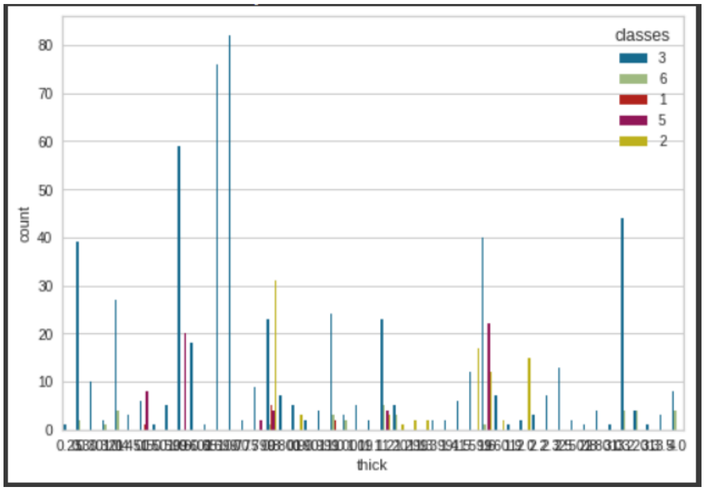
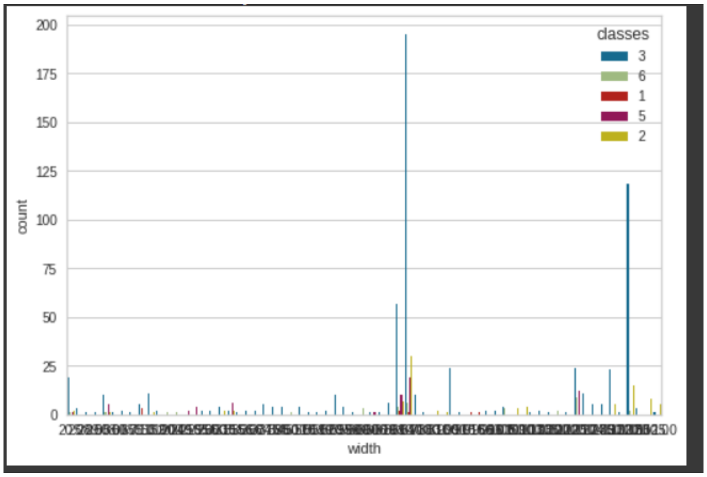
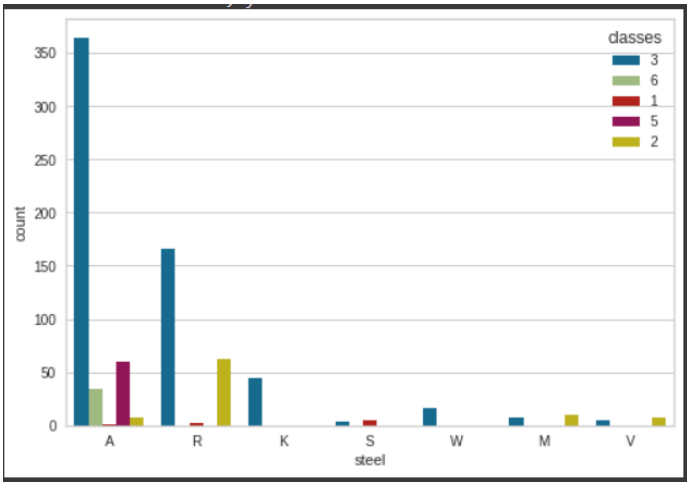
LAB Assignment 7
CSL 2050

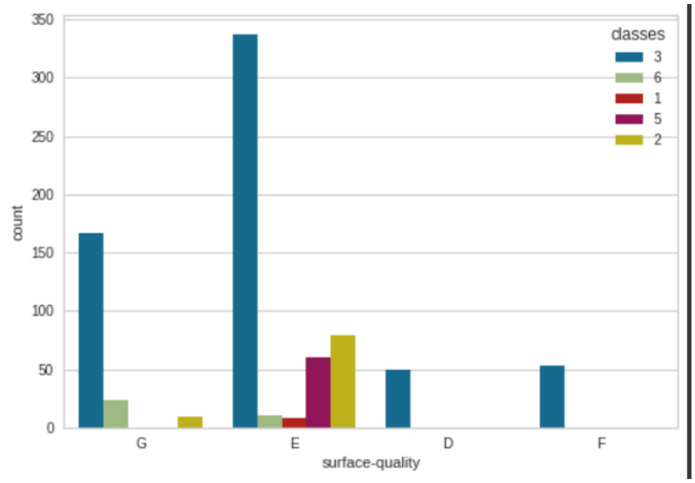
By :- Akshat Jain B21CS005

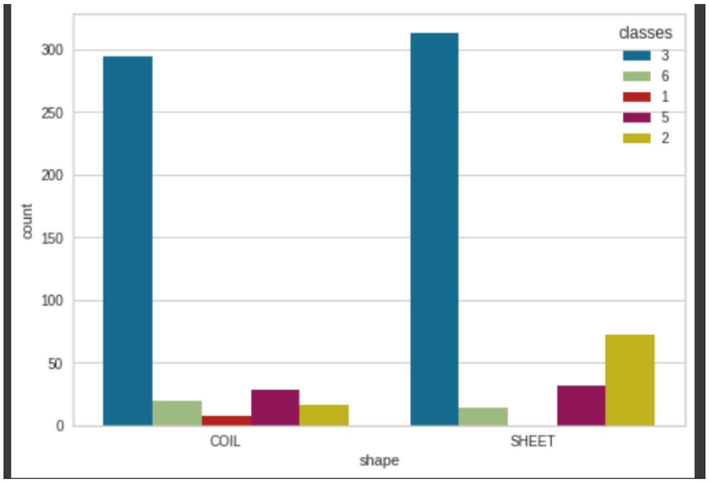
1. We work with the Annealing dataset here
   1.1. Combined with Part 1.2
   1.2. We first load the dataset. Then we systematically analyze the dataset to see which columns are actually usable as features for our problem. We then drop a majority of the columns due to a large no. of NaN values in them(roughly 25% or more) we are then left with 2 continuous features and 3 categorical features. We then visualize the data w.r.t classes already given in the data to get a feel for the distribution.







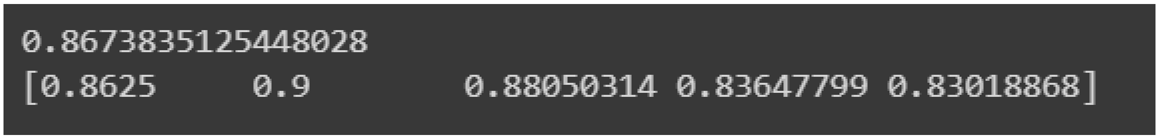




We then perform categorical encoding on the categorical features. After that we create a deepcopy of the entire dataset in which we then perform standard scaling on all the continuous features.

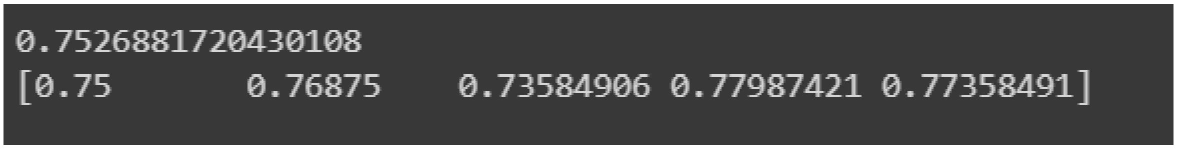
1.3. We then Train a Decision tree and an svm classification model. We perform 5- fold cross validation for both the models and check the accuracy for each split on the 2 different datasets.

For DTC and normal data:



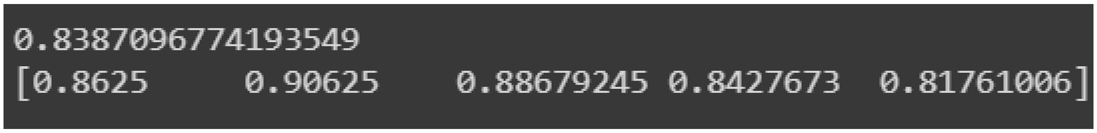
```
0.8673835125448028
[0.8625     0.9        0.88050314 0.83647799 0.83018868]
```

For SVC and normal data:



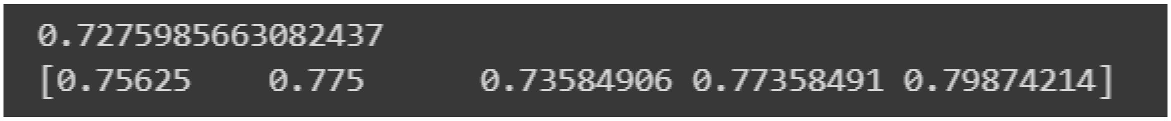
```
0.7526881720430108
[0.75       0.76875    0.73584906 0.77987421 0.77358491]
```

For DTC and standarized data:



```
0.8387096774193549
[0.8625     0.90625    0.88679245 0.8427673  0.81761006]
```

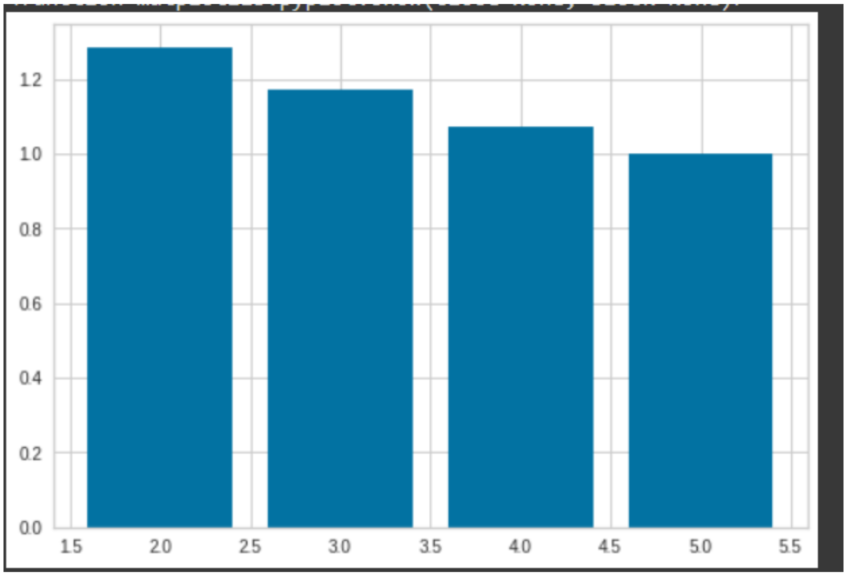
For SVC and standarized data:

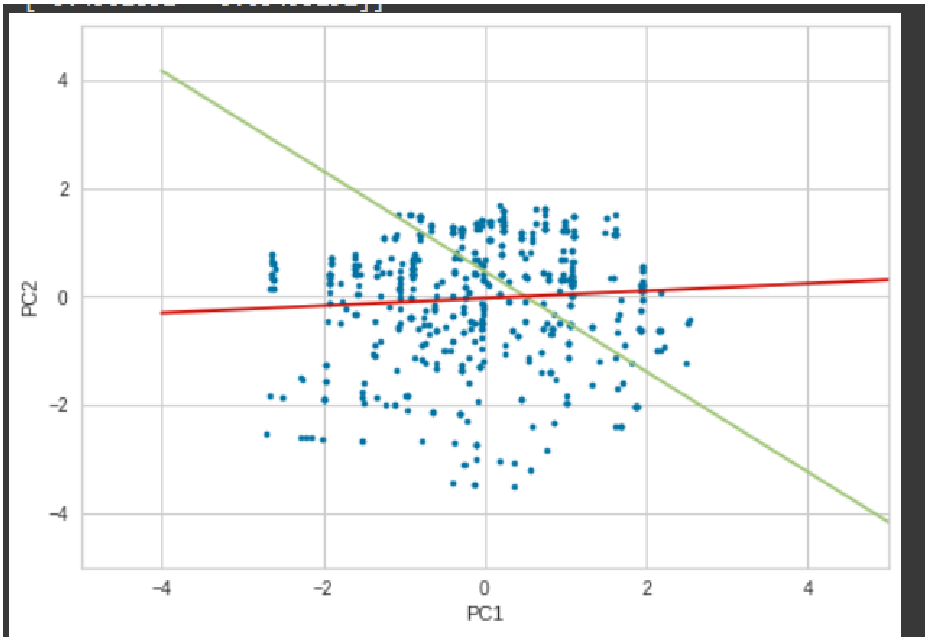


```
0.7275985663082437
[0.75625    0.775      0.73584906 0.77358491 0.79874214]
```

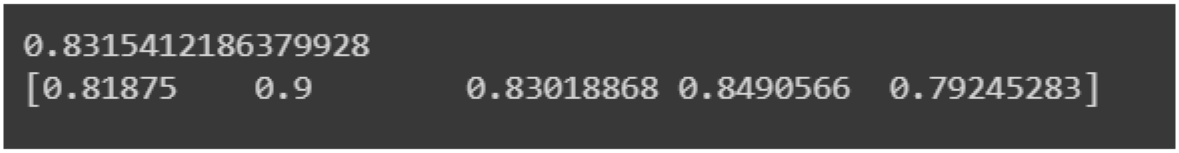
1.4. We First define a function to get the covariance matrix from scratch and then implement the PCA algorithm from scratch. The Features of this Algorithm are:- (1) Taking no of components as an input (2) Returns the relevant eigenvectors, cumulative variance and the transformed dataset (3) Centralizes the dataset first with the help of the Z value algorithm.

1.5. We plot the variance of the entire transformed dataset for each instance of increasing order of components. We also plot the transformed dataset for no. of Components = 2 along with the eigenvectors(the first 2 dimensions)



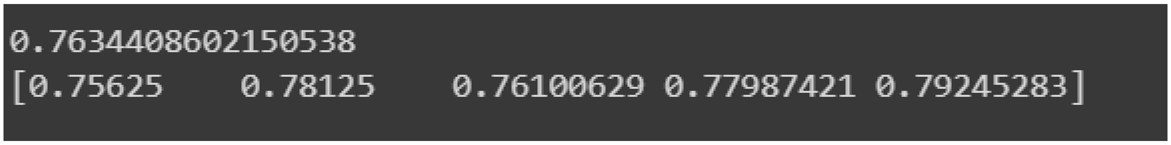


1.6. We then train the previous 2 classification models on the transformed dataset to compare their accuracies.
For DTC and transformed data:



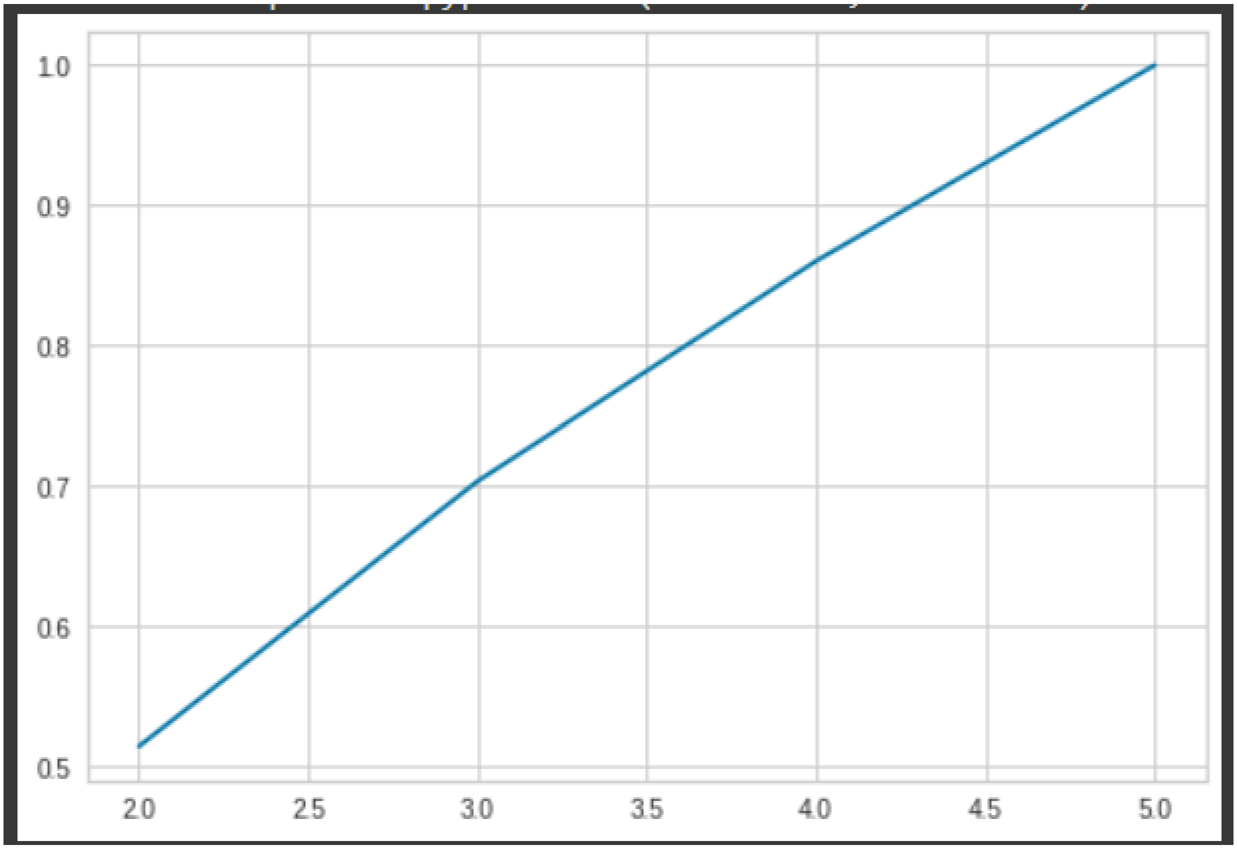
```
0.8315412186379928
[0.81875    0.9        0.83018868 0.8490566  0.79245283]
```

For SVC and transformed data:



```
0.7634408602150538
[0.75625    0.78125    0.76100629 0.77987421 0.79245283]
```

1.7. We plot the Cumulative variance for successive iterations of the PCA increasing the no. of components. We can see here that we can apply a basic threshold of 80% of the original variance for no. of components = 4. We will take this as our best case for this use instance.


1.0
0.9
0.8
0.7
0.6
0.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0