Introduction to Data Science

Course Project

Report Document

<Rida Malik >

<21L-5608>

<Section 3D>

**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:
   1. Code File (Jupyter Notebook): L210000\_Code.ipynb
   2. Report Document (This File): L210000\_Report.pdf
4. Project will not be evaluated if:
   1. You submit python (.py) files
   2. You submit multiple .ipynb files
   3. You submit compressed (.rar or .zip) files
   4. 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 😺

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

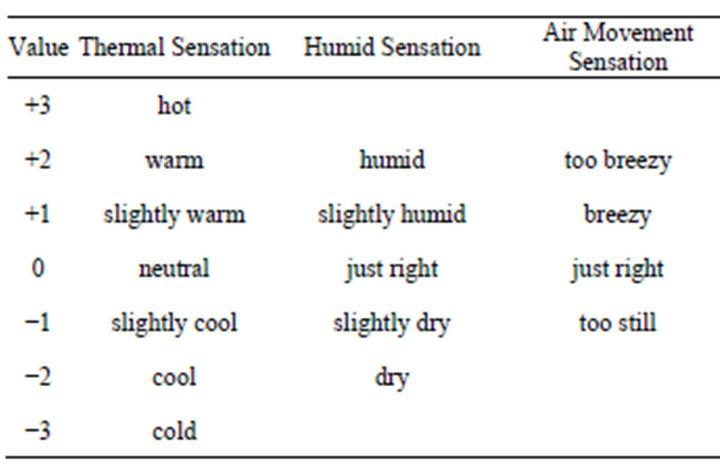
*TA Emails*

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:



**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 | Categorical | 12341 | 225 | 1349 | -1.0 | 292 | 1.0 | 17.73 | 9.36 | 1.88 | 64.5 |
| Clo | Categorical | 5677 | 6889 | 5616 | 15.96 | 29.81 | 23.0 | 23.18 | 23.19 | 4.00 | 1.24 |
| Met |  | 12509 | 57 | 384 | 0.0 | 1.94 | 0.62 | 0.62 | 0.60 | 1.95 | 0.21 |
| Dewpt |  | 5544 | 7022 | 5430 | -7.42 | 11.70 | 0.3 | 0.21 | 0.40 | 3.84 | 1.04 |
| PlaneRadTemp |  | 5677 | 6889 | 0 | 16.41 | 30.85 | 23.3 | 23.24 | 23.24 | 3.99 | 1.22 |
| Ta |  | 8865 | 3701 | 0 | 16.61 | 37.44 | 22.5 | 23.45 | 23.40 | 9.77 | 1.50 |
| Tmrt |  | 9338 | 3228 | 1417 | 16.639 | 31.80 | 22.5 | 23.45 | 23.38 | 1.08 | 1.56 |
| Vel |  | 5561 | 7005 | 5419 | 0.00 | 1.814 | 0.11 | 0.14 | 0.13 | 3.86 | 0.10 |
| AirTurb |  | 12531 | 35 | 0 | 7.40 | 79.3 | 64.0 | 42.5 | 42.9 | 1.95 | 15.06 |
| Pa |  | 7910 | 4656 | 0 | 0.00 | 27.7 | 2.1 | 5.12 | 3.55 | 7.80 | 8.15 |
| Rh |  | 12521 | 45 | 1709 | -0.62 | 32.9 | 22.8 | 20.74 | 22.4 | 1.95 | 7.40 |
| TaOutdoor |  | 12547 | 19 | 313 | 6.10 | 100 | 55.0 | 56.6 | 57.8 | 1.95 | 18.12 |
| RhOutdoor |  | 12546 | 20 | 27 | -22.70 | 37.8 | 0.0 | 18.74 | 19.9 | 1.96 | 12.80 |
| AMV |  | 12515 | 51 | 0 | 1.00 | 3.00 | 2.0 | 1.85 | 2.00 | 1.95 | 0.414 |
| PMV |  | 11177 | 1389 | 1222 | 0.00 | 1.0 | 0.05 | 0.112 | 0.10 | 1.56 | 0.104 |

**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).**

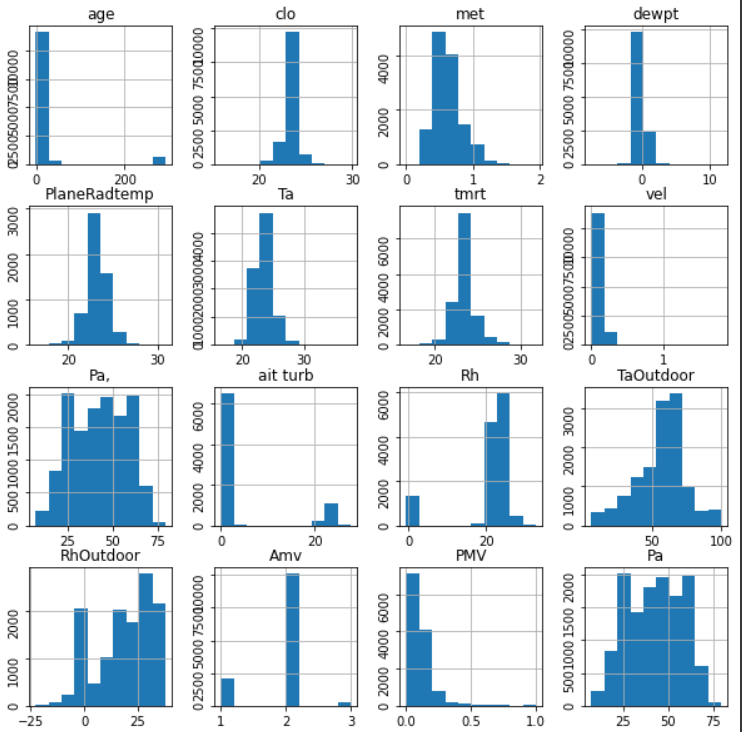
**Bos plot are present in IPYNB file.**

| **Age** | |
| --- | --- |
| Histogram | Box Plot |
| Age is right-skewed(positively skewed).  The histogram shows that it has very less number of outliers. | The whole data of age lies above the median. |
| Comments: | Comments: |
| **Clo** | |
| Histogram | Box Plot |
| Clo seems symmetric so it has normal distribution. | Mean and median seems to be almost equal. |
| Comments: | Comments: |

| **Met** | |
| --- | --- |
| Histogram | Box Plot |
| Met is positively skewed. | Data lies above median. |
| Comments: | Comments: |
| **Dewpt** | |
| Histogram | Box Plot |
| Dewpt is almost symmetric. | Mean equals to median. |
| Comments: | Comments: |

| **PlaneRadTemp** | |
| --- | --- |
| Histogram | Box Plot |
| Normally skewed | Mean equals to median. |
| Comments: | Comments: |
| **Ta** | |
| Histogram | Box Plot |
| Normally skewed | Mean equals to median. |
|  |  |
| Comments: | Comments: |
| **Tmrt** | |
| Histogram | Box Plot |
| Normally skewed | Mean equals to median. |
| Comments: | Comments: |
| **Vel** | |
| Histogram | Box Plot |
| Positively skewed | Data lies above median. |
| Comments: | Comments: |
| **Pa** | |
| Histogram | Box Plot |
| Normally skewed | Mean equals to median. |
| Comments: | Comments: |
| **Air Turb** | |
| Histogram | Box Plot |
| Positively skewed | Data lies above median. |
| Comments: | Comments: |
| **Rh** | |
| Histogram | Box Plot |
| Negatively skewed | Data lies below the median. |
| Comments: | Comments: |

| **Ta Outdoor** | |
| --- | --- |
| Histogram | Box Plot |
| Normally skewed | Mean equals to median. |
| Comments: | Comments: |
| **Rh Outdoor** | |
| Histogram | Box Plot |
| Negatively skewed | Data lies below the median. |
|  |  |
| Comments: | Comments: |
| **AMV** | |
| Histogram | Box Plot |
| Normally skewed | Mean equals to median. |
| Comments: | Comments: |
| **PMV** | |
| Histogram | Box Plot |
| Positively skewed | Data lies above median. |
| Comments: | Comments: |

****

**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 | 225 | Mode | It’s a categorical data so using mode is the best approach. |
| Clo | 6889 | Median | In the case of outliers, median is the best approach. |
| Met | 57 | Mean | The mean gives the average value of data that’s why it was used. |
| Dewpt | 7022 | Median | In the case of outliers, median is the best approach. |
| PlaneRadTemp | 6889 | Median | In the case of outliers, median is the best approach. |
| Ta | 3701 | Mean | The mean gives the average value of data that’s why it was used. |
| Tmrt | 3228 | Mean | The mean gives the average value of data that’s why it was used. |
| Vel | 7005 | Median | In the case of outliers, median is the best approach. |
| AirTurb | 35 | Mean | The mean gives the average value of data that’s why it was used. |
| Pa | 4656 | Mean | The mean gives the average value of data that’s why it was used. |
| Rh | 45 | Mean | The mean gives the average value of data that’s why it was used. |
| TaOutdoor | 19 | Mean | The mean gives the average value of data that’s why it was used. |
| RhOutdoor | 20 | Mean | The mean gives the average value of data that’s why it was used. |
| AMV | 51 | Dropped | They are response variables so null values are dropped so that the values remain real and give proper predictions. |
| PMV | 1389 | Dropped | They are response variables so null values are dropped so that the values remain real and give proper predictions. |

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

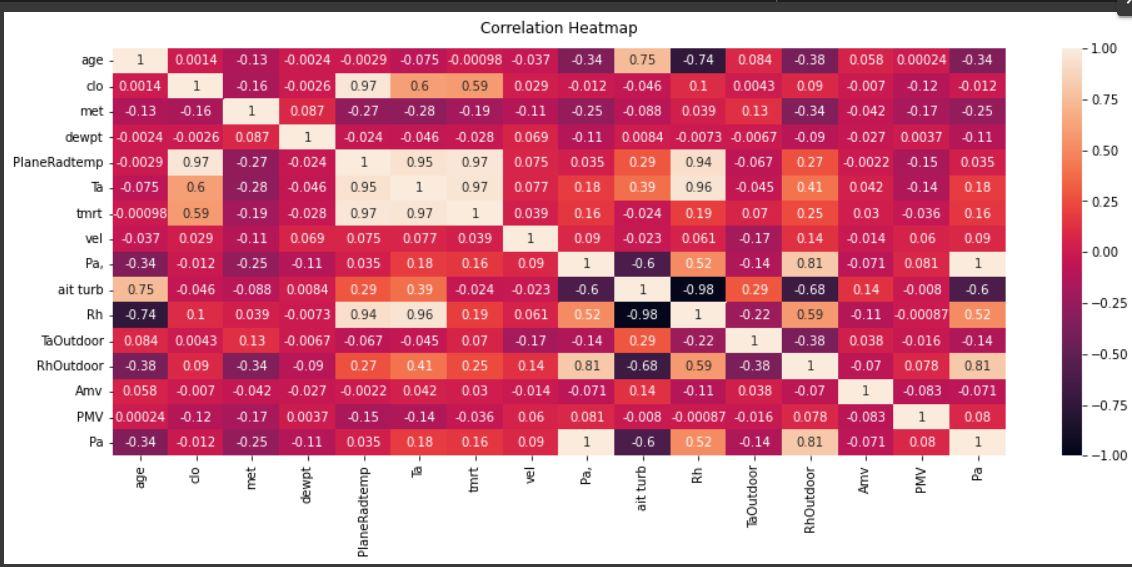
| Dim Name | Number of Outliers | Smooth using/ Dropped | Reason for selecting a certain approach |
| --- | --- | --- | --- |
| Age | 1349 | Smooth(Min Max Scaler) | Min-Max scaler is |
| Clo | 5616 | Smooth(Min Max Scaler) | the best approach |
| Met | 384 | Smooth(Min Max Scaler) | and it the the best |
| Dewpt | 5430 | Smooth(Min Max Scaler) | way to normalize |
| PlaneRadTemp | 0 | Smooth(Min Max Scaler) | the data. It scales |
| Ta | 0 | Smooth(Min Max Scaler) | the data between 0- |
| Tmrt | 1417 | Smooth(Min Max Scaler) | 1 due to which the |
| Vel | 5419 | Smooth(Min Max Scaler) | handling of data |
| AirTurb | 0 | Smooth(Min Max Scaler) | becomes easy. |
| Pa | 0 | Smooth(Min Max Scaler) |  |
| Rh | 1709 | Smooth(Min Max Scaler) |  |
| TaOutdoor | 313 | Smooth(Min Max Scaler) |  |
| RhOutdoor | 27 | Smooth(Min Max Scaler) |  |
| AMV | 0 | Smooth(Min Max Scaler) |  |
| PMV | 1222 | Smooth(Min Max Scaler) |  |

**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)**

**After normalization, variance threshold for the data is: 0.05. On this basis only Air Turb is dropped.**

| Dim Name | Variance | Apply filter or no, reason |
| --- | --- | --- |
| Age | 0.047 | Selected |
| Clo | 0.0036 | Selected |
| Met | 0.0117 | Selected |
| Dewpt | 0.0013 | Selected |
| PlaneRadTemp | 0.0071 | Selected |
| Ta | 0.0052 | Selected |
| Tmrt | 0.0079 | Selected |
| Vel | 0.0014 | Selected |
| AirTurb | 0.0866 | Not selected |
| Pa | 0.0438 | Selected |
| Rh | 0.0486 | Selected |
| TaOutdoor | 0.0371 | Selected |
| RhOutdoor | 0.0444 | Selected |
| AMV | 0.0430 | Selected |
| PMV | 0.0096 | Selected |

**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)**

| Dim Name | MOST INFORMATIVE | LEAST INFORMATIVE |
| --- | --- | --- |
| Age | YES |  |
| Clo |  | LEAST |
| Met | YES |  |
| Dewpt | YES |  |
| PlaneRadTemp | YES |  |
| Ta |  | LEAST |
| Tmrt |  | LEAST |
| Vel |  | LEAST |
| AirTurb | YES |  |
| Pa | YES |  |
| Rh | YES |  |
| TaOutdoor | YES |  |
| RhOutdoor |  | LEAST |
| AMV | - |  |
| PMV | - |  |

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

| Dim Name | Entropy | Info Gain | Reason |
| --- | --- | --- | --- |
| Age | 7.692 | 3.28 | Information gain is greater than threshold |
| Clo |  |  | Information gain is greater than threshold |
| Met | 7.43 | 5.706 | Information gain is greater than threshold |
| Dewpt | 1.052 | 4.22 | Information gain is greater than threshold |
| PlaneRadTemp | 8.341 | 7.22 | Information gain is greater than threshold |
| Ta |  |  | Information gain is greater than threshold |
| Tmrt |  |  | Information gain is greater than threshold |
| Vel | 4.286 | 3.424 | Information gain is greater than threshold |
| AirTurb | 10.168 | 8.76 | Information gain is greater than threshold |
| Pa |  |  | Information gain is greater than threshold |
| Rh | 6.981 | 5.273 | Information gain is greater than threshold |
| TaOutdoor | 7.03 | 5.554 | Information gain is greater than threshold |
| RhOutdoor |  |  | Information gain is greater than threshold |
| AMV |  | - | Information gain is greater than threshold |
| PMV |  | - | Information gain is greater than threshold |

***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.**

**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 |
| --- | --- |
| RH,DEWPT | The best performance is achieved through forward selection for PMV. |

**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 |
| --- | --- |
| Pa  DEWPT | The best performance is achieved through forward selection for PMV |

**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 |
| --- | --- |
| MET,AIR TURB | 50.26% is the performance achieved. |

**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 |
| --- | --- |
| PLANE RED TEMP,AIR TURB | 50.02% is the performance achieved. |

**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.**