Introduction to Data Science

Course Project

Report Document

<Eesha Tariq>

<21L-6269>

<Section 3A>

Instructions: Read These Carefully Before Starting

1. Due Date: Sunday 4th December 2022 – 11:59PM
2. Submission will be taken on Google Classroom
3. Submit only the following 2 files named like the following:
    a. Code File (Jupyter Notebook): L210000_Code.ipynb
    b. Report Document (This File): L210000_Report.pdf
4. Project will not be evaluated if:
    a. You submit python (.py) files
    b. You submit multiple .ipynb files
    c. You submit compressed (.rar or .zip) files
    d. You submit any files other than the required PDF and IPYNB
5. Upload data files directly to Google Colab - do not use Google Drive or GitHub linking method
6. All source files needed to complete this project are uploaded with it on Google Classroom.
7. Do not add the data file with your submission on Google Classroom.

Not following these instructions will lead to mark deduction.

Please try to use Microsoft Word instead of Google Docs to edit this document and to export it as a PDF file for final submission.

Happy Coding 😸

TA Emails

Section A, C - Muhammad Maarij l192347@lhr.nu.edu.pk

Section B, D - Hira Ijaz l192377@lhr.nu.edu.pk

For this project you will be applying machine learning models (both regression and classification) to the dataset which contains information about various individuals, their clothing, and its properties along with other atmospheric elements such as temperature, pressure humidity etc. The users also provided feedback on if they feel cold or not. The feedback (through AMV and PMV) which is based on the following mapping:

The following table shows the mapping of sensations:

Value	Thermal Sensation
+3	hot
+2	warm
+1	slightly warm
0	neutral
−1	slightly cool
−2	cool
−3	cold

The dataset is given in an excel file named CollectedData.xlsx, see sheet 2 of excel file. The dimension names (column headers) are not mentioned in the given file. The table below describes the columns which will be of your interest.

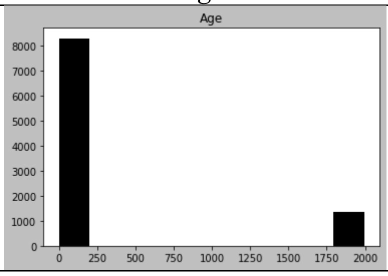
Column number	Feature Name	Feature Description
3	Age	Age
22	Clo	Clothing insulation
19	Met	Met Rate
26	Dewpt	Dewpt
27	PlaneRadTemp	plane radiant temperature
37	Ta	Average air temperature
38	Tmrt	Average mean radiant temperature
40	Vel	Air Velocity
42	AirTurb	Air Turbulance
43	Pa	Vapor Pressure
44	Rh	Humidity
74	TaOutdoor	Outdoor Air Temperature
77	RhOutdoor	Outdoor Humidity
8	AMV	Classification response variable
49	PMV	Regression response variable

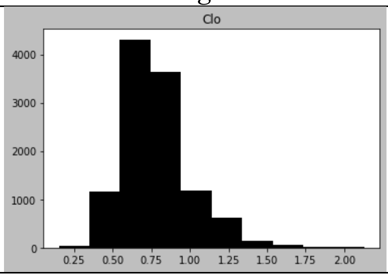
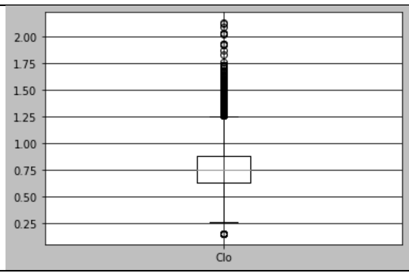
Part A. Preprocessing

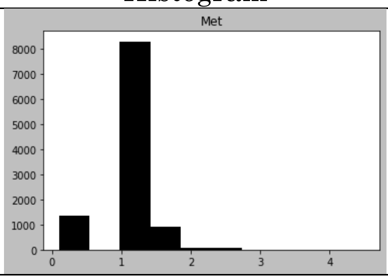
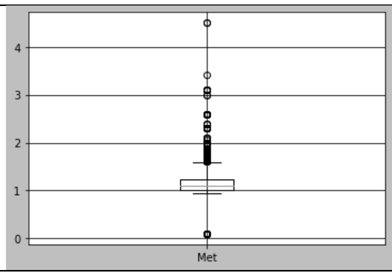
1. In this step, you are required to apply the preprocessing steps that you've covered in the course. Specifically, for each of the input dimension, fill in the following (add rows and complete the table for all input dimensions).

Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	STD
Age	Quantitative	9650	2916	1359	0.000	1996.0	24.0	308.637	35.0	462556.6	680.115105
Clo	Quantitative	11159	1407	373	0.150	2.13	0.77	0.77	0.7517	0.049284	0.221999
Met	Quantitative	10678	1888	1731	0.100	4.5	1.0	1.06591	1.1	0.18394	0.428882
Dewpt	Quantitative	9014	3552	0	-1.953	26.89675	17.4	13.621447	14.1	34.84593	5.903044
PlaneRadTemp	Quantitative	5544	7022	452	-7.420	11.7	0.3	0.217785	0.2	1.084022	1.041164
Ta	Quantitative	12545	21	539	15.960	31.0	23.2	23.179187	23.1367	2.053443	1.432984
Tmrt	Quantitative	8864	3702	343	16.610	37.445	22.5	23.450693	23.35	225.7473	15.02489
Vel	Quantitative	8865	3701	309	0.000	1.88	0.1	0.112445	0.1	0.006248	0.079044
AirTurb	Quantitative	6965	5601	2	0.000	102.45	0.5	18.26587	0.5	627.0571	25.041109
Pa	Quantitative	7910	4556	1352	0.000	27.7	2.1	5.123996	1.55	66.52255	8.156136
Rh	Quantitative	12530	36	0	7.400	79.3	64.0	42.528507	43.2768	226.848	15.061475
TaOutdoor	Quantitative	11197	1369	124	-24.900	32.35	27.55555556	17.175087	18.2	113.7511	10.665415
RhOutdoor	Quantitative	12546	20	1349	0.000	100.35	0.0	61.098939	68.7958	610.3056	24.704364
AMV	Qualitative	12510	56	0	-3.000	3.0	0.0	0.100584	0	1.21443	1.102012
PMV	Quantitative	11869	697	259	-4.170	2.5	0.1	-0.073711	-0.03	0.289471	0.538025

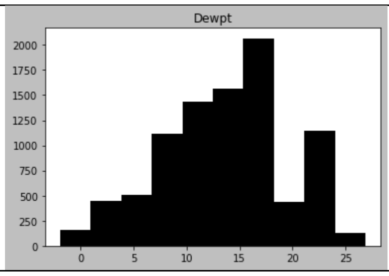
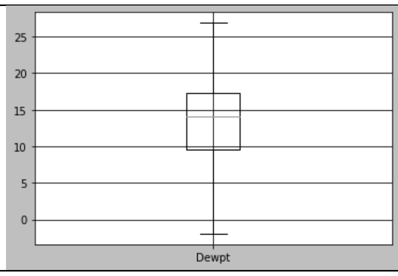
2. For each of the input dimension, plot histogram and comment the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimension, you're required to fill the following table (duplicate it for each of the 15 dimensions).

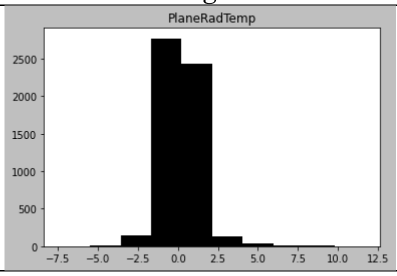
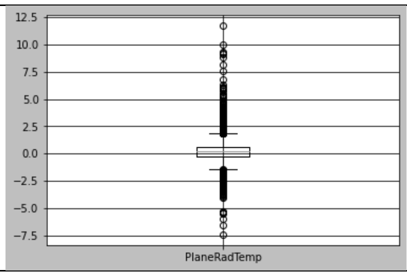
<table>
<tr><th colspan="2">Age</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments:</td><td>Comments: a couple of outliers</td></tr>
</table>

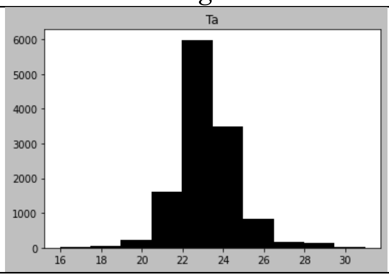
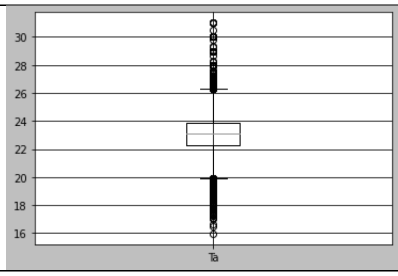
<table>
<tr><th colspan="2">Clo</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments:</td><td>Comments: literally everything is an outlier</td></tr>
</table>

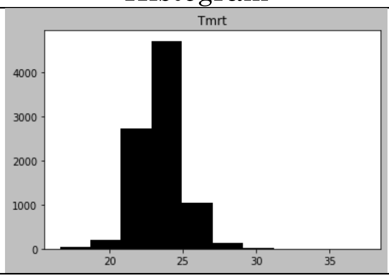
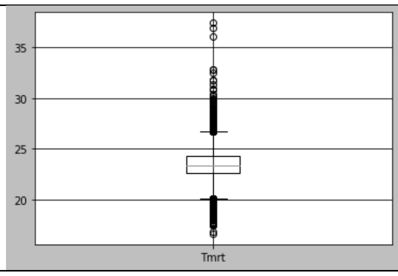
<table>
<tr><th colspan="2">Met</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments:</td><td>Comments: large amount of outliers</td></tr>
</table>

<table>
<tr><th colspan="2">Dewpt</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
</table>

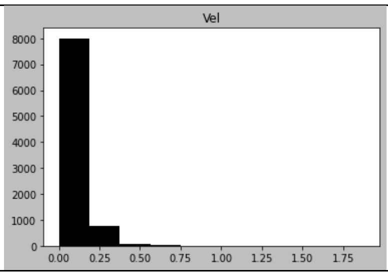
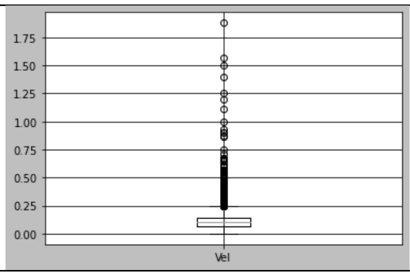
	
Comments:	Comments: no outliers

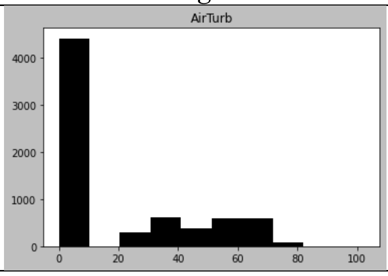
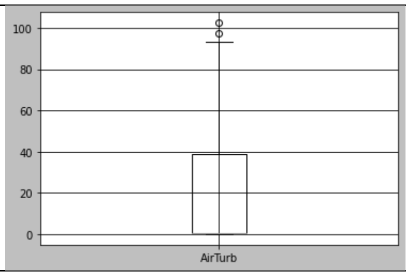
PlaneRadTemp	
Histogram	Box Plot
	
Comments:	Comments: only outliers at this point

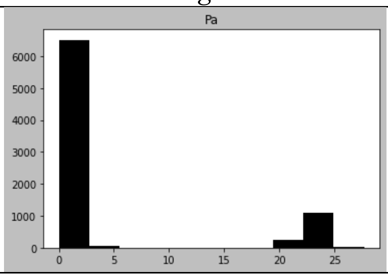
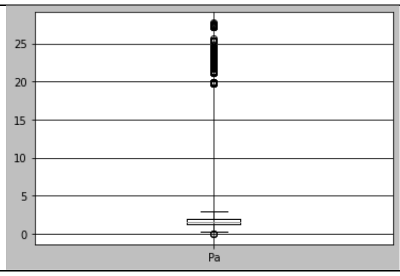
Ta	
Histogram	Box Plot
	
Comments:	Comments: many, many outliers

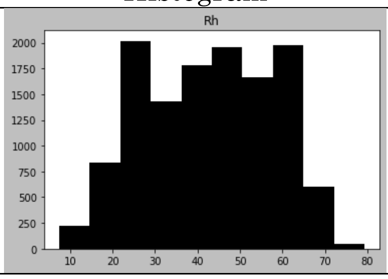
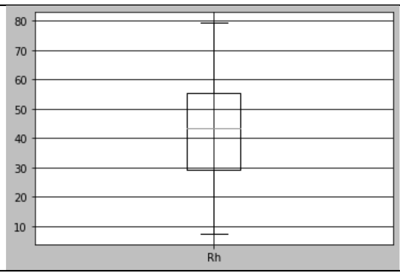
Tmrt	
Histogram	Box Plot
	
Comments:	Comments: too many outliers

Vel	
Histogram	Box Plot

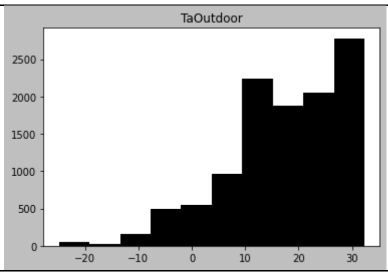
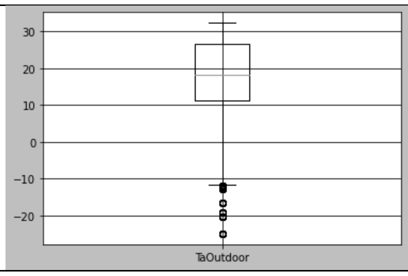
	
Comments:	Comments: data filled with outliers

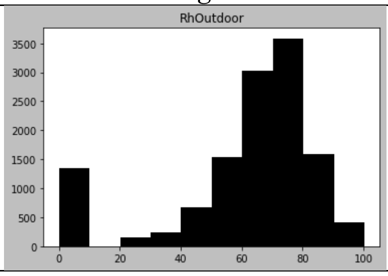
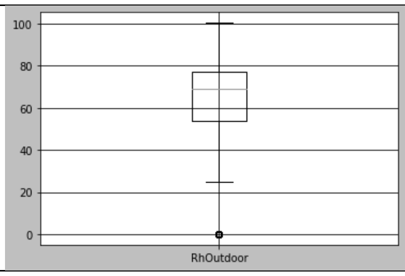
AirTurb	
Histogram	Box Plot
	
Comments:	Comments: very few outliers

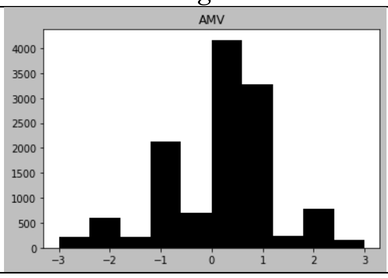
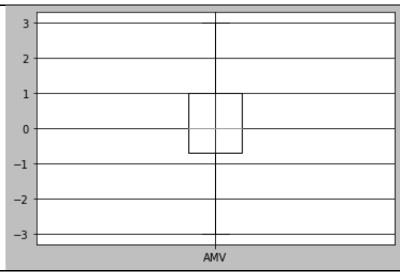
Pa	
Histogram	Box Plot
	
Comments:	Comments: a number of outliers

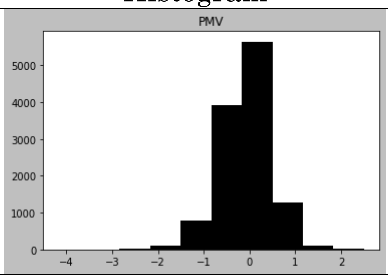
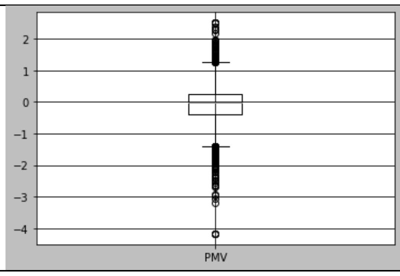
Rh	
Histogram	Box Plot
	
Comments:	Comments: no outliers

TaOutdoor	
Histogram	Box Plot

	
Comments:	Comments: some outliers

RhOutdoor	
Histogram	Box Plot
	
Comments:	Comments: one or two outliers

AMV	
Histogram	Box Plot
	
Comments:	Comments: no outliers

PMV	
Histogram	Box Plot
	
Comments:	Comments: many outliers

3. Find the missing values in each of the dimension (do this for both input and output dimensions), and fill these using an "appropriate" methodology that we've discussed in the class. You may also choose to drop a certain sample based on your analysis. Mention your approach and its justification.

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
Age	2916	Median	Simple and easy method
Clo	1407	Median	Simple and easy method
Met	1888	Median	Simple and easy method
Dewpt	3552	Median	Simple and easy method
PlaneRadTemp	7022	Median	Simple and easy method
Ta	21	Median	Simple and easy method
Tmrt	3702	Median	Simple and easy method
Vel	3701	Median	Simple and easy method
AirTurb	5601	Median	Simple and easy method
Pa	4556	Median	Simple and easy method
Rh	36	Median	Simple and easy method
TaOutdoor	1369	Median	Simple and easy method
RhOutdoor	20	Median	Simple and easy method
AMV	56	Median	Simple and easy method
PMV	697	Median	Simple and easy method

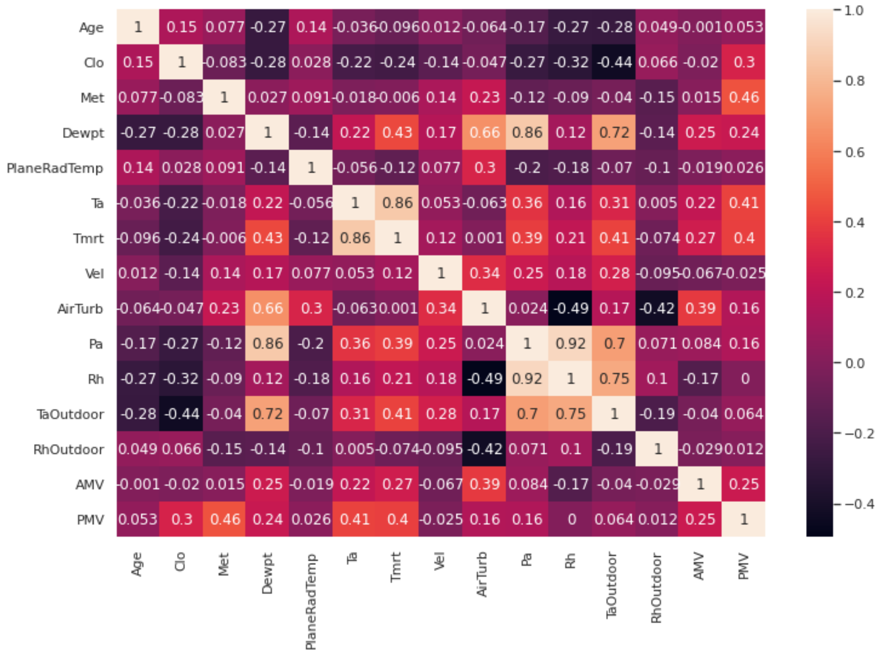
4. For each of the dimension, find out the outliers (noisy data) and handle these appropriately.

Dim Name	Number of Outliers	Smooth using/ Dropped	Reason for selecting a certain approach
Age	1359	using logical ranges to replace values with median	Visually possible
Clo	373	Using inter quartile range to replace values with median	Highly reliable method
Met	1731	Using inter quartile range to replace values with median	Highly reliable method
Dewpt	0	Using inter quartile range to replace values with median	Highly reliable method
PlaneRadTemp	452	Using inter quartile range to replace values with median	Highly reliable method
Ta	539	Using inter quartile range to replace values with median	Highly reliable method
Tmrt	343	Using inter quartile range to replace values with median	Highly reliable method
Vel	309	Using inter quartile range to replace values with median	Highly reliable method
AirTurb	2	Using inter quartile range to replace values with median	Highly reliable method
Pa	1352	Using inter quartile range to replace values with median	Highly reliable method
Rh	0	Using inter quartile range to replace values with median	Highly reliable method
TaOutdoor	124	Using inter quartile range to replace values with median	Highly reliable method
RhOutdoor	1349	Using inter quartile range to replace values with median	Highly reliable method
AMV	0	Using inter quartile range to replace values with median	Highly reliable method
PMV	259	Using inter quartile range to replace values with median	Highly reliable method

5. Using the variance that you've calculated above, for each dimension, comment whether you'll select the input dimension or no. (don't drop a dimension at this point)

Dim Name	Variance	Apply filter or no, reason
Age	462556	Yes. Values have too high variance, resulting in no worthy pattern in the data whatsoever.
Clo	0.049	Yes. Too less variation in values, i.e., could be considered similar.
Met	0.184	Yes. Too less variation in values, i.e., could be considered similar.
Dewpt	34.84	No. Ideal data.
PlaneRadTemp	1.084	No. Ideal data.
Ta	2.054	No. Ideal data.
Tmrt	2.258	No. Ideal data.
Vel	0.00624	Yes. Too less variation in values, i.e., could be considered similar.
AirTurb	627.522	No. Ideal data.
Pa	66.522	No. Ideal data.
Rh	226.836	No. Ideal data.
TaOutdoor	113.743	No. Ideal data.
RhOutdoor	610.30	No. Ideal data.

6A. Create a correlation matrix (Heat Map) for all the dimensions (input and output).



6B. Using the above correlation matrix, comment what are the most informative dimensions, and which are the least. Note that, be careful since we have two response variables in the dataset (i.e., PMV and AMV regression and classification respectively)

AMV: Dewpt, Ta, Tmrt, AirTurb, RhOutdoor, PMV

PMV: Clo, Met, Dewpt, Ta, Tmrt, AMV

8. **Apply entropy followed by information gain on the selected columns. Specify your selection criteria.**

AGAINST AMV

Dim name	Entropy	Info Gain	Reason
Dewpt	7.84361	1.7132	Relatively high correlation against AMV
Ta	8.27324	1.31277	Relatively high correlation against AMV
Tmrt	8.20147	1.77703	Relatively high correlation against AMV
AirTurb	6.25414	1.44382	Relatively high correlation against AMV
RhOutdoor	7.2004	0.84255	Relatively high correlation against AMV
AMV	-	-	AMV itself
PMV	3.4737	0.70471	Relatively high correlation against AMV

Dim name	Entropy	Info Gain	Reason
Clo	7.40891	2.67315	Relatively high correlation against PMV
Met	4.99779	0.1.68797	Relatively high PMV against AMV
Dewpt	7.84361	3.78205	Relatively high correlation against PMV
Ta	8.27324	3.42146	Relatively high correlation against PMV
Tmrt	8.20147	3.93297	Relatively high correlation against PMV
AMV	3.4737	0.68441	Relatively high correlation against PMV

Part B. Applying Algorithms

1. For this part, split the data randomly into 80/20 percent. Where 80% represents the training data. Also normalize the dataset as you see fit.

[done in colab]

2A. Apply forward selection, considering PMV as response variable and Multilinear regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector	Performance achieved (using MSE)

2B. Apply backward selection, considering PMV as response variable and Multilinear regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector	Performance achieved (using MSE)

3A. Apply forward selection, considering AMV as response variable and Logistic regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector	Performance achieved (using accuracy)

3B. Apply backward selection, considering AMV as response variable and Logistic regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector	Performance achieved (using accuracy)

4. Using the optimal feature vector that you've figured out from your analysis above, apply 3-fold cross validation for both regression and classification problems (PMV and AMV respectively). Write down the optimal parameters values for each of the model. Further, plot confusion matrix for the classification part.