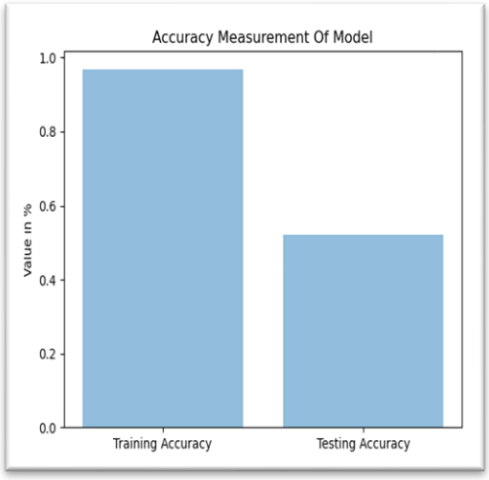
MACHINE LEARNING COURSE

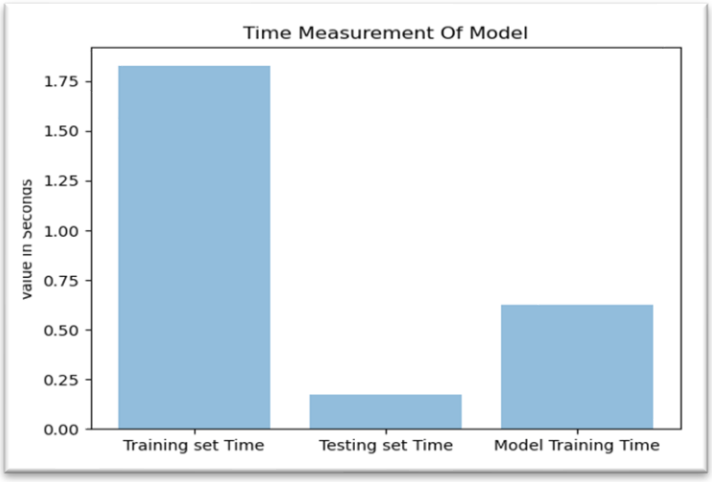
MILESTONE 2 | CLASSIFICATION

➢ Drawing Bars for Each Classification Model

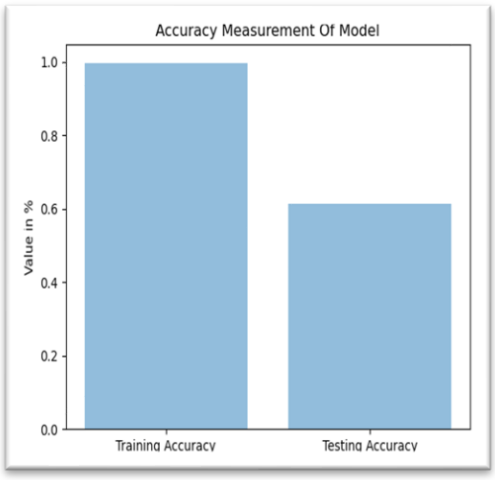
For bars I know that it requested only 3 bars, but I preferred to show Bar for each one e.g., accuracy in training, in testing and so on so the bar will be consisting of five columns as defined below as two for accuracy and three for time measurements.

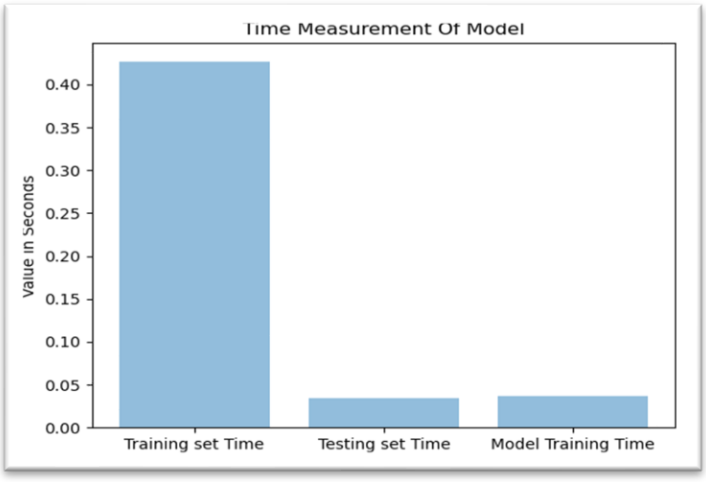
- The Bars for SVM Model [Support Vector Machine]



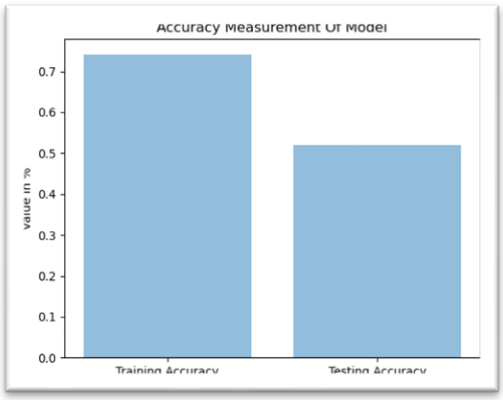


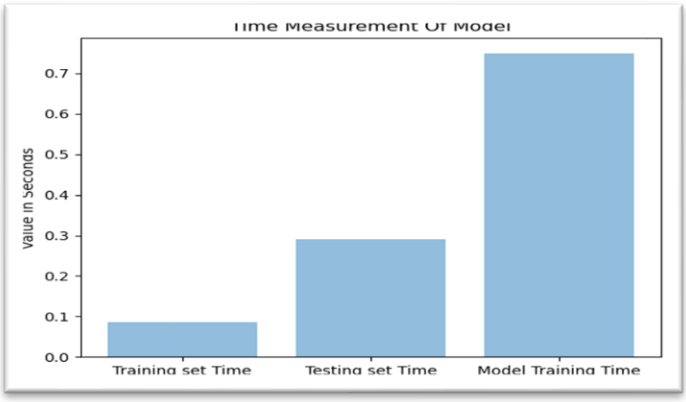
- The Bars for RFC Model [Random Forest Classifier]





- The Bars for KNN Model [K Nearest Neighbors]





➢ Feature Selection

Here we depend on correlation of each feature [after making Dummy variable] with the output variable according to specific percentage so the returned features considered the best to work with and that happen by using function TopKPrecentage() that return the features according to given percentage and graph the values of it for visualization as shown in the figure below.

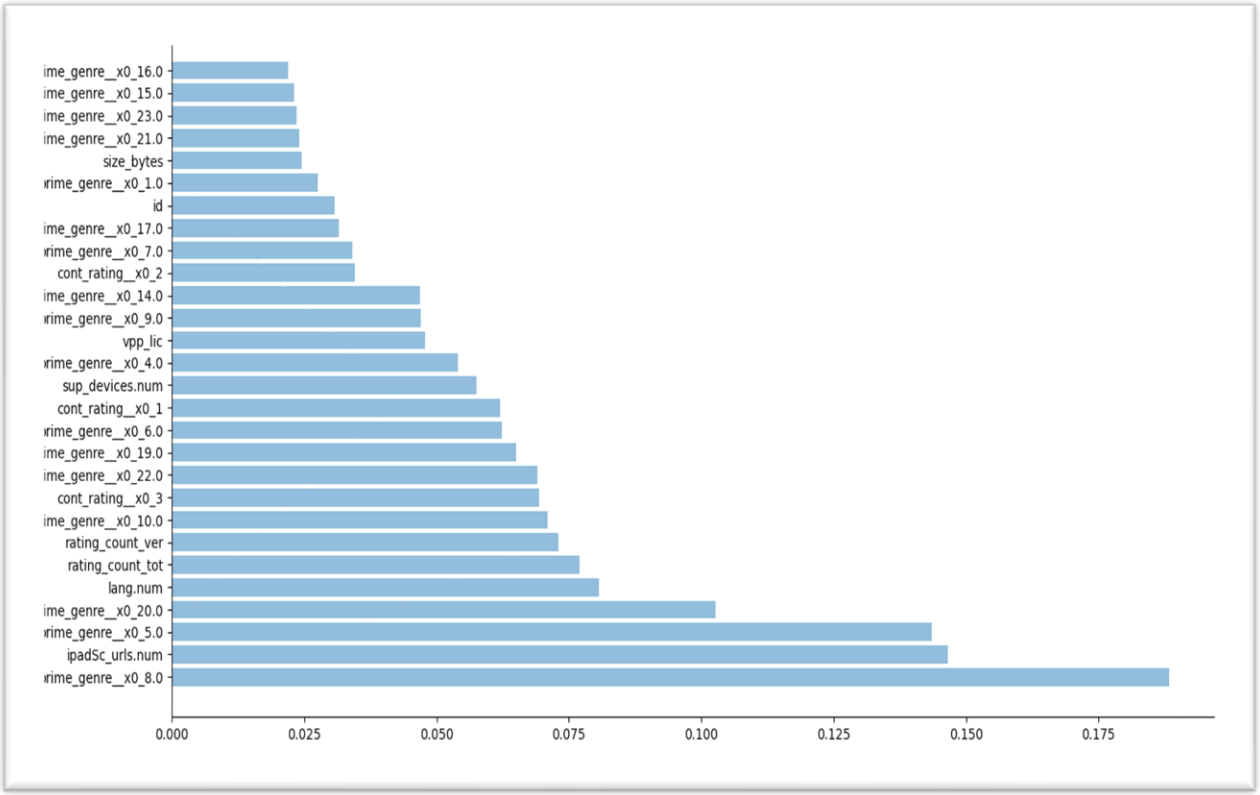
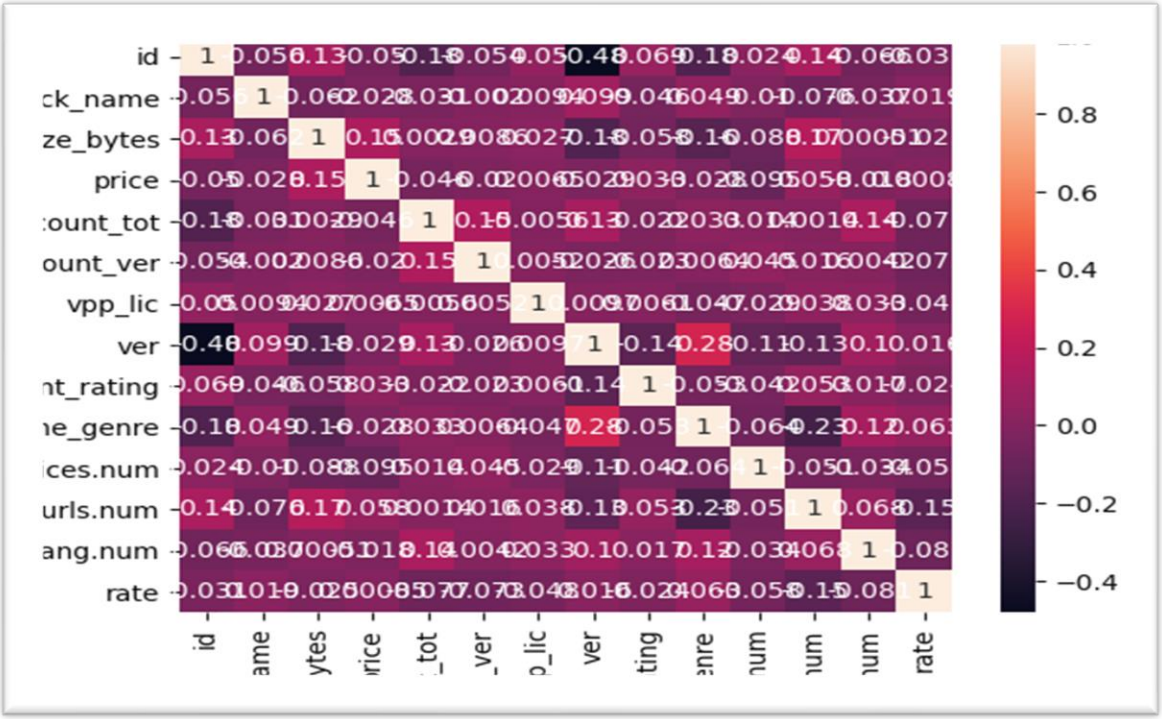


Figure to show features more than 0.02

Another thing I want to talk about which is dummy variable you see that there are more than one feature more than 0.1 while with given data there is only one feature is more than 0.1 as figured below so it's enough to say that the way we depend on for feature selection is trustful enough.



Heatmap correlation figure

- **Explain how hyperparameters effect model performance.**
  Here we will talk about the tuning parameters we will change and watch the change of performance on doing this.

- For SVM Model

Note	Mode	The Measurements	C=8 Gamma=0.9
Here We Work with Default Values of C and Gamma	Training	MSE	0.173
		Accuracy	0.857
	Testing	MSE	0.539
		Accuracy	0.561

Notes	Mode	The Measurements	C=8 Gamma=0.1	C=8 Gamma=20	C=8 Gamma=50
Here we will fix C and Change Gamma	Training	MSE	0.264	0.001	0.000
		Accuracy	0.780	0.999	0.999
	Testing	MSE	0.524	0.657	0.664
		Accuracy	0.560	0.512	0.515

Notes	Mode	The Measurements	C=1 Gamma=0.9	C=25 Gamma=0.9	C=50 Gamma=0.9
now we will fix Gamma and Change C	Training	MSE	0.179	0.014	0.009
		Accuracy	0.866	0.986	0.991
	Testing	MSE	0.584	0.613	0.166
		Accuracy	0.516	0.481	0.477

So, Conclusion from previous test cases we watch that when C or Gamma is too large as 25 or higher it causes high accuracy in training But, low accuracy in Testing and high value of MSE and that is overfitting while when C or Gamma it too small as 1 or low it leads not to train model very will which cause underfitting.

- For RFC Model

Note	Mode	The Measurements	Estimator=50 Max Depth=20
Here We Work with Default Values of Estimator and Max Depth	Training	MSE	0.007
		Accuracy	0.993
	Testing	MSE	0.521
		Accuracy	0.597

Notes	Mode	Measurements	Estimator=1 Depth=20	Estimator=25 Depth=20	Estimator=100 Depth=20
Now fixing Depth and Change Estimator	Training	MSE	0.237	0.006	0.0
		Accuracy	0.809	0.994	1.0
	Testing	MSE	0.698	0.485	0.483
		Accuracy	0.493	0.586	0.592

Notes	Mode	Measurements	Estimator=50 Depth=1	Estimator=50 Depth=10	Estimator=50 Depth=50
now fixing Estimator and Change Depth	Training	MSE	0.638	0.219	0.0
		Accuracy	0.529	0.814	1.0
	Testing	MSE	0.628	0.43	0.491
		Accuracy	0.516	0.612	0.581

So, Conclusion from previous test cases we watch that when we use very low value for estimator or max depth as 1 occurs very high MSE in Test and low prediction in training so it considers underfitting as data does not fit will in data in another hand when estimator or depth reach for a specific value it never changes again, or it will change with very small value and sometimes cause overfitting as in 100 in estimator and 50 in max depth.

- For KNN Model

Note	Mode	The Measurements	Neighbor=3 algorithm='auto'
Here We Work with Default Values of neighbors and algorithm	Training	MSE	0.347
		Accuracy	0.743
	Testing	MSE	0.544
		Accuracy	0.546

Notes	Mode	Measurements	neighbor=3 algorithm='ball_tree'	neighbor=3 algorithm='kd_tree'	neighbor=3 algorithm='brute'
Here we will fix neighbor and Change algorithm	Training	MSE	0.362	0.632	0.632
		Accuracy	0.739	0.739	0.739
	Testing	MSE	0.571	0.571	0.571
		Accuracy	0.535	0.535	0.535

Notes	Mode	Measurements	neighbor=1 algorithm='auto'	neighbor=9 algorithm='auto'	neighbor=19 algorithm='auto'
now we will fix algorithm and Change neighbor	Training	MSE	0.0	0.466	0.465
		Accuracy	1.0	0.625	0.605
	Testing	MSE	0.575	0.551	0.510
		Accuracy	0.525	0.533	0.558

So, Conclusion from previous test cases we watch that when fix neighbor and changing the type of algorithm MSE and accuracy almost the same while when changing value of neighbor and fix the algorithm type there is overfitting occur when using only one neighbor while in using more than 3 Neighbors accuracy starts being lower than before, so we see that it the worst model could fit the data.

➢ Conclusion

In Classification there is exist no model always solve a problem but, the used model depends on the type of this problem also when you get high accuracy it doesn't mean that your model is so good but it determined also by the accuracy of Test and how your model fit the data will so in our milestone 2 we do our best to get best accuracy and keep not falling in overfitting and underfitting so the stopping condition for us in accuracy of model that mean square error in testing not be more than the accuracy in testing. We try to do our best and that what we get ... Thanks.