BIRZEIT UNIVERSITY

Computer Science Department
Dr. Radi Jarrar
Project Two

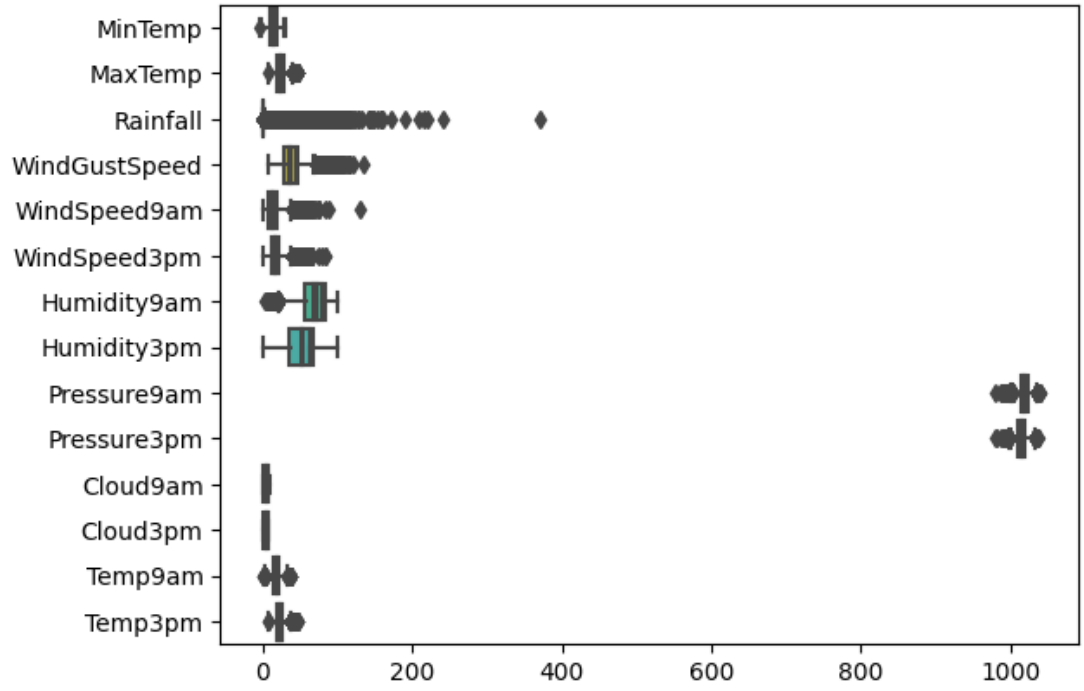
Name: Musab Abuasi ID: 1190954
Name: Rayyan Asia ID: 1192126

Date: 2/7/2023

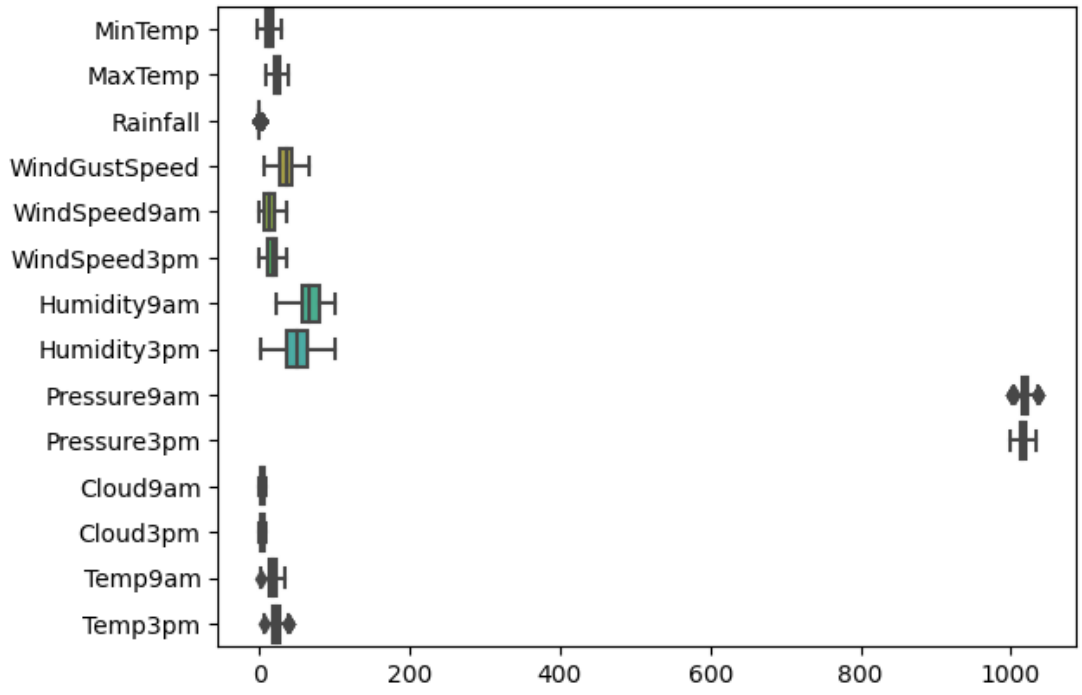
Introduction:

In a time where artificial intelligence and machine learning is practically dominating the world, we have been given the task to create a model given a dataset. After looking at the given dataset, we decided to set our target classification to be whether it will be raining the next day or not. To make sure we didn't choose the wrong model, we built multiple models and compared their performances. The models we used were decision tree, logistic regression, neural network, k means clustering, and finally the support vector machine model.

- Pre-Processing/EDA:
  - Feature Scaling:
    - First, we took all numerical columns in the dataset and scaled them using z-score standardization. This is achieved by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation.
  - Removing Outliers:
    - Here we plotted our features using box plots and detected our outliers by visualization. Then we removed any values that fell below the 25th percentile minus 1.5 times the IQR or above the 75th percentile plus 1.5 times the IQR.
    - Before:



- After:



➢ Handling Missing Values:

- For numerical values, we either replaced them with the minimum value in the column or the mean value in the column
- For categorical values, we just removed those rows all together.

➢ Changing Categorical Columns to Binary and Ordinal Columns

- We just changed the categorical columns to binary or ordinal depending on their nature (Yes or No= 0 or 1, Region1, Region2… = 1, 2…

❖ Decision Tree:

➢ Decision Trees use learned rules to form a tree structure starting with a root node down to many leaf node each consisting of classification results depending on the input.

➢ Feature Selection:

- SelectKBest is a feature selection method in scikit-learn that selects the k best features based on a score function. The score function calculates the importance of each feature, and the k best features are those with the highest scores.
- The score function chosen was mutual information which is calculated by

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \left[\frac{p(x,y)}{p(x) \cdot p(y)} \right]$$

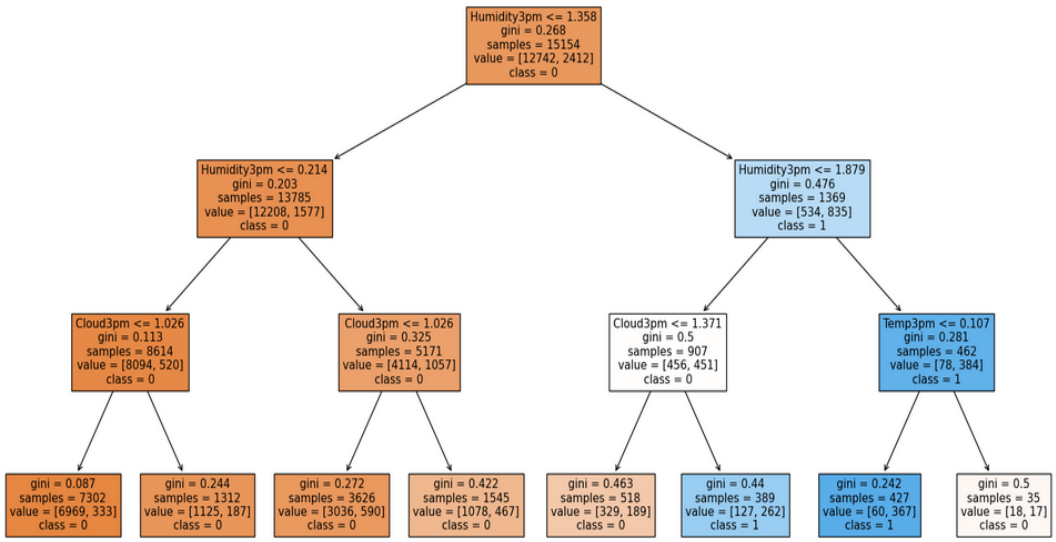
where p(x) and p(y) are the marginal probabilities of X and Y, and p(x, y) is the joint probability of X and Y. The summation is over all possible values of X and Y. The logarithm is base 2, so the mutual information is expressed in bits.

- Features Selected based on this: 'Rainfall', 'WindGustSpeed', 'Humidity3pm', 'Cloud3pm', 'Temp3pm'

➢ Accuracy measures:

Rain Tmr.	Precision	Recall	F1 Score	Total
No	.89	.99	.93	1771
Yes	.78	.30	.43	316

➢ Decision Tree Image:



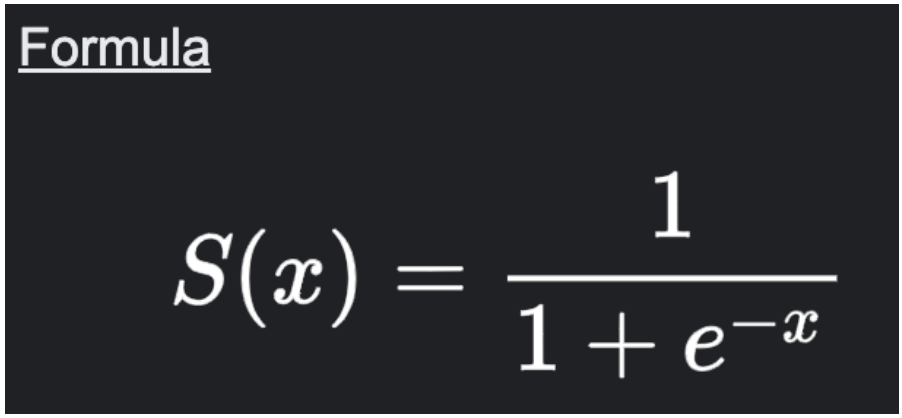
■

❖ Logistic Regression:

➢ After Deciding that one model wasn't enough, we decided to make more models, starting with logistic regression.

➢ Logistic Regression is a model that is similar to linear regression however results in very different outcomes.

➢ It uses the sigmoid function which creates the outputs from a range of 0 and 1:

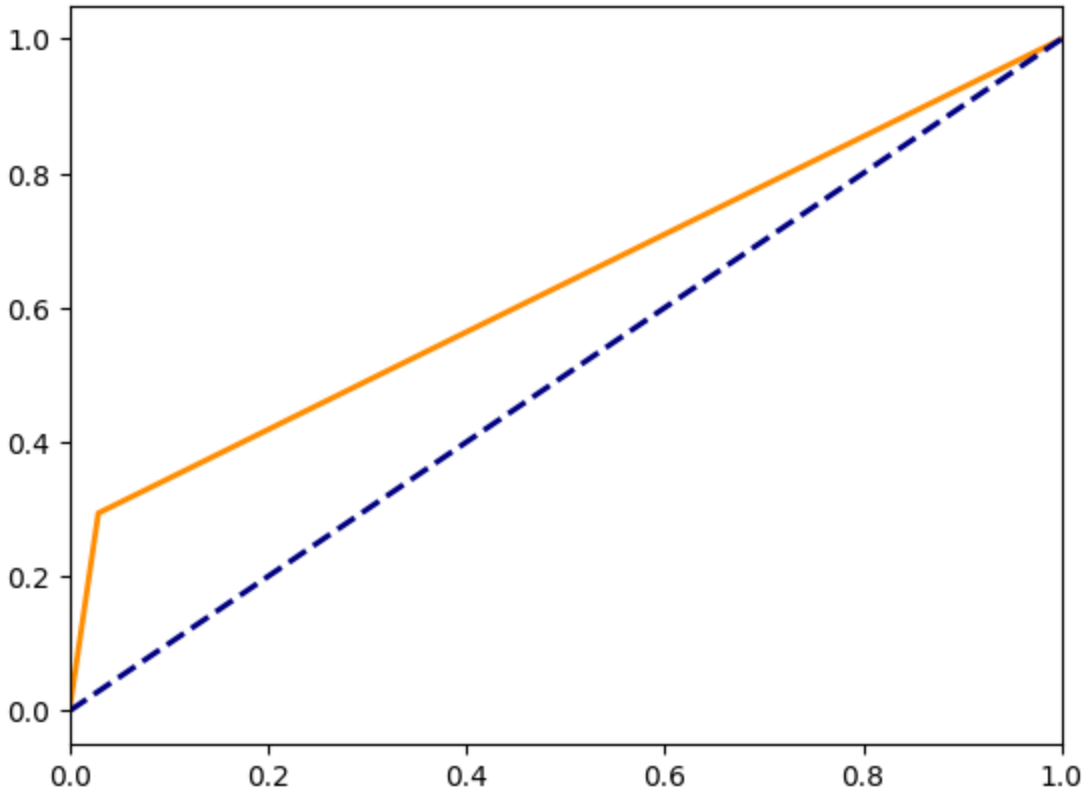
- 

➢ As well it uses a threshold value to classify the given unlabeled data.

➢ In our model we decided that 150 iterations was more than enough for the model to be created and tuned.

➢ In the end we had a model that had 88% accuracy.

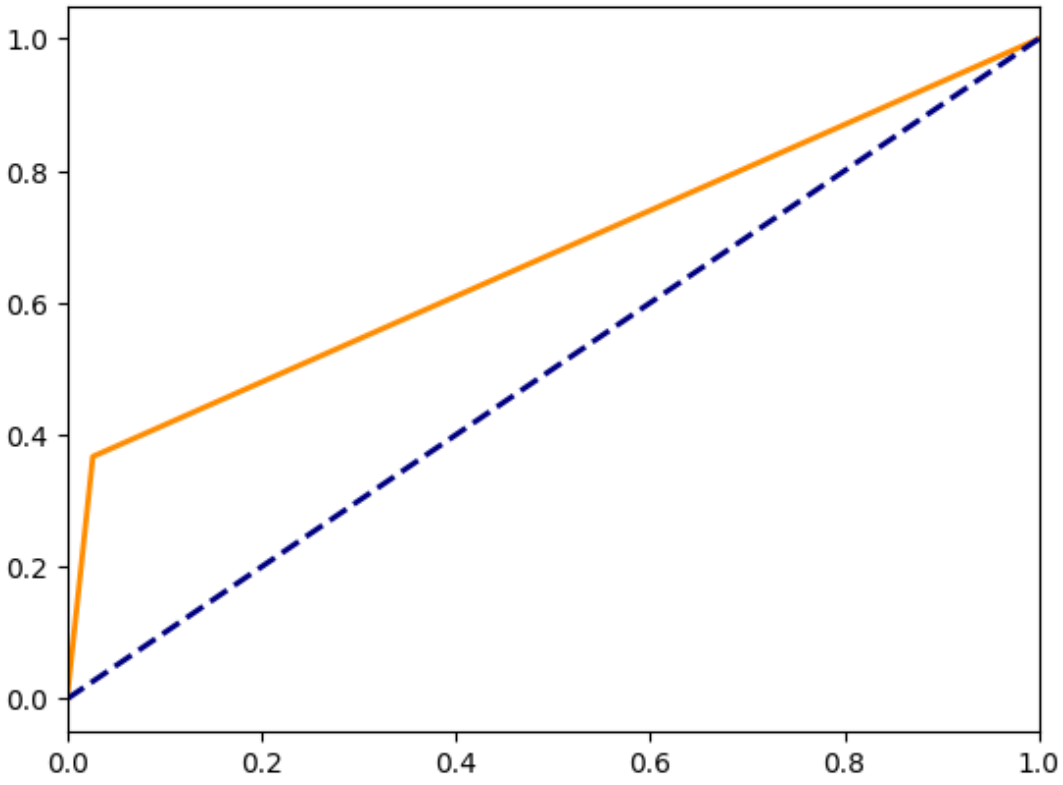
Rain Tmr.	Precision	Recall	F1 Score	Total
No	.89	.98	.93	1771
Yes	.70	.30	.43	316



- Neural Network:
  - It wouldn’t be a full machine learning if we didn't at least dabble in some neural networks.
  - Neural Networks is a machine learning algorithm modeled after the structure and function of the human brain.
  - It consists of interconnected nodes, known as artificial neurons, that process and transmit information through weighted connections.
  - Trained using large amounts of data and adjusting the weights to minimize the difference between the predicted output and the actual output.
  - Each neuron has an activation function that determines the outcome unto the next neuron.
  - The amount of hidden layers and amount of nodes in each layer directly affect the complexity of the model.

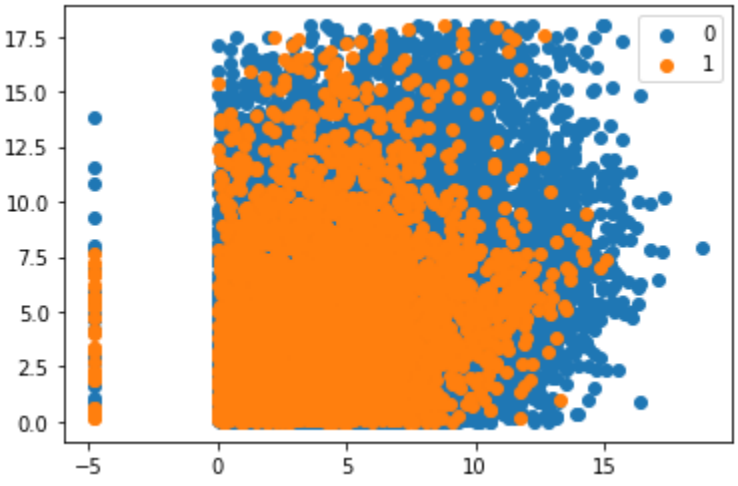
- ➢ If the model becomes too complex we may face the problem of overfitting.
- ➢ In our Neural Networks model, we gave it a dataset of 3081 data entries, and gave it 150 iterations.
- ➢ The model still wants to continue modifying its weights, however we noticed that 150 iterations is where progress relatively stops.
- ➢ In conclusion, The model ended with 88% accuracy just like our logistic regression and decision tree models.

Rain Tmr.	Precision	Recall	F1 Score	Total
No	.90	.97	.93	1771
Yes	.71	.37	.48	316



❖ K Means Clustering:

- As of the nature of the course, many things are results of trial and error. We decided to explore how the K Means would perform with our given tasks.
- Not to our surprise the model didn't prove itself to be the correct model for the job.
- It works by iteratively assigning each data point to the cluster with the nearest mean, and then calculating the mean of each cluster within every iteration.
- This process continues until the assignment of data points to clusters no longer changes, or if the developer sets an iteration limit. The result ends with K clusters describing the data.
- However in some cases, just like ours, the clusters overlap, making it hard for the model to predict the label correctly. Which makes sense since K means clustering is not typically used when you have a target value.
- Take a look of the clusters our model generated:



- As we could see the clusters are fairly the same
- The model gave an accuracy of 60%

	Precision	Recall	F1 Score	Total
No	.95	.55	.70	17626
Yes	.26	.84	.39	3242

- SVM:
  - Support Vector Machine is a supervised machine learning algorithm, the main idea is to find the hyperplane that best separates the data into separate classes.
  - In a two-class problem, like ours, the hyperplane is chosen to maximize the margin between them, as well as between the two closest data points of each class(support vectors).
  - For problems with more than two classes, the data is typically transformed into multiple two-class problems.
  - In some cases, such as non-linearly separable data, the data is transformed into a higher-dimensional space where a linear separation can be found.
  - This model proved accurate with 88% accuracy by using the Radial Basis Function Kernel, choosing a penalty size of 10 and a gamma value(The gamma parameter defines how far the influence of a single training example reaches) of 0.1
    -

	Precision	Recall	F1 Score	Total
No	.82	.50	.62	1624
Yes	.10	.34	.16	271

Models Summary:

	Rain Tommorow	Decision Tree	Logistic Regression	Neural Network	K Means Clustering	Support Vector Machines
Accuracy %		88	88	88	60	88
Precision	No	.89	.89	.90	.95	.82
	Yes	.78	.70	.87	.26	.10
Recall	No	.99	.98	.97	.55	.50
	Yes	.30	.30	.37	.84	.34
F1 Score	No	.93	.93	.93	.70	.62
	Yes	.43	.43	.48	.39	.16
Bias		0.132	0.132	—	—	—
Variance		0.009	0.004	—	—	—

Conclusion:

After trying all these different models, we believe the Neural Network showed the most impressive results amongst all the models, if you think otherwise then feel free to look at the statistics. The neural network model has practically the highest precision, recall, and F-1 score. As well as sharing the highest accuracy with the others. The reason we believe it outperformed the other models is because the neural network is composed of layers of nodes, with each node interconnected with the nodes from the previous layer and next one, all with individual weights on each connection. As well as the constant update of the weights through the iterations. Thus the model takes

practically almost all factors into account. Resulting with giving us the best performance.