Cairo University - Faculty of Engineering

Computer Engineering Department

Machine Learning

Rain Prediction in Australia

Team 6

Name	Section	B.N.
Abdallah Hussien	2	3
David Raafat	1	24
François Adham	2	10
Kareem Mohamed	2	11

Submitted to:

Eng./ **Yehia Zakaria**

Academic Year

2021 / 2022

Name	Workload
Abdallah Hussien	• Preprocessing • Visualisation
David Raafat	• Training • Tuning Hyperparameters
François Adham	• Preprocessing • Visualisation
Kareem Mohamed	• Tuning Hyperparameters • Training

Problem Definition and motivation:

Predict whether the next day will have rain in Australia or not.

Dataset:

Rain Prediction in Australia:

[Dataset]

Evaluation Metrics:

- Model Accuracy

Analysis and solution of the problem:

- **Preprocessing**
    1. First we show data shape and the features in it.
    2. Check if data has any NAN values.
        a. Remove all rows with NA in the dependent variable.
        b. Remove features with a high percentage of NA values.
    3. Then we see how many features are categorical in our dataset.
    4. Check how many unique values in these categorical features.
    5. After seeing that we removed the “date” and “location” features from the data, because they have many unique values and don’t give new information.

- **Visualisation**
    1. First we show a histogram of our target RainTomorrow which shows that data is not balanced.
    2. Show a histogram for each categorical feature and see how it affects our target variable.
    3. For numerical features we show the correlation between it and our target to see how they affect it.
    4. For more information and visualisation check the code notebook.

- **Extracting insights**

  We have five main features in our data related to whether state which we analysed independently :

  1. **Temperature (MaxTemp, MinTemp, Temp9am, Temp3pm)**
     a. After seeing the correlation between each one of them with our target, we found they don't affect it significantly.
     b. After searching we found that the difference between temperatures can affect rain probability, so we tried to construct two new features "DiffMinMax" and "DiffTemp" which were more correlated with our target.
     c. We left only "DiffMinMax", because it's the most correlated one with our target and removed "DiffTemp" because it highly correlated with "DiffMinMax".
  2. **Pressure (Pressure3pm, Pressure9am)**
     a. After seeing the correlation between each one of them with our target, we found they both affect it.
     b. We tried to do the same as temperature but we found that pressure difference doesn't affect rain probability.
  3. **Humidity (Humidity3pm, Humidity9am)**
     a. After seeing the correlation between each one of them with our target, we found they both affect it, but 3pm has almost the double effect of 9am.
     b. We tried to do the same with temperature but we found that Humidity difference affects rain probability slightly.

4. **Wind Speed (WindSpeed9am, WindSpeed3pm, WindGustSpeed)**
    a. After seeing the correlation between each one of them with our target, we found that only “WindGustSpeed” effects correlated to our target.
5. **Wind Direction (WindDir9am, WindDir3pm, WindDirSpeed)**
    a. These are categorical variables which we visualise its effect on our target and show that there are many outliers.
    b. Then we convert them to numerical variables to see its correlation with or target.
    c. After seeing the correlation between each one of them with our target, we found that no one of them affects our target.

- **Model/Classifier training**
    1. First we split the independent variables together and the dependent variable alone.
    2. Then we split our data into 60% training data, 20% validation data and 20% testing data.
    3. Our problem is to predict a binary (Classification) variable (RainTomorrow), so we used different classifiers to evaluate performance as follows:
        a. Support Vector Machine (SVM)
        b. Multilayer Perceptron (MLP)
        c. Logistic Regression
        d. K Nearest Neighbour (KNN)
        e. Random Forest
        f. Decision Tree

4. For each classifier we applied a grid search to tune hyperparameters on the validation data.
5. After hyperparameters tuning, we train each classifier with training data.
6. After training each model, we test it by the 20% testing data and record the accuracy as follows.

5- Results and Evaluation

Model	**Accuracy**
Random Forest	84.64 %
Multilayer Perceptron (MLP)	82.92 %
Support Vector Machine (SVM)	79.8 %
K Nearest Neighbour (KNN)	83.84 %
Logistic Regression	81.43 %
Decision Tree	77.44 %

6- Unsuccessful trials

1. After Seeing that data is not balanced we tried to make downsampling/oversampling but that affects the accuracy negatively
2. We tried to choose the best 3 features only and work with it but that decreases the accuracy, and that makes sense because we have selected only features that affect our target from the beginning.

7- Conclusion and Future Work

1. We can notice that data is not very good and needs many improvements so we can collect new and better data.
2. We can find more clean data entries in the dataset by noticing new days and recording its information to make better predictions in the future.
3. We can ask experts to help us to find more important features to the dataset which can affect rain probability.