Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Program: Bachelor of Science in Computer Science and Engineering

Course No. : CSE4108

Course Title : Artificial Intelligence Lab

Title : Report of the project “Rain Prediction”

Submitted to :

Mr. Md. Siam Ansary
Lecturer, Department of CSE, AUST.

Ms. Tamanna Tabassum
Lecturer, Department of CSE, AUST.

Submitted by,

Safwan Muntasir,	180104084
Mashfiq Rahman,	180104087
Arifur Rahman Jawad,	180104097

Introduction: In this project, we will predict the chance of raining the next day by training classification models on a constructed dataset. We collected the data from Australian climate report websites [1][2]. We preprocessed our collected data and made it convenient for our target. After dataset collection, we trained our model with traditional machine learning algorithms in order to predict whether it's going to rain the next day or not. In the 'performance table' section we have compared the performance of each algorithm and concluded the best classifier model for our task.

Construction of Dataset: From weather forecasting and climate report websites[1][2] we collected a huge collection of data consisted of 22 independent variables(columns) to predict the chance of the next day's rain (last column). We deleted most of the columns and the kept 8 independent variables that might be the most important variables to predict the result. The columns are :

- Minimum Temperature
- Maximum Temperature
- Rainfall
- Evaporation
- Sunshine
- Wind Speed
- Humidity
- Rain Today

And the final column is named 'RainTomorrow'. We kept total 366 rows to scrutinize our model. Finally, we deleted the column name row to build our concluding dataset. There are 300 0s(not raining) and 66 1s(raining) in the 'RainTomorrow' column.

Methodology: We implemented five machine learning classifiers to train our model. A brief discussion on our used algorithms is stated briefly.

1. **Random Forest Classification:** The first model is used in my project is Random forest classification. For the training, 80 percent of data is used and the rest is used in the test set. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes

the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

2. **Naive Bayes:** Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. For the training, 80 percent of data is used and the rest is used in the test set.

3. **SVM:** "Support Vector Machine" (SVM) is a supervised machine learning algorithm. that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. For the training, 80 percent of data is used and the rest is used in the test set.

4. **Logistic Regression:** Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique. For the training, 80 percent of data is used and the rest is used in the test set.

5. **Decision Tree Classification:** Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

Performance Table :

After implementing the experiment on different algorithms on our dataset using python and scikit learn library we gained the following results :

Name of algorithm	**Accuracy**	**Recall**	**Precision**	**F1 score**
Random Forest	**0.80**	**0.08**	**0.25**	**0.12**
Naive Bayes	**0.76**	**0.15**	**0.22**	**0.18**
SVM	**0.80**	**0.15**	**0.33**	**0.21**
Logistic Regression	**0.81**	**0.15**	**0.40**	**0.22**
Decision Tree	**0.70**	**0.09**	**0.21**	**0.10**

Table1: Performance table

Discussion: As we can see our dataset is an imbalanced dataset, we see a fluctuation in the f1 score despite the accuracies of all the classifiers are almost the same. We can conclude that Logistic Regression showing the best f1 score measure. Through preprocessing in a planned approach and constructing with balanced train set values we can increase the accuracy of each models.

References :

1.http://www.bom.gov.au/climate/data
2.http://www.bom.gov.au/climate/dwo/