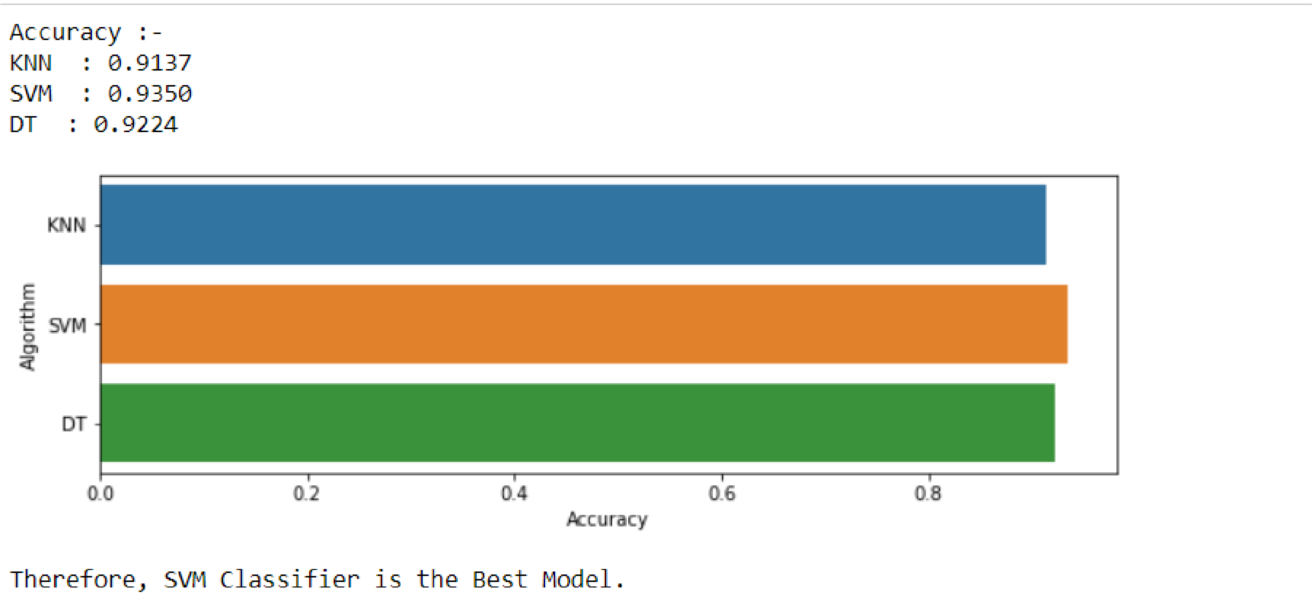
Firstly, the data is split into training and testing data then different classifier models are build using different algorithms: k-nearest neighbour, support vector machine and decision tree.

1. K-Nearest Neighbor (KNN) Algorithm is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and puts the new case into the most similar category to the available categories.
2. Support vector machine (SVM) Algorithm, in this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features), with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.
3. Decision Tree Algorithm, in this algorithm, for predicting a class label for a record, we start from the tree’s root. We compare the values of the root attribute with the record’s attribute. Based on the comparison, we follow the branch corresponding to that value and jump to the next node and this process is continued.

ACCURACY AND CONFUSION MATRIX: -

Finally, using accuracy and confusion matrices of all models, the best model is determined.

OUTPUT AND CONCLUSION: -



```
Accuracy :-
KNN  : 0.9137
SVM  : 0.9350
DT  : 0.9224
```

```
Therefore, SVM Classifier is the Best Model.
```