**1. Exploratory data analysis (half page)**

* What is the proportion of cloudy days to clear days.?
  + Not cloudy = 519, Cloudy = 1120
* Graphical user interface

  Description automatically generated with medium confidenceAre there any attributes you need to consider omitting from your analysis? (1 Mark)
  + Attributes to omit : Location, Day, Month, Year
  + Why
* descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-valued attributes
  + skewness
    - To summarize, generally **if** the distribution of **data is skewed** to the left, the mean is less than the median, which is often less than the mode. **If** the distribution of **data is skewed** to the right, the mode is often less than the median, which is less than the mean
  + Mean
  + Sd
  + lots of na values

boxplot

Chart

Description automatically generated

* + Highest range.
  + lots of outliers

**2. Data Cleaning and Pre-processing (half page)**

**2.1. handling Missing values**

**2.2. dealing with Categorical values**

**2.3. Divide the dataset into Dependent & Independent variable**

**2.4. Split the dataset into training and test set**

**2.5 Feature Scaling ?**

but there could be some value in standardizing your variables if you're interested in predictor importance scores. RF will tend to favour highly variable continuous predictors because there are more opportunities to partition the data. A better way to deal with this issue, however, is to use particular approaches ([cite](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-25))

**3. Original modelling**

**3.1 – 3.5 5 models**

**Evaluation:**

**3.6 confusion matrix**

**Accuracy is higher for DT**

**- Random dataset is good for DT to learn**

**- setting Mfinal in its best size, still the replicated dataset, so called bags, is not good for boosting to learn**

**- Boosting and bagging are overfitted. When I tried to use taining dataset the accuracy went up to 81%**

**3.7.1 ROC 3.7.1 AUC**

**Chart, line chart

Description automatically generated**

**3.8 variable importance**

* Which variables could be omitted from the data with very little effect on performance? Give reasons.
* Which are the most common and least common for VarImportnace?

**4. model improved :**

**4.1. based on simplicity**

* Describe your model, either with a diagram or written explanation.
* How well does your model perorm, and how does it compare to those in Part 4?
* What factors were important in your decision and why you chose the attributes you used.

**4.2 based on performance**

* How to improve
  1. prune trees
  2. CV
  3. boosting
* You may do this by adjusting the parameters, and/or cross-validation of the basic models in Part 4, or using an alternative tree-based learning algorithm.
* Show that your model is better than the others using appropriate measures.
* Describe how you created your improved model, and why you chose that model.
  1. What factors were important in your decision and why you chose the attributes you used.

**5.ANN**

* Comment on attributes used and your data pre-processing required.
* How does this classifier compare with the others? Can you give any reasons?