LAB ASSIGNMENT 7
Lab Report

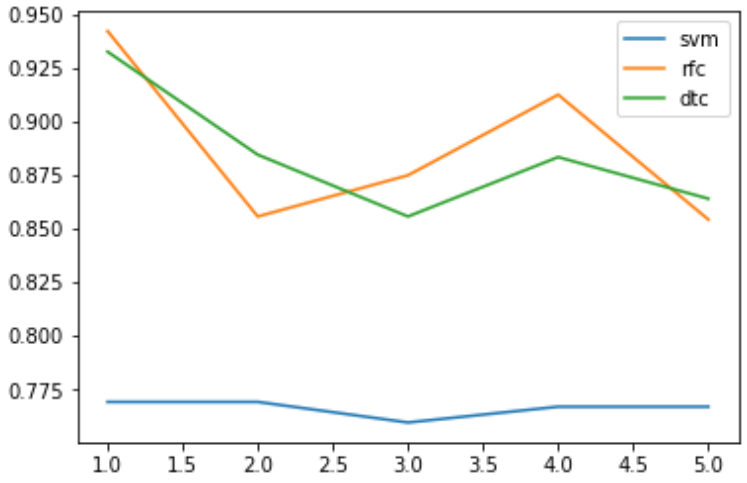
Shubh Goyal (B21CS073)

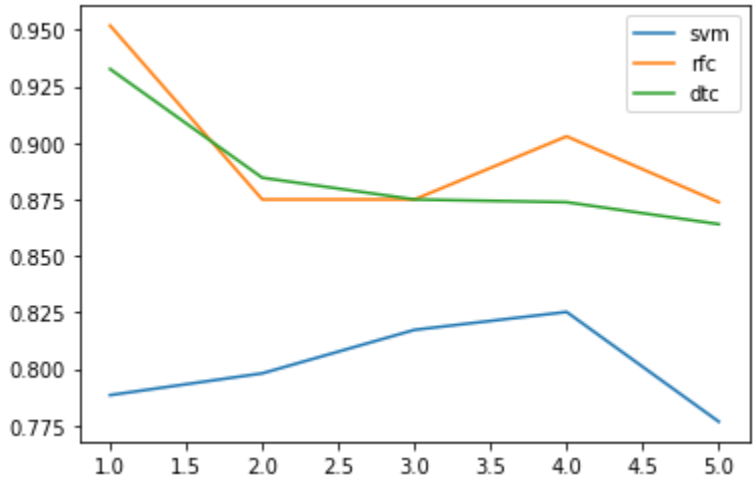
QUESTION 1-

Task 1 and 2: It was found that except 10 features, all other features had more than 50% values missing from the dataset and one feature had more than 200 values missing. Those columns were removed from the dataset.
For the remaining columns, ‘steel’ had missing values which were replaced by the mode of the column. Also, one hot encoding was applied on ‘product-type’, ‘steel’, ‘shape’ and ‘bore’. There was no specific reason for this rather categorical encoding could also have been done.

Task 3: SVM Classifiers, Decision Tree Classifiers and Random Forest Classifiers were trained along with five fold cross validation on both original and standard data.

The following graphs represent the accuracies of each validation in all three classifiers for original and standardized data respectively:





Also, accuracies were obtained on the test sets and the result are as follows:

	Original Data	**Standardized Data**
SVM	75.35714285714286	80.71428571428572
Random Forest	88.57142857142857	88.57142857142857
Decision Tree	87.14285714285714	86.78571428571429

The following observations were made:

1. From the graphs, it can be concluded that Random Forest Classifiers performed around 5-10% better than SVM's in both the cases. Also the performance of Random Classifiers were better than Decision Trees which was expected.
2. There is no significant difference between accuracies of Decision Trees and Random Forest Classifiers and thus we can also go with Decision Trees rather than Random Forests in case of time trade-off.
3. There is a significant improvement in performance of SVMs when using standardized data.

SVMs were used due to better performance in high dimensional space and its memory efficiency. Decision Trees were used because the data seemed to have some linearly separable boundaries. Also, Random Forest Classifiers were used to see if any improvements can be made to the performance of SVMs.

Task 4: A class named 'PCA' was made for scratch implementation of Principal Component Analysis. It takes 'n_components' as an attribute which is the number of principal components needed to be given as output. It fulfills the task using the following method:

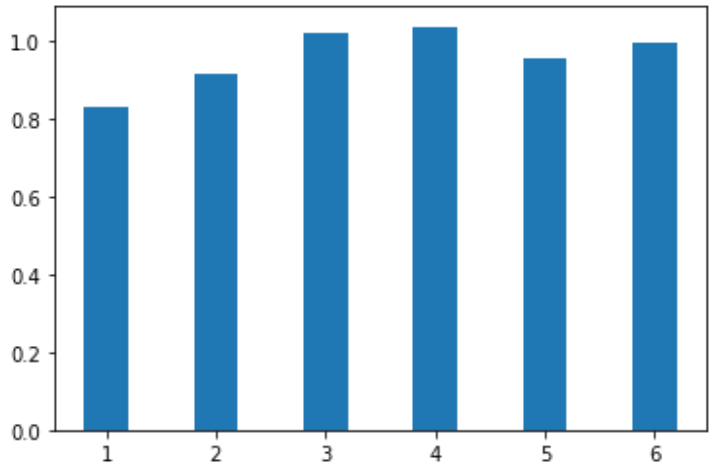
1. fit(): It takes as input the feature vectors in the form of panda framework or numpy matrix and gives as output a matrix of principal components.

4 helper functions were implemented in the class:

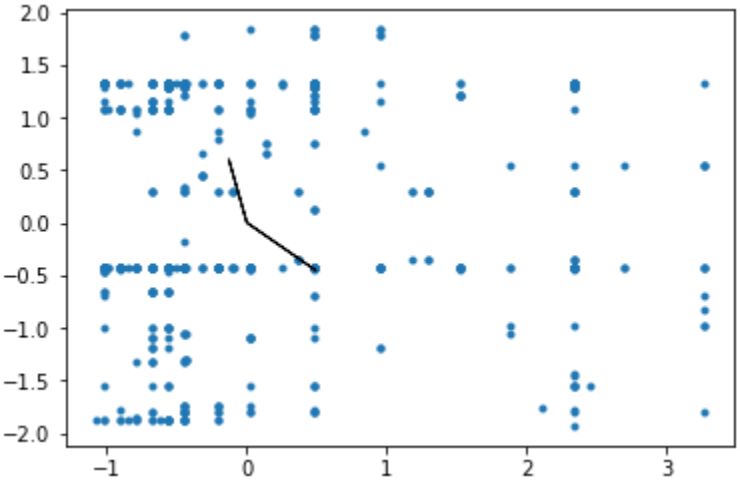
1. centralize(): It normalizes the given feature vectors to be transformed in the PCA.

2. calc_covar(): It calculates the covariance matrix of the normalized feature vector from scratch.
3. eigen(): It gives the eigenvalues and eigenvectors of the covariance matrix. Eigenvalues are being calculated from scratch using an orthogonal matrix. The dot product of the matrix, the covariance matrix and the transpose of the matrix is being taken and the diagonal elements of the obtained matrix represent the eigenvalues. Eigenvectors were calculated using the eig() function of numpy.
4. prin_comp(): It is used to make the principal component features to be returned by the class.

Task 5: The following is the bar plot of variance as we increase the number of principal components:



A scatter plot along with eigenvalues was made taking ‘thick’ at x-axis and ‘width’ at y-axis. The following results were obtained:



Task 6: The same models were trained on the PCA data in this task.

Task 7: The following accuracy and f1 scores were obtained for the three classifiers on the original data and the pca data (with n_components=1):

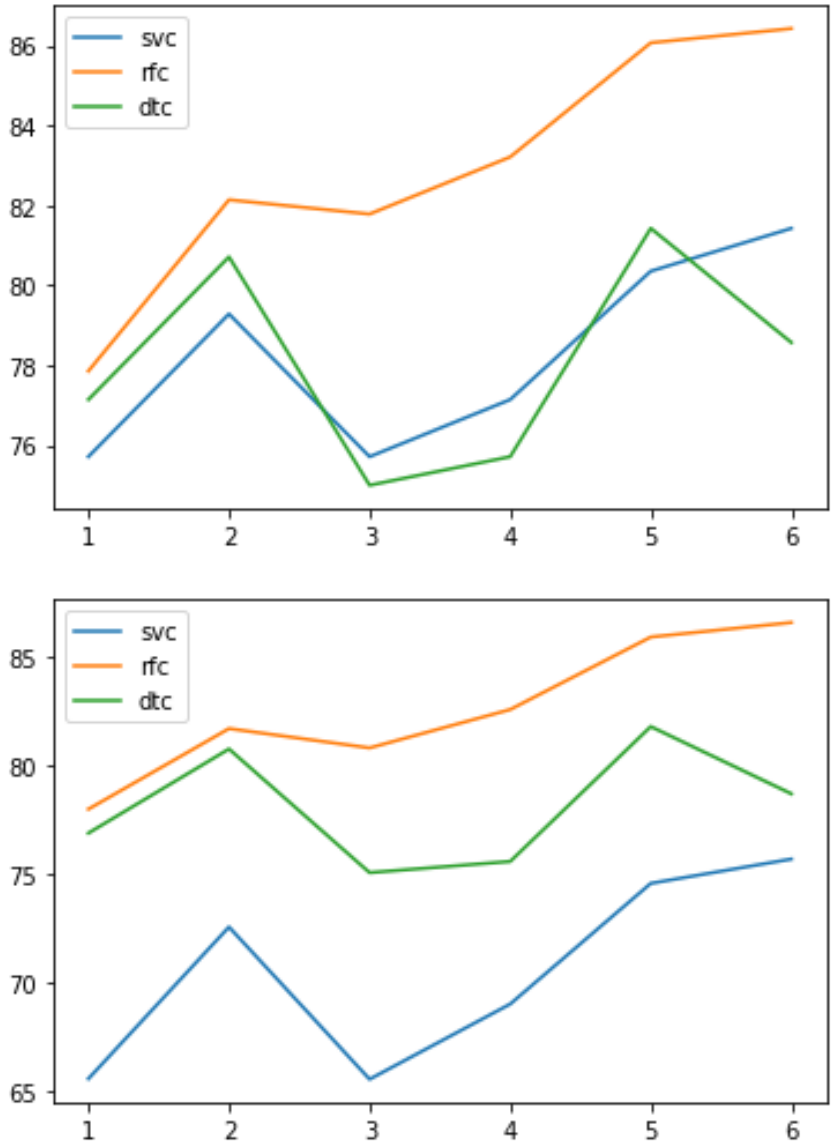
ACCURACY SCORE		
	Original Data	**PCA Data**
SVM	75.35714285714286	76.07142857142857
Random Forest	87.85714285714286	76.42857142857142
Decision Tree	86.07142857142858	74.64285714285714

F1 SCORE		
	Original Data	**PCA Data**
SVM	64.76723887110852	68.04722304722304
Random Forest	87.62776148094356	75.91995620149758
Decision Tree	86.61474469305796	74.06485923725099

It can be observed that the scores decreased for Random Forest and Decision Tree in case of PCA data which can be due to loss of information in reduction of dimensions. Also, we can see that the scores of SVMs increased for PCA due to feature reduction.

Task 8: As observed from the scatter plot between ‘thick’ and ‘width’ made in Task 6, there is no change in the distribution of data other than the reduction in magnitude of the data.

The graphs for accuracy and f1 scores for all the classifiers were plotted taking different number of principal components and the following results were obtained:



From the following plots, the optimal number of principal components observed is 5.