Machine learning disposition

* Introduction
  + Aim of the project
  + What is machine learning
* Description of data
  + Data gathered from weather stations around Edinburgh
  + Gather from 24th of October to 7th of November
  + 45178 datapoints
    - 9 observational features
    - Locational features
  + Missing data
    - Replaced with random representation of that feature
* Description of classifiers
  + Classification methods used
    - Decision tree
      * Whitebox method
        + Can see inside and understand
      * Creates a tree with leaves
        + Each leaf can produce either a true or false
        + Test variable propagates down the tree until it reaches a final leaf
    - Neural network
      * Black box method
        + Cant see inside and understand
      * Creates a network of “neurons” with different weights between them
        + The sequence of neurons that fire determines the outcome
    - Nearest neighbours
      * Groups data together in n-dimensional space (n features)
        + Tries to define boundaries in space that determines what class the result should be
    - Random forest
      * Creates many decision trees with different weights
        + Uses the results from all of them to determine the result
  + Optimisation used
    - Cross-validation
      * F
    - Standardising
      * Setting mean as 0 and variance as 1
    - Normalising
      * Changing the feature interval to -1,1
    - Hyper-parameter tuning
      * Loops through a given set of initializing variables for a classifier to determine what the best variables are
    - Oversampling
      * Increases the number of datapoints for data that is underrepresented
        + Tries to combat over-fitting
    - Result comparison
      * Compares the result from multiple classifiers to try and find similarities between them
        + Idea is to keep the good results and discard the bad
* Results performance and comparison
* Summary conclusion
  + How to define best results
    - Avg/total precision and recall
    - Over fitting
  + Are the results good enough
    - 0.8 isnt very good for a general classifier
    - 0.8 is good when it comes to predicting weather but it is hard to say in this case
  + Is the data good enough
    - Some very rare cases in the advanced file that make them hard to incorporate in a classifier
    - More data could improve the model