A Comparison of Decision Tree (DT) and Random Forest (RF) for Weather Classification

INM431 Machine learning Project | Sahan Chowdhury | City University of London

Motivation & Description

- The weather is arguable one of the most crucial elements of our daily lives and the preservation of our planet. Agriculture, tourism, emotional wellbeing are only a few of the many aspects that are influenced by weather. In fact, it was reported there were over 55million monthly active users in 2020 whom used The Weather Channel App [1]. Foreseeing the weather has become crucial for humans to arrange their day-to-day activities.
- The objective of this machine learning task is to build two models for a supervised classification task using Decision Tree and Random Forest, to assess the performance for weather prediction, and compare that to a previous study of Ismaila Oshodi (2022).

Dataset Description & Preliminary analysis

- The dataset is a weather dataset for Seattle, Washington, USA, obtained from Kaggle.
- The dataset contains 1461 entries, dated from 01/01/2012 – 31/12/2015, and 6 characteristics (columns); 'date', 'precipitation', 'temp_max', 'temp_min', 'wind', 'weather'.
- 4 numeric characteristics were used as predictive variables ('precipitation', 'temp_max', 'temp_min', 'wind') to predict the target variable ('weather'). Table 1 shows the summary statistics.
- The date column was extracted into day, month and year.
- There was no missing values or duplicates.
- The target variable ('weather') was assigned from 1-5, each number corresponding to a distinct weather type.
- A new column temp_range was created, which calculated the range of daily temps (temp_max – temp_min).
- Binary columns were created for every season (Winter, Summer, Autumn, Spring), assigning each month to a specific season.
- Months 1,2,3,12 -> winter, months 6,7,8 -> summer, months 9,10,11, -> Autumn, months 4,5 -> Spring
- Performed standardization to ensure all predictor variables were scaled equally.
- There seemed to be no correlation between the predictor variables except for temp_max and temp_min which had a Pearsons correlation of 0.8757 as shown on figure (1) and figure (4).
- Box plots were plotted to visualize distribution of the data for the main predictors. Precipitation box plot had the most 'outliers', however these were not the usual outliers, but rather extreme rainy days. (figure 2)
- Plotted histogram of counts vs variable as shown on figure (3). The precipitation variable is right skewed, which explains why it has so many extreme values from the box-plot.
- Running initial decision tree and Random Forest modelling, produced near identical results to the reference papers Ismaila Oshodi (2022) and C. B. S, B. Shreegagana, B. H. S, I. Karanth, A. R. K. P and G. S

Summary statistics of predictors

predictor	min	median	max	std
precipitation	0	0	55.9	6.6802
temp_max	-1.6	15.6	35.6	7.3498
temp_min	-7.1	8.3	18.3	5.023
wind	0.4	3	9.5	1.4378

Table 1

Figure 1:

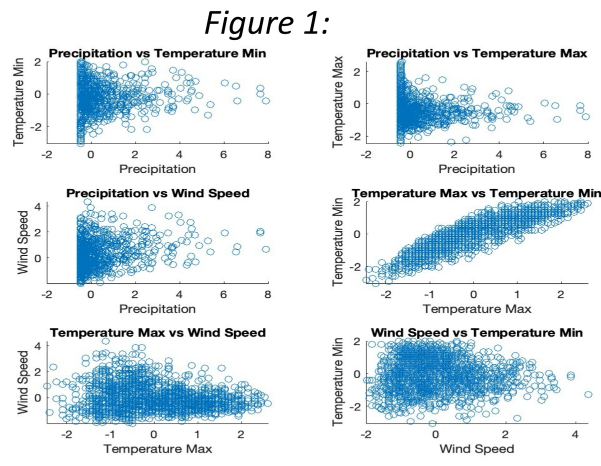


Figure 2:

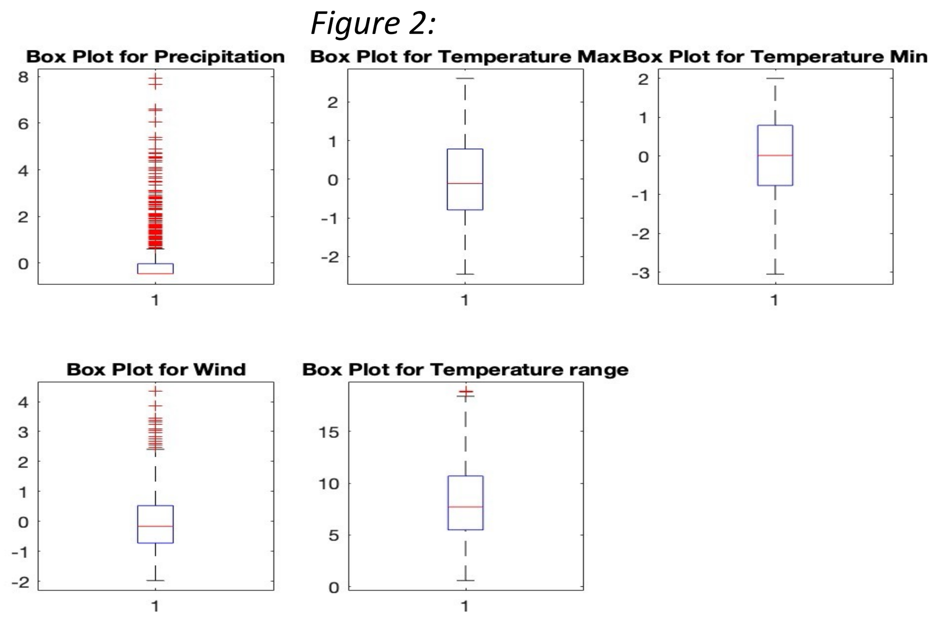


Figure 3:

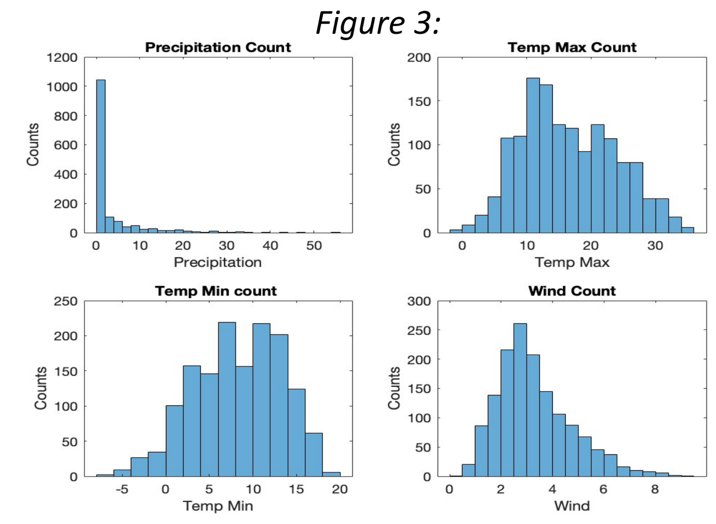
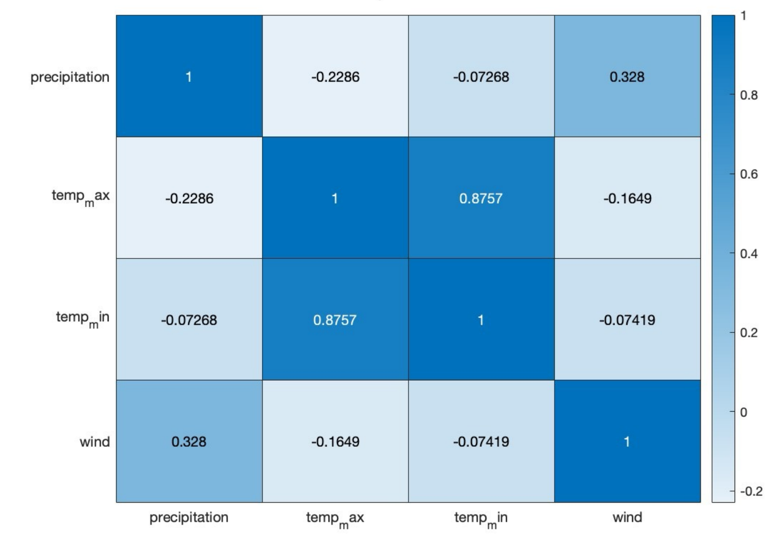


Figure 4:



Decision Tree

- Decision Trees is a supervised machine learning algorithm used for classification and/or regression tasks
- Decision Trees is like a flowchart, it contains of a root node, branches, internal nodes and leaf nodes.
- The way a decision tree works is by by dividing the source data set into subsets according to an attribute value test, a tree can be trained to learn. This procedure, known as recursive partitioning, is carried out iteratively on every derived subset. When splitting stops adding value to the predictions or when the subset at a node has all the same value for the target variable, the recursion is said to be finished. [4]

Pros

- Easily interpretable – The model is simple and can be easily interpreted
- Little to or no data preparation required – Decision Trees are more versatile than most other models, it can handle range of data types and data structures. [5]
- Very good in handling multi-classification tasks.

Cons

- Susceptible to overfitting – Decision Trees can become very large and complex, they can include noise when used on training data, and this may lead to low accuracy when tested on testing data or unseen data.
- High variance estimators – A slight variation within the data can alter the decision tree completely. [5]
- Certain level of bias – If there are dominant classes , Decision Tree can tend to creating biased trees.

Random Forest

- Random Forest is a supervised machine learning algorithm that uses ensemble learning for classification and/or regression tasks
- It's applied by first training the dataset to generate numerous Decision Trees, and then identifying the classification modes of each tree separately. [6]
- Random Forests have three main hyperparameters which need to be set before training; node size, number of trees and number of features sampled. [7]

Pros

- In training data, they are less sensitive to outlier data.
- The ease with which parameters can be set removes the necessity for tree pruning.
- It reduces the likelihood of overfitting – Random Forests aggregate predictions over multiple trees, hence overfitting is substantially less likely than Decision Tree.

Cons

- Takes longer time to train when compared to Decision Trees, as it builds numerous trees to combine outputs, rather than a single Decision Tree.
- Black Box Model - There is not much transparency or control among the decision making of the model.
- Not suitable for sparse data.

Hypothesis Statement

- From the associated research papers [3],[2], it was evident that both Random Forest and Decision Trees performed well. Both papers found Random Forest had the higher accuracy of 82.69% and 79.50% respectively [3],[2]. Similarly for Decision Tree it was 76.54% and 72.40%. [3],[2]
- The aim of this Machine Learning task is to obtain comparable results to the reference research papers
- It's expected that Random Forest should outperform Decision Trees.
- Random Forest usually have a higher accuracy than Decision Trees , and the margin of error and overfitting is also likely to be lower, however it's expected to have a higher train time.
- Precipitation is anticipated to become the most important predictor in this classification task.

Methodology

1. The dataset is split using holdout method. Holdout is a formed through a non-stratified partition.
2. The dataset is split into 80:20 split respectively for testing and training. The training set has 1169 instances and testing set has 292 instances.
3. Measures such as precision, recall, fscore are calculated for comparison.
4. Hyperparameter Optimization will used to improve model predictive performance and determine whether they are optimal models.
5. Train and test times will also be considered.
6. Final evaluation will determine the most optimal model

Decision Tree Parameters and Results

- Hyper-parameter model was set to auto, so that the model will automatically tune using internal optimization algorithm
- The optimized model reduced the error by 2.73%
- Also note for each weather type the accuracy was very high as shown in Table 2
- Model training evaluated 30 functions.
- Performance evaluations calculated yielded similar results to that off the research papers

Accuracy

Weather	Decision Tree	Random Forest
Drizzle	97%	96%
Fog	94%	93%
Rain	96%	96%
Snow	98%	98%
Sun	88%	88%

Table 2

Random Forest Parameters and Results

- The time taken to train model is almost three times more than Decision Tree.
- Fitcensemble was used for the classification task, and specifically the 'Bag' ensemble approach was applied
- Hyperparameter selection was also set to auto, so the model finds the best tuned model.
- The Random Forest model with hyper-parameter tuning performed the best among all models

Figure 6:

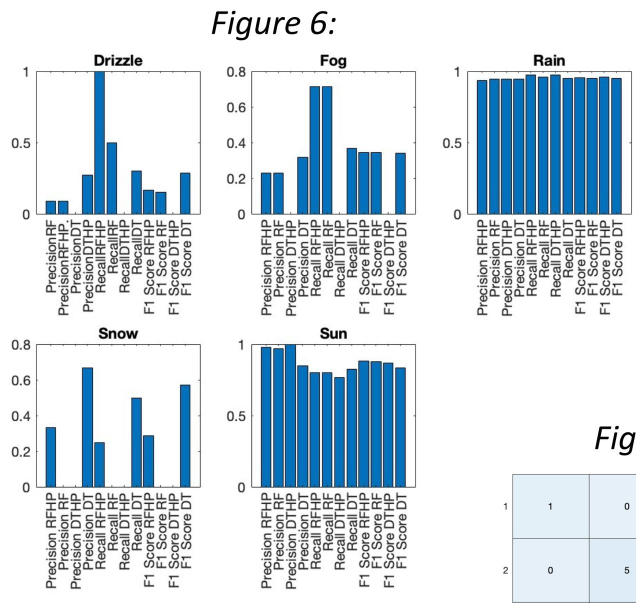


Figure 5:

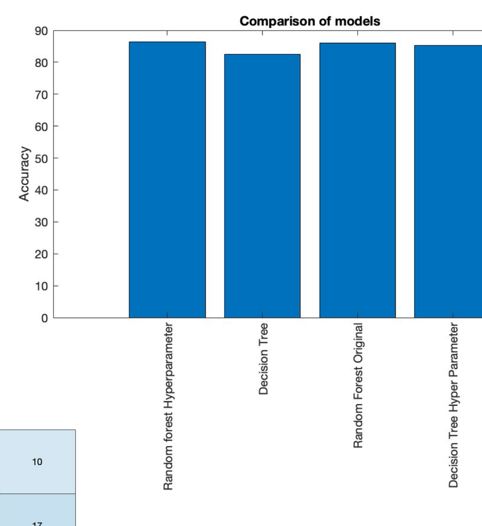


Figure 7:

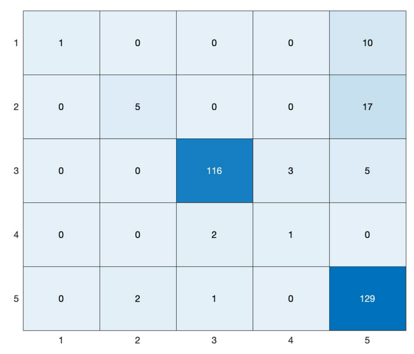
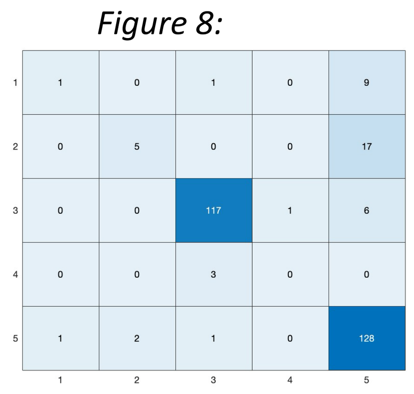


Figure 8:



Analysis and Evaluation of Results

- Analyzing the testing accuracy, both optimized models by hyper-parameter tuning for both Decision Trees and Random Forest performed very high, with an accuracy of 85.27% for the Optimized Decision Tree model and 86.64% for the Random Forest model. In comparison to Ismaila Oshodi (2022) which obtained 79.50% for Random Forest and 72.40% for the Decision Tree model . Figure (5) represent all model accuracies. This difference in performance may be due to incorporating binary seasonal predictor variables, rather than considering daily changes like the research paper did. The hyper parameter tuned Random Forest model performed the best as expected.
- The Accuracy scores which are visible from table 2, indicate that both Decision Tree and Random Forest are great classifiers for different weather Types, especially the weather type 'Snow'
- Looking at performance metrics its evident that 'Sun' and 'Rain', performed the best across Precision, Recall and F1 scores, this can be visually seen on Figure(6) . This suggests that the models are great at being precise and capturing positive instances for those weather types. The optimized Random Forest model, for instance achieved an f1 score of 0.96 for 'Rain' , which resembles the result from Ismaila Oshodi (2022) . This strong correlation may be down to the effectiveness of the Precipitation predictor, given the close association between precipitation and rain. Similarly for Decision Tree 'Rain' again a notably high f1 score of 0.95 this was also very similar to that off Ismaila Oshodi (2022) which was 0.91. However, its worth noting that the other weather types 'Drizzle', 'Snow', 'Fog', had much lower scores, with some almost negligible by the optimized models , this issue was also persistent in Ismaila Oshodi (2022) research paper, and this may be due to the fact there is similarities among these weather types with 'Sun' and 'Rain', there is a lack predictor variables to clearly differentiate between weathers, hence making it more difficult to predict obscure weathers such as fog.
- While both Random Forest and Decision Trees proved to effective classifiers, there were notable challenges during the process. Decision Trees can become too used to the training data, and this could cause overfitting , in the context of weather prediction this could mean the model might pick up peculiarities in the training set and represent this in true patterns in the broader meteorological phenomena. While Random Forest models may struggle to pick up intricate relationships in weather patterns, computationally more expensive and take much longer to train, it is the tradeoff for a model that has a lower rate of error and yields higher accuracy. Random Forest seems like the best model to predict weather despite of having no precise knowledge of atmospheric physics. [8]

Lessons Learnt

- While Decision Trees performed well, there was a noticeable issue of overfitting among the data, this should be considered in future.
- The ensemble nature of Random Forest is effective for a wide range of classification task
- Balanced dataset is vital to prevent biases and model performance overall

Future Work

- Addressing mixed weather conditions, i.e., Rain and Fog simultaneously .
- Use K-fold cross validation on the dataset to compare with Holdout.
- Handle imbalanced dataset using SMOTE.
- Use normalization and remove potential noise from the data and enhance accuracy.
- Larger dataset - to acquire weather patterns over a broader period and recognize the impact of Global Warming on weather conditions.
- Incorporating more variables/predictors to get a more realistic model.

References

[1] S. Bergman, "In this economy, the winner is weather apps," Business Insider. https://www.businessinsider.com/weather-apps-dont-work-that-well-but-we-are-addicted-2023-7?r=US&IR=T#:~:text=And%20then%20there%27s%20the%20big (accessed Dec. 22, 2023).

[2] Ismaila Oshodi (2022). Machine Learning-Based Algorithms for Weather Forecasting. International Journal of Artificial Intelligence and Machine Learning, 2(2), 12-20. doi: 10.51483/IJAIML.2.2.2022. 12-20.

[3] C. B. S, B. Shreegagana, B. H. S, I. Karanth, A. R. K. P and G. S, "Weather Prediction Analysis using Classifiers and Regressors in Machine Learning," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 894-900, doi: 10.1109/ICSSIT55814.2023.10061073.

[4] R. Kumar, "Decision Tree for the Weather Forecasting," International Journal of Computer Applications, vol. 76, no. 2, pp. 31–34, Aug. 2013, doi: https://doi.org/10.5120/13220-0620.

[5] IBM, "What is a Decision Tree | IBM," www.ibm.com. https://www.ibm.com/topics/decision-trees

[6[A. Mathew and J. Mathew, "Weather Forecasting Using the Random Forest Algorithm Analysis," Mar. 2022, doi: https://doi.org/10.5281/zenodo.6361990.

[7] IBM, "What is Random Forest? | IBM," www.ibm.com. https://www.ibm.com/topics/random-forest

[8] Meenal, Rajasekaran & Angel, Prawin & Pamela, D. & Rajasekaran, Ekambaram. (2021). Weather prediction using random forest machine learning model. Indonesian Journal of Electrical Engineering and Computer Science. 22. 1208. 10.11591/ijeecs.v22.i2.pp1208-1215.