Introduction to Data Science

Course Project

Report Document

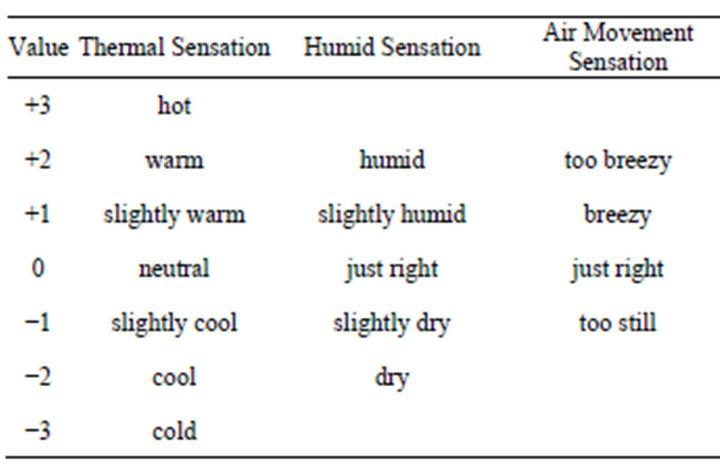
Murtaza Ahmed Bhatti

21L-6234

Section: BDS-3C

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 dimensions, 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 | **float** | **9650** | **2916** | **1359** | **0.0** | **1996** | **24.0** | **308.637** | **35.0** | **77.39** | **680.11** |
| AMV | **float** | **12511** | **55** | **0** | **-3.0** | **3.0** | **0.0** | **0.10** | **0.0** | **1.05** | **1.10** |
| Met | **float** | **10679** | **1887** | **1732** | **0.1** | **4.5** | **1.0** | **1.06** | **1.1** | **0.013** | **0.42** |
| Clo | **float** | **11160** | **1406** | **373** | **0.15** | **2.13** | **0.77** | **0.77** | **0.7517** | **0.027** | **0.22** |
| Dewpt | **float** | **9014** | **3552** | **0** | **-1.953** | **26.89675** | **17.4** | **13.62** | **14.1** | **5.59** | **5.90** |
| PlaneRadTemp | **float** | **5544** | **7022** | **452** | **-7.42** | **11.7** | **0.3** | **0.21** | **0.2** |  | **1.04** |
| Ta | **float** | **12546** | **20** | **540** | **15.96** | **31.0** | **23.2** | **23.17** | **23.13** | **0.83** | **1.43** |
| Tmrt | **float** | **8865** | **3701** | **344** | **16.61** | **37.445** | **22.5** | **23.45** | **23.35** | **0.51** | **1.50** |
| Vel | **float** | **8866** | **3700** | **309** | **0.0** | **1.88** | **0.1** | **0.11** | **0.1** | **0.00094** | **0.079** |
| AirTurb | **float** | **6965** | **5601** | **2** | **0.0** | **102.45** | **0.5** | **18.26** | **0.5** | **0.014** | **25.04** |
| Pa | **float** | **7910** | **4656** | **1352** | **0.0** | **27.7** | **2.1** | **5.12** | **1.55** | **0.032** | **8.15** |
| Rh | **float** | **12531** | **35** | **0** | **7.4** | **79.3** | **64.0** | **42.52** | **43.28** | **204.93** | **15.06** |
| PMV | **float** | **11870** | **696** | **259** | **-4.17** | **2.5** | **0.1** | **-0.073** | **-0.03** | **0.14** | **0.53** |
| TaOutdoor | **float** | **11198** | **1368** | **124** | **-24.9** | **32.35** | **27.55** | **17.17** | **18.2** | **99.89** | **10.66** |
| RhOutdoor | **float** | **12547** | **19** | **1349** | **0.0** | **100.35** | **0.0** | **61.1018.74** | **68.79** | **154.94** | **24.70** |

**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).**

|  |  |
| --- | --- |
| **Age** | |
| Histogram | Box Plot |
|  |  |
| The data is not normally distributed it is right skewed | We can wee there are many outliers |

|  |  |
| --- | --- |
| **AMV** | |
| Histogram | Box Plot |
|  |  |
| The data almost lies in the middle but still nt normally distruibuted | The data contain no outliers |

|  |  |
| --- | --- |
| **Met** | |
| Histogram | Box Plot |
|  |  |
| It is right skewed | Too many outliers |

|  |  |
| --- | --- |
| **Clo** | |
| Histogram | Box Plot |
|  |  |
| Right skewed | there are many outliers |

|  |  |
| --- | --- |
| **Dewpt** | |
| Histogram | Box Plot |
|  |  |
| bimodal distribution | No Outliers |
|  |  |

|  |  |
| --- | --- |
| **PlaneRadTemp** | |
| Histogram | Box Plot |
|  |  |
| unimodal histogram | Too many outliers and is not suitable for input dimension |

|  |  |
| --- | --- |
| **Ta** | |
| Histogram | Box Plot |
|  |  |
| unsymmetrical unimodal histogram | Many Outliers are there |
|  |  |

|  |  |
| --- | --- |
| **Tmrt** | |
| Histogram | Box Plot |
|  |  |
| Right skewed | Outlier on both the edges of the data |
|  |  |

|  |  |
| --- | --- |
| **Vel** | |
| Histogram | Box Plot |
|  |  |
| Right Skewed | Many Outliers in the distribution towards Q3 |

|  |  |
| --- | --- |
| **AirTurb** | |
| Histogram | Box Plot |
|  |  |
| Right Skewed with outliers | Many Outliers |
|  |  |

|  |  |
| --- | --- |
| **Pa** | |
| Histogram | Box Plot |
|  |  |
| Right Skewed | Outliers and median high far away |

|  |  |
| --- | --- |
| **Rh** | |
| Histogram | Box Plot |
|  |  |
| The data is uniformally distributed | Data contained almost no outliers |

|  |  |
| --- | --- |
| **PMV** | |
| Histogram | Box Plot |
|  |  |
| The data is weakly left skewed | Outliers on Both upper and lower quartile |
|  |  |

|  |  |
| --- | --- |
| **TaOutdoor** | |
| Histogram | Box Plot |
|  |  |
| Strong Left Skewed | Comments: Not many outliers but mean will be affected due to outliers |
|  |  |

|  |  |
| --- | --- |
| **RhOutdoor** | |
| Histogram | Box Plot |
|  |  |
| modal Distribution | There exist outliers but very few |

**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.**

**Comment:**

**Column with more than 50% of the missing data is not considered good to be filled so that’s whole column had to be dropped.**

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Number of Missing Values | Filled using OR Dropped | Reason for selecting a certain approach |
| Age | **2916** | median | The data is replaced with median because it is right skewed and the outliers have disrupted the mean |
| AMV | **55** | median | The data is spread normally so it is suitable to replace null values with mean median or mode |
| Met | **1887** | median | The data is right skewed and replacing with median is considered best approach in case of skewed distribution |
| Clo | **1406** | median | The data is right skewed and replacing with median is considered best approach in case of skewed distribution |
| Dewpt | **3552** | median | The data is leftt skewed and replacing with median is considered best approach in case of skewed distribution |
| PlaneRadTemp | **7022** | dropped | the data contained more than 50% of null values and is dropped |
| Ta | **20** | median | The data is left skewed and replacing with median is considered best approach in case of skewed distribution |
| Tmrt | **3701** | median | As the data is skewed it the best practice to replace skewed data with median |
| Vel | **3700** | mode | The histogram shows the data is not too widespread and most of the data set is appearing in one class ence it is best to use mode |
| AirTurb | **5601** | median | right-skewed and replacing with median is considered best approach in case of skewed distribution |
| Pa | **4656** | median | The data is right-skewed and replacing with median is considered best approach in case of skewed distribution |
| Rh | **35** | mean | The data is spread normally so it is suitable to replace null values with mean but we can also replace it with median or mode |
| PMV | **696** | median | The data is left-skewed and replacing with median is considered best approach in case of skewed distribution |
| TaOutdoor | **1368** | median | The data is left-skewed and replacing with median is considered best approach in case of skewed distribution |
| RhOutdoor | **19** | median | The data is left-skewed and replacing with median is considered best approach in case of skewed distribution |

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

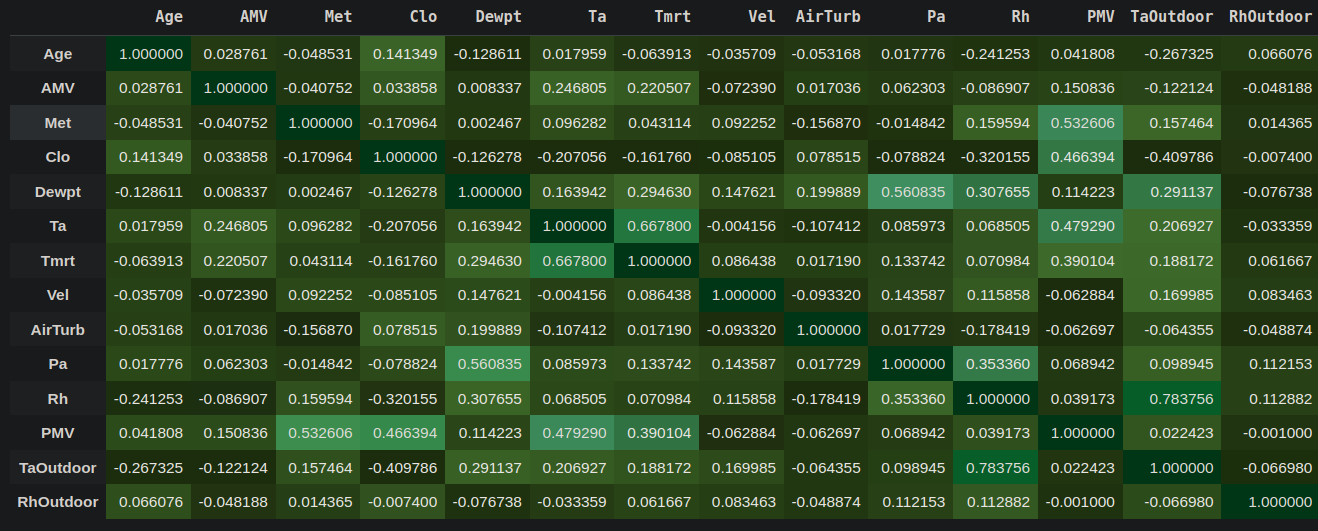
Note: Outliers from data set are removed

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Number of Outliers | Smooth using/ Dropped | Reason for selecting a certain approach |
| Age | **1359** | dropped | After removing the outliers we still have many samples so its te best practice in this situation to drop all the outlier |
| AMV | **0** | dropped |  |
| Met | **1732** | dropped |  |
| Clo | **373** | dropped |  |
| Dewpt | **0** | dropped |  |
| PlaneRadTemp | **452** | dropped |  |
| Ta | **540** | dropped |  |
| Tmrt | **344** | dropped |  |
| Vel | **309** | dropped |  |
| AirTurb | **2** | dropped |  |
| Pa | **1352** | no outliers |  |
| Rh | **0** | dropped |  |
| PMV | **259** | dropped |  |
| TaOutdoor | **124** | dropped |  |
| RhOutdoor | **1349** | dropped |  |

**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)77.39839317852865, 1.0538105006519092, 0.013049765927104256, 0.02710804066225975, 5.590438868501571, 0.83120021768928, 0.515658494774394, 0.0009482748593283201, 0.014493890243015834, 0.03234172724877394, 204.93991671878777, 0.14816228694181824, 99.89213403677209, 154.94392036932717**

|  |  |  |
| --- | --- | --- |
| Dim Name | Variance | Apply filter or no, reason |
| Age | **77.39** | Selected because the variance is very low or high |
| AMV | **1.05** | Ouput dimension |
| Met | **0.013** | Very low variance |
| Clo | **0.027** | Very low variance |
| Dewpt | **5.59** | Selected because the variance is very low or high |
| PlaneRadTemp | column was dropped as it had more than 50% as null data | Dropped |
| Ta | **0.83** | Selected because the variance is very low or high |
| Tmrt | **0.51** | Selected because the variance is very low or high |
| Vel | **0.00094** | Very low variance |
| AirTurb | **0.014** | Very low variance |
| Pa | **0.032** | Very low variance |
| Rh | **204.93** | Selected because the variance is very low or high |
| PMV | **0.14** | Output dimension |
| TaOutdoor | **99.89** | Selected because the variance is very low or high |
| RhOutdoor | **154.94** | Selected because the variance is very low or high |

**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) ?**

PMV is our response variable which has most positive correlation with Met,Clo,Ta and Tmrt Features.

AMV has highest positive correlation with Tm ,Tmrt and PMV the response variable.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Dim name | Entropy | Info Gain | Reason |
| Age |  |  |  |
| … |  |  |  |
| 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.**

|  |
| --- |
|  |
|  |

**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.**

**Optimal feature set:**

**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 |
|  |  |

**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 |
|  |  |

**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 |
|  |  |

**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.**