Rain Prediction Based on Machine Learning

Ye Zhao[1,*], Hanqi Shi[2], Yifei Ma[3], Mengyan He[4], Haotian Deng[5], Zhou Tong[6]

[1]*The University of Ningbo Nottingham, NingBo, 315100, China, scyyz6@nottingham.edu.com*
[2]*Dana Hall School, MA, USA, hanqi.shi@danahall.org*
[3]*Hunan University, Changsha, China, 1250152196@qq.com*
[4]*Xi'an Jiaotong University, Xian, China, 1913004851@qq.com*
[5]*Chongqing BI Academy, Chongqing, 401120, China, 1559117929@qq.com*
[6]*Xi'an Jiaotong-Liverpool University , Jiangsu, China*
**Correspondence author: scyyz6@nottingham.edu.com*

ABSTRACT

Our purpose is to try to use machine learning algorithms to predict the weather of the next day, since whether it will rain tomorrow is a very important indicator of the weather. In order to find the most predictable attributes of rain, The researcher use line charts, matrix graphs, and scatterplot graphs for visualization and analysis. The researcher find that several pairs of attributes have a high degree of similarity and correlation. In the fitting stage, the researcher used simple models such as KNN, decision tree, and ridge regression to evaluate its basic prediction quality and found that the accuracy rate is around 0.78. Since in the visualization stage, the researcher found that the samples that rained today have a slightly higher probability of raining the next day, the researcher tried to use LSTM to analyze the impact of historical weather and found that the relationship is not strong. Finally, logistic regression turns out to have the highest accuracy of 0.85, followed by adaboost with an accuracy of 0.82. Whether it will rain remains unpredictable to some extent.

Keywords: *machine learning, Rain prediction, LSTM*

1. INTRODUCTION

Our goal is to predict whether it will rain the next day in Australia based on what the researcher learned in the field of data mining. And the researcher hopes to build high accuracy prediction models to deal with practical problem [1]. To achieve this goal, first the researcher finds a dataset containing almost 10 years of daily weather observations from many locations across Australia, with approach 142,193 lines of data. The data set has 24 weather attributes such as wind speed, humidity, temperature and so on. It's a huge enough satisfy our need of data. Second, the researcher think that climate change actually has certain rules in it. In China, there is something called the 24 Solar terms. It's twenty-four periods in traditional Chinese calendars that signifies some natural phenomenon. It is a relatively accurate prediction of China's climate change for a whole year. So, when the researcher looked at this data set, the researcher also wanted to see if there was any pattern in Australia's climate. And the researcher tried to find what attributes have effects on whether it rains the next day. Nowadays, climate-changing is a hot topic around the world, since serious disasters keep occurred in recent years. Climate change may be caused by natural internal processes or by continued anthropogenic changes to the composition of the atmosphere and land use. These all have big and small impacts on our planet. For example, the melting of glaciers around the world is accelerating, which means that millions of people are facing the threat of floods, droughts and reduced drinking water. And the extreme climate is patronizing the earth with unusual frequency, which is related to the global climate change background. Global warming is reducing food production, like droughts, water shortages, rising sea levels, floods, heatwaves and temperature upheavals all disrupt food production around the world. The fourth assessment report of the United Nations Intergovernmental Panel on Climate Change (IPCC) released in 2007 pointed out that climate change will lead to the extinction of a large number of species in the next six to 70 years. All of these are problems that climate change will bring, and humans should face them. The researcher wants to use what the researcher knows about how to process data and some machine learning models

to predict weather changes, a practical problem that people care in daily life.

The researcher also wants to figure out which attribute has closer relationship with whether it will rain the next day, using some basic viualization skills.

2. BACKGROUND

In the late 1930s, during World War II, the British innovative discovery radar could not only be used to monitor enemy aircraft, but also could receive echoes from raindrops of certain specific wavelengths (5-10 cm) [2]. This technique can be used to track and study individual showers and observe the precipitation structure of larger storms. This is the early use of scientific and technological means by modern scientists to predict the possibility of future rainfall, but it is clear that this can only predict rainfall in a short period of time. With the rapid development of technical scientific data and technology, scientists can begin to use geophysical knowledge and use a large amount of research data to make judgments. The first choice is to use machine learning in artificial intelligence. Senior January, Google engineer Jason Hickey introduced an application of machine learning to weather radar charts. The idea is to convert the radar chart into a "computer vision" problem that machines are good at. He tried to use a large amount of data to drive machines to learn physics principles from algorithms, and later used U-Net in neural neural networks (CNNs). In comparison, his machine learning algorithm is better than the three traditions-HRRR numerical forecasting, optical flow method (optical flow method) and persistent modeling (persistent model), in ultra-short-term changes [3]. When the experiment uses decision tree to predict rain in Australia, some data processing methods and models are needed. Below, the actual skills will be described.

2.1. Initial Stage

The Z-socre method standardizes the dataset. Before data analysis, the researcher usually needs to analyze (standardize) the data first and use the data analysis after the data. The z-score standard is based on the mean (mean) and standard value (standard deviation) of the original data for each indicator of the data. Each column says that all data is around 0, which is 1.

2.2. Encoding

In many machines learning tasks, features are not always continuous values, but may be personal values. In this way, for each feature, if it has m possible values, it becomes m binary features after individual hot coding, and these features do not affect each other, and there is only one activation at a time. So, the benefits of these data will become different categories. Therefore, the main benefits of solving data are: 1. Solving the problem of bad attributes 2. To a certain extent, the role of the characteristics of the universe

2.3. Balance the data

Category imbalance refers to the situation of the spherical bottle mouth of the training sample tasks of different categories in the classification. There are three commonly used situations, namely 1. sampling, 2. over sampling, and 3. threshold shift. This time the prediction is mainly due to the following reasons, that is, to understand some counter-example explanations, counter-example problems, and then learn. Following is visualization, which is processing data: it is observable and easy to observe when a clean data set after analysis is obtained, the next step is Exploratory Data Analysis (EDA). EDA can help discover data, and can also be used to find patterns, relationships, or anomalies to guide students' analysis. One of the effective start-up tools is the scatter plot matrix. The scatter plot matrix allows two separate distributions and the relationship between them. The histogram of the diagonal position allows us to see the distribution of each variable, while the scatter plot on the diagonal shows the relationship between the two. Another important thing is the line chart. The visualization form of many visualization tools is the line chart. By drawing the data changed by the tool into a line, the size and trend of the line chart can be intuitively changed. Set, especially those occasions where the trend is more important.

2.4. Models Applied

Logistic Regression is a machine learning method used to solve classification problems and is used to estimate the possibility of something. Logistic Regression and Linear Regression are both a generalized linear model. The simple process is to first write the maximum likelihood function and perform logarithmic processing. Then use the gradient descent method to find the minimum value of the cost function. The KNN classifier, as a relatively easy-to-understand classification algorithm in supervised learning, is often used in various classification tasks. The core idea of the KNN model is very simple. It calculates the Euclidean distance between each sample point of the test set and each sample in the training set, and then takes the K nearest Euclidean distance. Point (k is the number of neighbors that can be delineated artificially, and the determination of K will affect the results of the algorithm) and count the category frequencies of the K training set sample points and convert the category with the highest frequency to the test sample point Forecast category [4]. The method of LSTM and Ensemble Algorithm—boosting bagging are also used.

3. VISUALIZATION

3.1. Line Charts

A line chart is a visualization tool that helps demonstrate the distribution of features by classes, which in our experiment, is whether it will rain tomorrow. In our construction of a line chart, the researcher aims primarily to determine the important features in classification, the features in which the two groups have the most distinctions. Some challenges appear when constructing a line chart, and the researcher use several methods to overcome them. First, the dataset contains some categorical data, which cannot be represented in a quantitative way directly. The researcher encodes them with LabelEnoder from the Scikit-learn package. This function allows us to assign each label in one variable an integer from 0 to the number of labels minus one. Because RainToday is binary, either yes or no, encoding it actually gives a clearer representation. Another problem is the different ranges of data that makes it hard to effectively present all attributes in one graph. The researcher draws the graph in Figures that shows the ranges for each feature. Based on their ranges, the researcher divides the features into six groups:

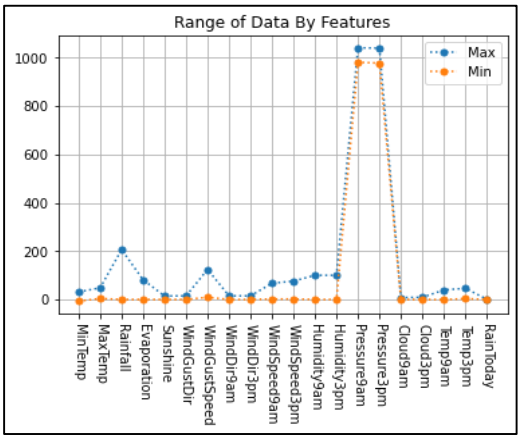


Fig. 1 This graph shows the ranges of data in each feature.

Group A includes MinTemp, MaxTemp, Temp9am, Temp3pm

Group B includes Sunshine, WindGustDir, WindDir9am, WindDir3pm

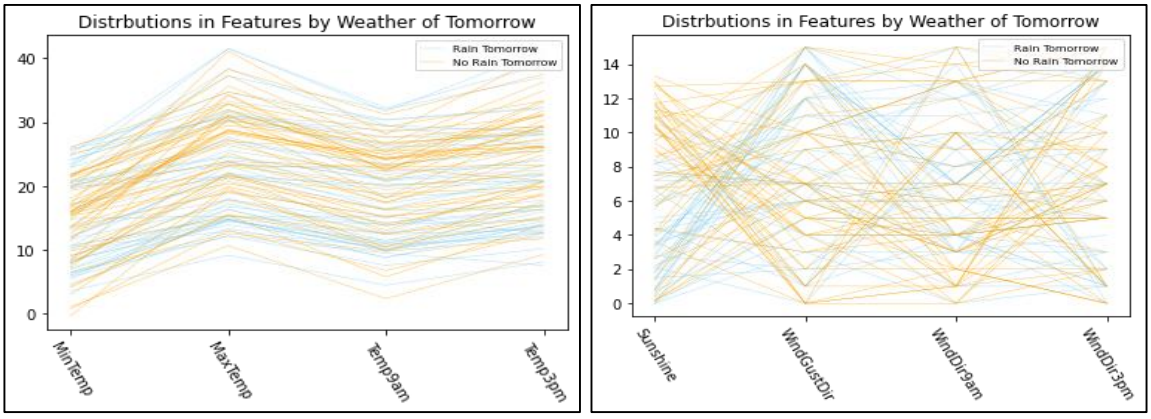
Group C includes Rainfall, Evaporation, Humidity9am, Humidity3pm

Group D includes WindGustSpeed, WindSpeed9am, WindSpeed3pm

Group E includes Pressure9am, Pressure3pm

Group F includes Cloud9am, Cloud3pm, RainToday

Lastly, was the problem on unbalanced data. The researcher have many more cases of no rain tomorrow than cases of rain tomorrow. To make the data more balanced, the researcher draw 50 random samples without replacement from each class. This may have caused inaccuracy because the data the researcher use to graph is just part of the original data. However, since they are random samples, the researcher believe they still can represent the whole dataset in some degree. After graphing each group of features, the resulted graphs are as follow (Fig 2).



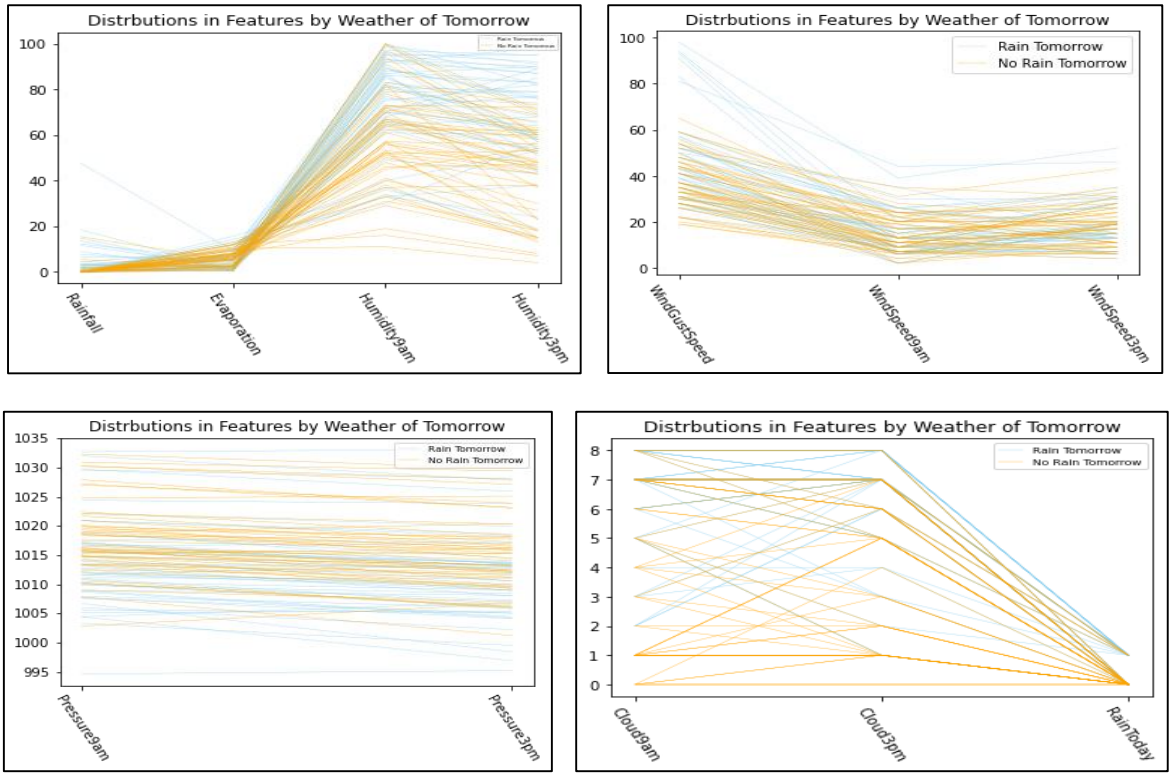


Fig. 2 This group of line charts shows the distributions in features by weather of tomorrow.

Comparing the features in Figure 2, the researcher notices that there is no significant distinction in temperature between the two classes. Similarly, in figure 2 there are no obvious patterns of wind directions. Nevertheless, there is a clearer distinction in sunshine: if it is going to rain tomorrow, the sunshine is going to be lower than if it is not going to rain. Rainfall is just the opposite to sunshine. Moreover, if the next day is a raining day, the current day will be more humid in both the morning and afternoon. On the contrary, the pressure will be lower. Cloud does not display a great distinction, but if cloud is really low, around 0 to 1, it is not likely to rain on the next day. For rain today, if it true, the value is 1, or it is 0. Rain today seems to be a key determining feature of tomorrow's weather, for if it is raining tomorrow, there is higher chance that today rained as well, and if it is clear tomorrow, there is higher chance that today was clear. Further, there seems to be more cases of rainy days followed by clear days than that of clear days followed by rainy ones. This can be explained by the fact that clear days are indeed more common weathers than rainy days. From the line charts, the researcher concludes that the features that contribute more to determining the class are sunshine, rainfall, humidity, pressure, cloud, and whether it rained today. The researcher acknowledge that our result contains certain limitations due to the sampling. Other visualization methods can complement the findings in line charts.

3.2. Matrix Graph

Matrix graphs are effective when analyzing the representativeness of attributes for datasets. Reduce unnecessary features in the training process can be helpful for increasing accuracy and saving running scales, the latter thus could be varied in index level. The researcher uses a matrix graph to observe the features of attributes and hope to find valuable information. According to the matrix graph as shown in Figure 1.2.1, attributes max temperature, rainfall, sunshine, Humidity is more important for classification, which can be the identity by color distributions in the matrix heatmap. At the same time, other attributes, min temperature, evaporation, and others are less unified on the aspect of color. Relationship between attributes, Linear relationship has also been reflected by couples of these attributes. The numbers of “Maxtemp” and “rainfall” seems to have a negative correlation, the opposite phenomenon, positive correlation exist between two pressure, two "cloud", and two temperatures of 9 a.m. and 3 p.m..

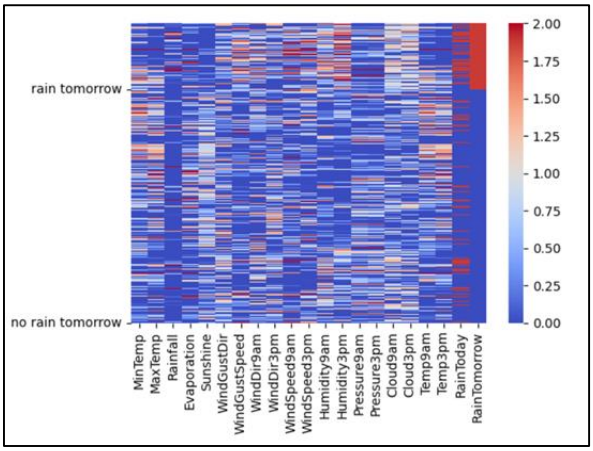


Fig.3 : There is no significant connection between "RainToday" and "RainTomorrow", which obeys common knowledge. In conclusion, "MaxTemp", "Rainfall"," sunshine" might play catalyst roles in the model prediction, the group consider adjust their weights to get better accuracy marks.

4. M0DELS

4.1. Decision Tree

Decision tree is a basic classification and regression method. This article mainly discusses decision trees for classification. The decision tree model has a tree structure. In a classification problem, it represents the process of classifying instances based on features. It can be considered as a collection of if-then rules, or as a conditional probability distribution defined in feature space and class space. Its main advantage is that the model is readable, and the classification speed is fast. When learning, use the training data to establish a decision tree model according to the principle of minimizing the loss function. When forecasting, use the decision tree model to classify the new data. Decision tree learning usually includes 3 steps: feature selection, decision tree generation and decision tree pruning. These decision tree learning ideas are mainly derived from the ID3 algorithm proposed by Quinlan in 1986 and the C4.5 algorithm proposed in 1993, and the CART algorithm proposed by Breiman et al. in 1984.

4.2. Decision Tree algorithm process

(1) Initialize the root node and have all the attributes in the sample

(2) According to the current partitioning rules, select the best attributes for partitioning, and get two sub-nodes

(3) Return to (2) and continue to divide the sub-nodes until:

a All samples of this node are of the same type

b Reach the pre-set conditions, such as: reach the maximum depth of the tree, and the number of samples is less than the minimum number required by the split node. Or feature has been used up.

4.3. Pros and Cons

Advantages

(1) The decision tree is easy to understand and explain, can be visualized and analyzed, and it is easy to extract rules.

(2) It can process nominal and numerical data at the same time.

(3) When testing the data set, the running speed is relatively fast.

(4) The decision tree can be well extended to large databases, and its size is independent of the size of the database.

Disadvantages

(1) It is difficult to deal with missing data.

(2) Prone to overfitting.

(3) Ignore the correlation of attributes in the data set.

(4) When the ID3 algorithm calculates the information gain, the result tends to be more numerical.

4.4. Adjusting the parameters

Here I plan to choose the following parameters: Criterion, splitter, max_depth, min_samples_leaf, min_samples_split, max_features to adjust the classifier. Because these parameters can be used to adjust and improve the effectiveness of decision tree best. As in the classification process of decision trees, the classification criterion directly affects the shape of the entire tree, and the information gain in the ID3 algorithm has shortcomings that may cause over-fitting. So, the quality of criterion and splitter will directly affect the structure of the tree. Moreover, according to the characteristics of decision trees that are prone to overfitting, approaches to limit branching are necessary. The depth of the tree can limit overfitting at a macro level, and the minimum leaves and minimum branches can determine the stopping criteria in a more detailed way.

The first parameter to adjust is a criterion which is the basic and the soul of a decision tree. It is the function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain. By, analyzing the chart, the researcher can find that using information gain is a more effective way to classify. Since information gain has better expressive power for more chaotic collections, but for purer collections, the Gini index will distinguish more clearly. So it may show that our data has a relatively high level of confusion.

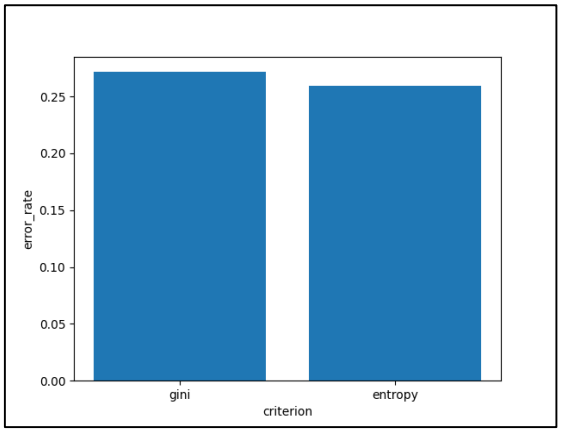


Fig 4: Next comes the splitter, which is the strategy used to choose the split at each node. 'Best' is to choose the best split, and 'Random' is to choose one from the best splits randomly. The researcher can see that using best is more stable and overall, its performance is better than random. So, the researcher decided to use the 'Best' as the splitter. Assuming that maybe it is due to the random_state of the computer.

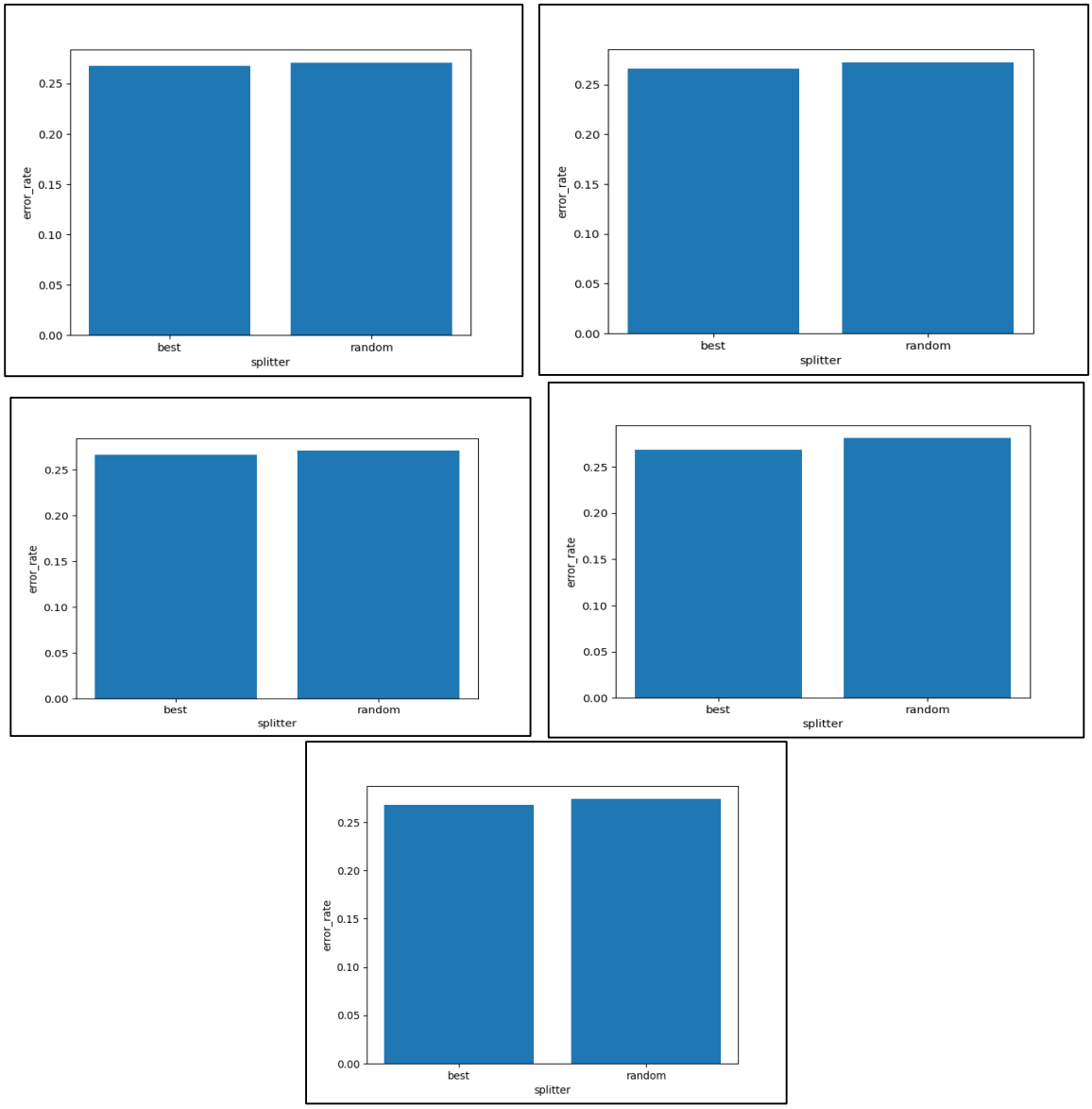


Fig. 5: Then the researcher decided the max depth of the tree. Because if a tree is too deep, it is easy for it to overfit the train set. So, the researcher need to find a suitable depth to find the compromise of overfitting and underfitting. And eight is the best depth.

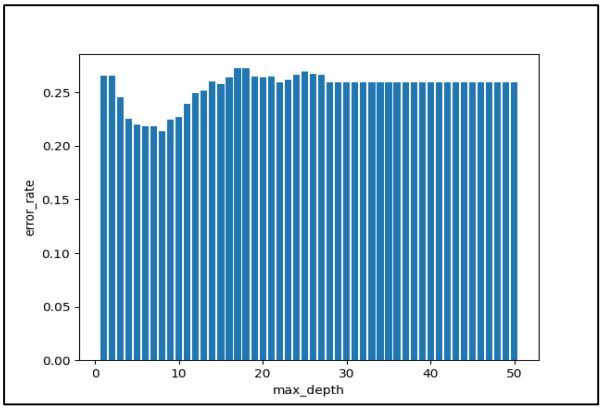


Fig. 6: About min_samples_leaf and min_samples_split. Min_samples_leaf is the minimum number of samples required to be at a leaf node. This is also used to prevent overfitting, as a node contains too few points means that it can even be an outlier and using it to predict is useless. But too many points means that it may still have the potential to split more precisely. Min_samples_split is the minimum number of samples required to split an internal node. It is a little familiar with the min_sample_leaf, if one node contains less than the number of you required, it will be a leaf.

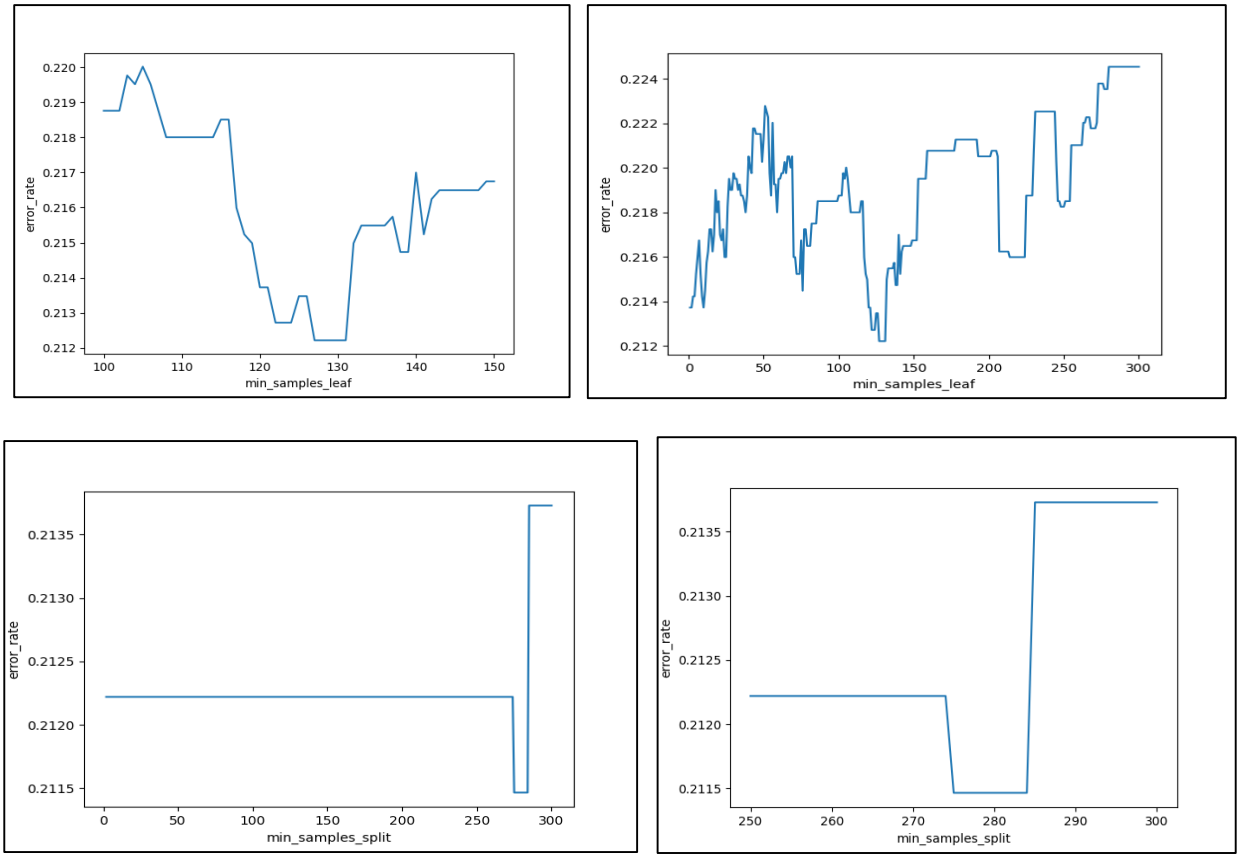


Fig. 7: At last, is the max_features which is the number of features to consider when looking for the best split. It limits the number of features considered when branching, and features that exceed the limit will be discarded. This is kind of a way to reduce dimensionality. But the method is too violent, without knowing the importance of each feature in the decision tree, forcibly setting this parameter may lead to Insufficient model learning. And the chart shows that if the features are less than 20, its overall performance is worse. But the researcher can see that a lower max_features can even lead to a lower error_rate, such as 16. Therefore, the researcher can find that there are many attributes that are not useful and even distort the prediction, so our inference: only by using a few attributes can predict whether it will rain tomorrow is correct.

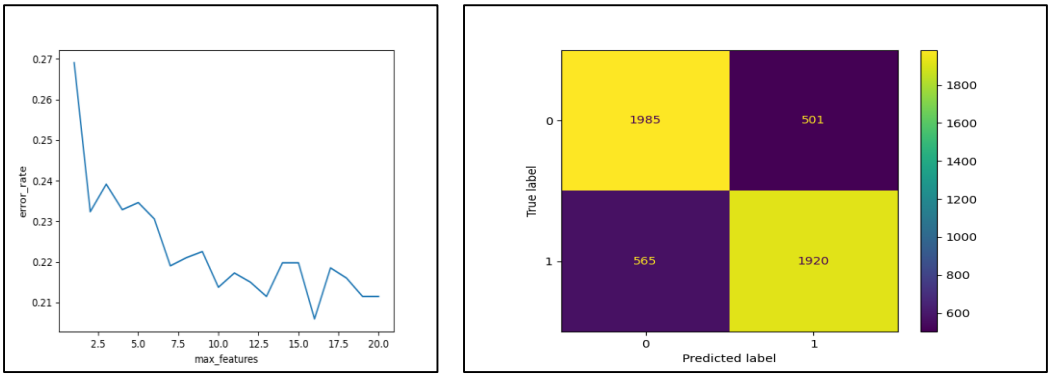


Fig. 8: Here is the overall performance of the decision tree. Its accuracy rate approximately is 0.78757. The confusion matrix is as follows. 0 means no rain tomorrow, and 1 means that it will rain tomorrow.

5. LOGISTITIC REGRESSION

Another model the researcher uses to predict the weather is logistic regression with penalties from the Scikit-learn package. The history of logistic regression traces back to the early 19th century (Cramer). Penalties, on the other hand, help decrease the complexity by adjusting the coefficients of inputs. If there is no penalty, the parameter for penalty is "none," and the model does not assign to different features different weights. In a journal published in 1996, Robert Tibshirani and his colleagues proposed the "least absolute shrinkage and selection operator" with the abbreviation Lasso. Lasso regression allows the model to remove some less important features completely by reducing their coefficients to zero [5]. The penalty parameter for Lasso regression is "l1." Ridge regression also produces weighted coefficients, and its penalty parameter is "l2." In this model, another hyperparameter the researcher experiment with is C. As c decreases, the regularization strength increases. For the solver parameter, the researcher chooses "saga" for all the experiments because it works for most combinations of parameters.

5.1. Parameter Adjustation

The dataset is split into three groups: 56% is training data, 24% is validation data, and 20% is test data. In the first test, the researcher uses Lasso and Ridge regression with eight different C values: 0.5, 0.1, 0.07, 0.05, 0.01, 0.007, 0.005, 0.001. As shown in Figure 2.2.1, in Lasso regression, the accuracy on the training dataset remains around 0.852 out of 1.00 when c is 0.5, 0.1, 0.07, and 0.05. However, it decreases at c = 0.01, drops a little more at c = 0.005, and drop significantly at c = 0.001. In the ridge regression result, displayed in Figure 2.2.2, the accuracy on the training dataset remains similar from c = 0.5 to c = 0.005, with relatively lower values at c = 0.05. It also drops significantly at c = 0.001. An interesting phenomenon is that in both regressions, the model predicts the validation dataset better than the training dataset. Based on this experiment, the researcher is able to select the optimal parameters for our Lasso regression model and Ridge regression model, which allows us to further choose the better penalty. For l1 penalty, c = 0.1 seems to work the best on the training dataset; for l2 penalty, c = 0.07 and c = 0.01 work equally well, but because the validation dataset is predicted more accurately at c = 0.01, the researcher decide this is a better c value.

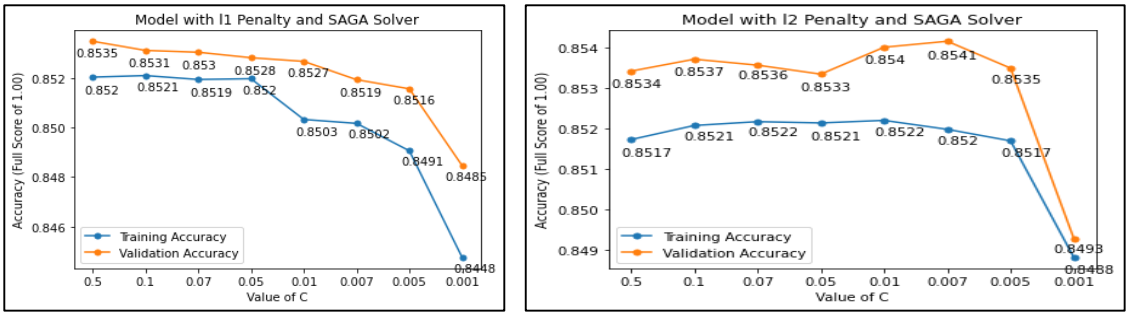
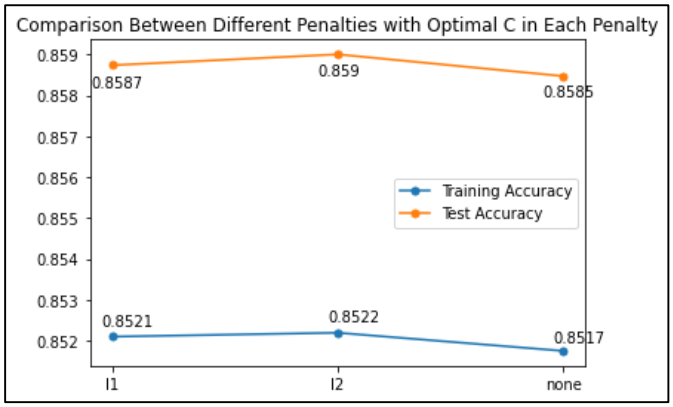


Fig. 9: In the next experiment, the researcher compares the parameter combinations between: l1 penalty and c = 0.1; l2 penalty and c = 0.01; no penalty. The result is presented in figure and again shows the intriguing feature of higher accuracy on validation dataset. The graph indicates that l2 penalty with c = 0.01 is the most accurate and no penalty is the least accurate in prediction. It can further be inferred that because ridge regression works better than Lasso regression and regression with no regularization, all features are used to determine the weather of tomorrow, yet some are less important than others. Because c is relatively small, there is a huge discrepancy between the contributions of the important attributes and of the unimportant attributes in this classification.



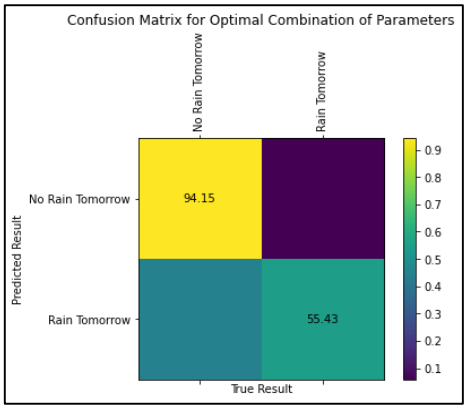
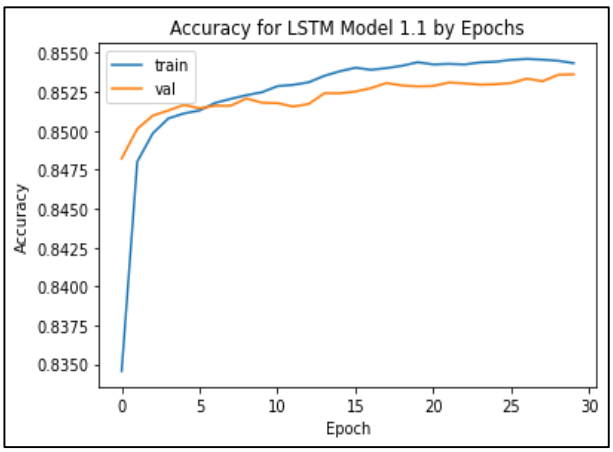


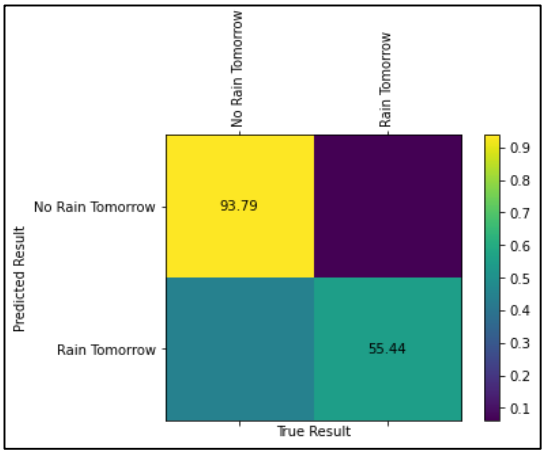
Fig.10: The confusion matrix for l2 penalty and c = 0.01, our optimal parameter, shows that the model much more accurate in predicting not raining circumstances than predicting raining circumstances. One possible explanation can be that because the model consists of an imbalance amount of raining and not raining classes, it learns how to classify not raining circumstances better.

5.2. LSTM

The dataset is not only organized by location in Australia but also by dates. This inspires us to postulate that weather predictions may be dependent on the previous weather of the week - there can exist a subtle pattern of rain in a period of time. To test this hypothesis, the researcher decides to use Long Short Term Memory, a type of recurrent neural network introduced by Sepp Hochreiter in 1997[6].

The unprocessed dataset, though containing much more data, is not perfect. There are void numerical and categorical values. For void numerical values, the researcher fills them with the mean of all their column. For categorical variables, the researcher uses logistic regression to make predictions and fill the voids with the results. Then, the researcher design three models to test the hypothesis. Model 1.1 uses data from the last seven days without data about whether it rains on each day to predict the rain status for the next day. Model 1.2 uses data from the last fourteen days without data about whether it rains on each day to predict the rain status for the next day. Interestingly, in both models, the prediction is not as accurate as the researcher have imagined. In fact, the two LSTM models perform similarly to the ridge regression model. It is nonetheless pleasing to see that in model 1.2 the prediction for no rain tomorrow is slightly more accurate than both the logistic regression and LSTM Model 1.1. Although a 2% increase may be ignorable, it still implies that perhaps rain displays a pattern in longer periods of time, but past data does not drastically improve the model.





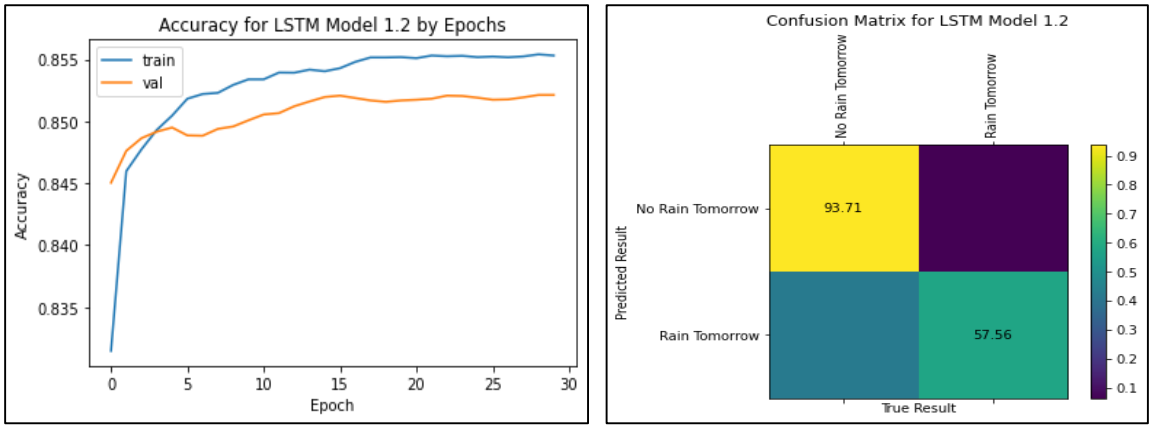


Fig 11: In the line chart and confusion matrix, the researcher has noticed that rain today is very different between if it is going to rain tomorrow or if it is not. To test if the prediction is directly related to this feature, the researcher excludes this feature in LSTM model 1.1 and model 1.2. LSTM Model 2 however includes this feature. It is a model that predicts the rain status of the next day based on environmental data from the last seven days, including if it rained on each day. Again, the result is not so much as the researcher have pictured it, and the accuracy is almost the same as model 1.1. Whether it rained is not as important as the researcher have assumed. It is just as likely to predict whether it will rain correctly with environmental data. Perhaps, even though the changes of rainy days and clear days are visible, they are just a visual presentation of rain, whereas other subtle environmental factors can also indicate the likeliness of rain.

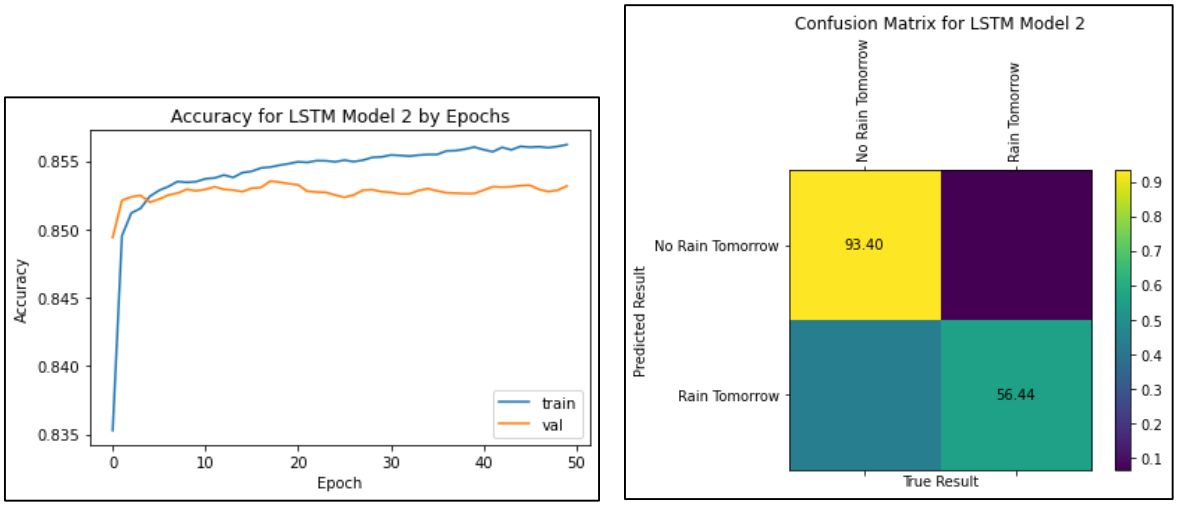


Fig 12: The lstm model, no matter using seven days of pat data or fourteen days of past data, is not much different from the logistic regression model. The implication here is that whether it will rain is not dependent on whether it rained on the past days, but rather can be predicted from just data on the previous day. However, because the initial data is incomplete, the researcher have filled void data with results based on statistics, not what is recorded. This can compromise the accuracy of the whole dataset, and thus lower the accuracy. Moreover, knowing if it rains on each day does not help the model improve, suggesting that what the researcher see as rain is just a direct representation of what has happened in the environment, for instance, the change in humidity and temperature.

5.3.Ensemble methods

Both bagging and boost algorithms are inspired by Breiman, who found gains in accuracy by combined several weak classifiers by sequencing way with weights calculated by evaluating their results [7]. For the purpose of forming more accurate models, the researcher decides to use ensemble algorithms. This Chapter will introduce several ensemble algorithms and their prediction result followed by simple analysis.

5.4. Boost Algorithm

Boosting is an ensemble method, which can promote the prediction rate of a machine learning algorithm, by train weak learners in a sequential way, then correct the model by the correct from the previous one [8]. Algorithms that adaptively change the distribution include Adaboost, Logist Boost and gradient Boosting, which can significantly lower the error rate on decision tree and other base classifiers [7]. The researcher chooses the Adaboost to decrease the error rate.

5.5. AdaBoost based on Decision Tree

The first parameter that needs to be considered is the iteration number for the method. It plays an important part in determining how big the scale of weight adapting. Excessive iteration takes huge time complexity, while a small iteration number may limit the performance of the model.

To find out the best fitting iterations, the line chart in Figure 2.4.1 is helpful.

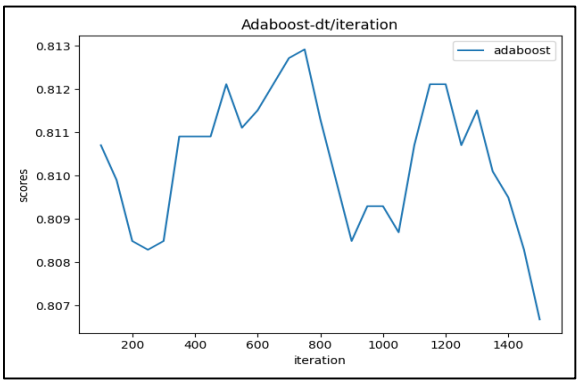


Fig. 13: In this Figure 2.4.1, axis_x represents the iteration times, axis_y represents the accuracy of the formed models. The score raised along with a few floating, then reach peak at 750 times iteration, then the model seems to get start to be overfitting. As a conclusion, 750 might be the suitable number.

Learning rate is also an indispensable role in AdaBoost methods. The influences from learning rate are, as shown in the following line charts.

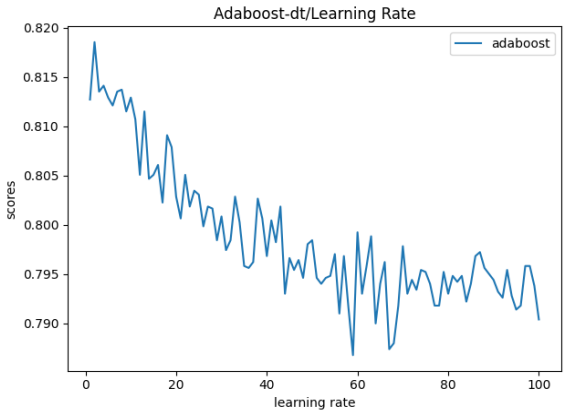


Fig. 14: The accuracy floating from 0.790 to 0.820, at the same time, the best score occurs in 0.02 learning rate. A guessing for the unsteady line is that the algorithm is iterated enough of time, thus the accuracy affected less from the learning rate.

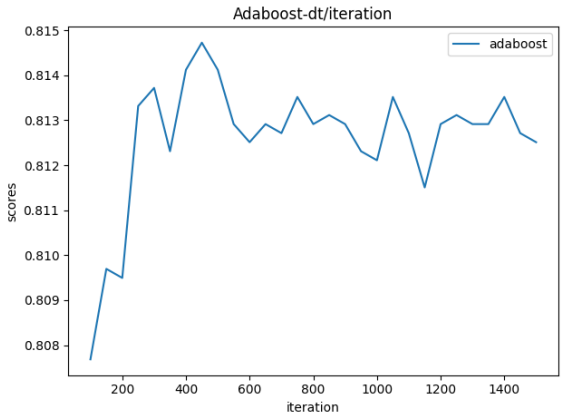


Fig. 15 To get rid of the possible influence existing between parameters, the researcher adjust the iteration times with learning rate 0.02. The figure shows that when AdaBoost using learning rate 0.02, The best result, 82%, comes with 250 iterations.

5.6.Bagging Algorithm

Bagging algorithms are methods that do not adaptively change the weight of the weak classifiers, Bagging, means bootstrap aggregating, is one of the simplest and earliest ensemble methods, with wonderful performance. The deference between bagging and boosting algorithms mainly reflects in sampling scheme. Bagging methods split the training datasets by sampling, then apply each to base model, second, synthesizing the predicted results of all base models to get the final predict result [7]. Bagging methods simply merge the decisions of simple learner by votes, and these trees are expected to be similar and to implement the same classification for each test instance (Subasi and Qaisar). The main advantage of Bagging is balancing the instability of base models. In each stage, the method alters the training datasets to decimating certain instances and interpolating others.

The researcher use bagging algorithms to promote the accuracy scores of Decision Tree.

5.7.Bagging based on Decision Tree

The number of the base learner affecting the accuracy of the final model. In extreme case, a classifier consists of infinite different weak learners which combined by bagging is able predict with 100% accuracy.

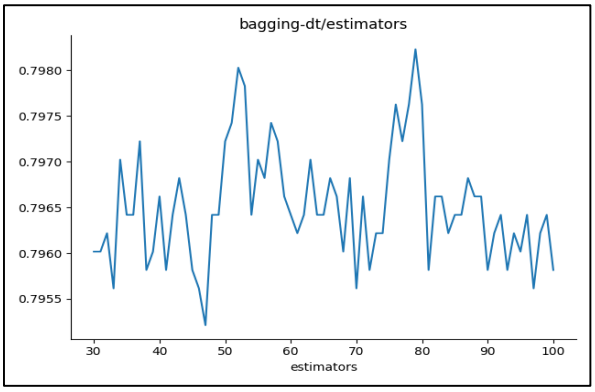


Figure 16: 75 decision trees are the best choice for the bagging in range one to a hundred. Surprisingly, the value of predication result is not changing intensely. To found out the possible reason for this phenomenon, a line chart has been created.

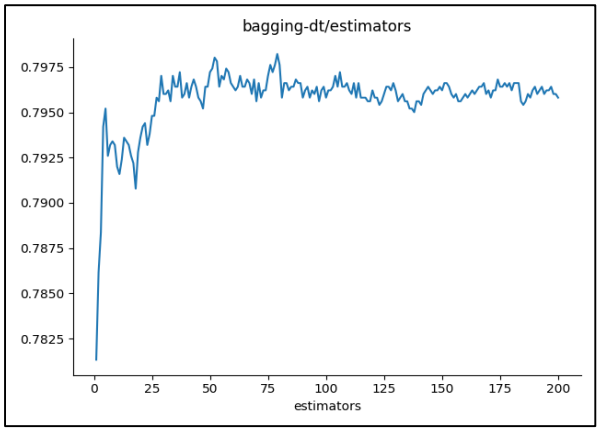


Fig 17: The accuracy reached a dropped after 75, at the same time, the accuracy keep floating in a small range, which indicate that this line will approach to straight line, since the bagging is work as an mean of all the weak leaner, which means the accuracy will reach a stable number as the increasing of the weak learner in same type.

5.8.KNN Classifier

The KNN classifier is based on a KNN algorithm, which is the simplest algorithm of the area of machine learning. The main idea of a KNN classifier is that the category of every simples can be delegated by its k-nearest neighbors. The KNN classifier mainly depends on the surrounding limited adjacent samples, rather than on the method of discriminating the class domain to determine the category. Therefore, KNN method is more suitable than other methods for the sample sets to be divided with more overlapping or overlapping class domains. According to the scatter plots metrics, most plots are very close to each other, and the plots of different category have large area overlapping, indicating that a KNN method is a suitable algorithm for our topic. In addition, many classification problems are successfully solved by KNN algorithm. Using four different manners to train KNN method to detect DTMF tones perturbed by noise, frequency and time variations. Using KNN method to applicate in handwritten digit recognition and the recognition rate reaches 94.6% [9]. Using KNN method to determine accurate classification of proteins and the classification accuracy is above 95% [10]. Above all, KNN algorithm is a simple but practical method for the classification problems. The researcher wrote KNN algorithm in python, spilt the dataset into training set and testing set, depending on which to train the model. Finally, use this particular model to predict whether it will rain tomorrow.

5.9.Parameter adjusting

Since every simple is represented by its k nearest neighbors in KNN algorithm, to determine k value is particularly important. If a smaller value of K is selected, it is equivalent to smaller training simples in the field of forecast, so approximation error will decrease, only with the input instance is close or similar training simples will only work on forecast results, at the same time the problem is that the estimation error will increase. The decrease of the K value means the whole model is complicated and overfitting is easy to occur; If a larger value of K is selected, it is equivalent to using training examples in a larger field to make predictions. Its advantage is that it can reduce the estimation error of learning, but its disadvantage is that the approximation error of learning will increase. At this time, training simples far from the input instance will also act on the predictor, making the prediction error, and the increase of K value means that the overall model becomes simple (Li). Here the researcher use a looping statement to test the prediction accuracy of different K values on the test base one by one. The KNN algorithm is achieved by using Kneighborsclassifier function in sklearn library. All the parameters in Kneighborsclassifier function except K value are the default value. K values range from 1 to 20.

It can be seen that when k value is small, the accuracy of the model is low. However, when k value is greater than 3, the accuracy of the model is significantly improved, exceeding 78 percent. When k value continues to increase, the accuracy of the model can be maintained at a good level, about 80 percent, although it does not continue to increase.

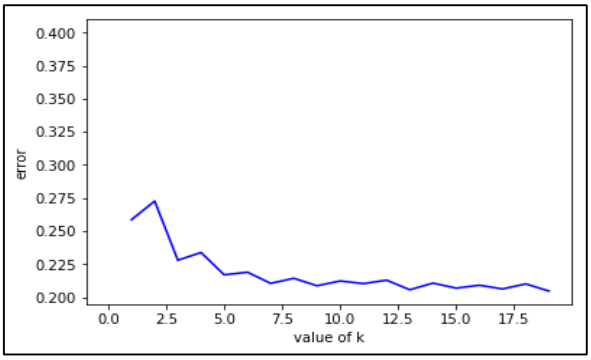


Fig. 18: K-value is the most important and influential parameter. Furthermore, the Kneighborsclassifier function have many different parameters, including weights and p. They can determine how a simple count in the algorithm and how to calculate the distance. The researcher still needs to adjust these parameters out of necessity. Weights includes 'uniform' where all point in each neighborhood are weighted equally and 'distance' where points are weighted by the inverse of distance and closer point will have a greater influence than others. The researcher can find out that in many point weights 'uniform' and weights 'distance' have same effect of prediction, but weights 'distance' is overall better than 'uniform'. So, weights 'distance' have a better prediction accuracy.

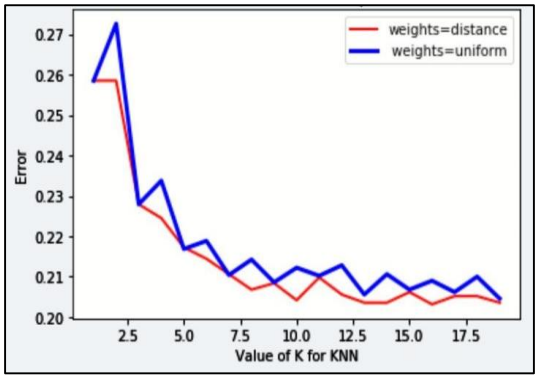


Fig. 19: P determine the calculation of the distance. When p is equal to one, the researcher uses Manhattan distance. When p is equal to two, the researcher uses Euclidean distance. Figure 2.4. shows that the prediction accuracy generally decreases with the p value rising, so the researcher chooses Manhattan distance in the final model.

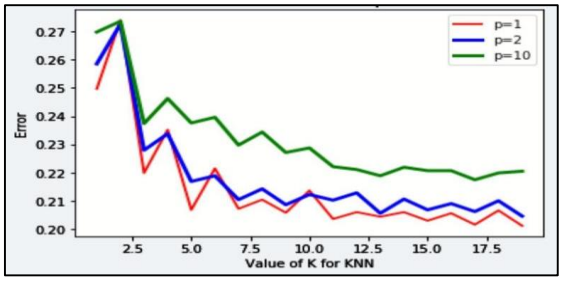


Fig. 20: Furthermore，to avoid the parameters affecting each other，the researcher used method of exhaustion to test every combination of above three parameters. The algorithm was achieved by GridSearch function in sklearn. Finally，the researcher got the best choice of parameter by testing the accuracy. K value:19. Weights: distance. P:1. The accuracy of the final model can reach 80 percent.

The researcher used these parameters in the final model.

Table 1: Results and Future work

	Decision Tree	Logistic Regression	LSTM	AdaBoost	bagging method	KNN
Accuarcy (out of 1.00)	0.78	0.85	0.85	0.82	0.798	0.80

6. Conclusion

After testing all models, the researcher can find that the models the researcher have trained can't predict perfectly. Although the researcher has tried ensemble method like AdaBoot, the effect of ensemble method is only slightly better than some weak learners. The current model can achieve an accuracy of 0.82, which is not outstanding enough as our expectations.

The researcher has several hypotheses about the reason why our accuracy cannot be improved furtherly. After searching for some papers and research reports the researcher find that combining several models for specific issue is an effective way to classify. So one possible reason why our accuracy cannot be improved is that our models are too weak and original. The researcher needs to enhance our model to find a better one. And another reason may be that predicting rain itself is an inaccurate behavior, so it is impossible to determine

whether it will rain tomorrow just by some simple attributes. Or maybe the attributes the researcher choose is just cannot represent whether it will rain tomorrow.

AACKNOWLEDGMENT

Hanqi Shi and Yifei Ma are both the second authors.

REFERENCES

[1] Young, Joe. "Rain in Australia." *Kaggle*, 2018, www.kaggle.com/jsphyg/weather-dataset-rattle-package/version/2. Accessed 20 Sept. 2021.

[2] Rafferty, John P., editor. "Numerical Weather Prediction (NWP) Models." *Encyclopaedia Britannica Online*, Encyclopaedia Britannica, www.britannica.com/science/weather-forecasting/Numerical-weather-prediction-NWP-models. Accessed 19 Sept. 2021

[3] Hickey, Jason. "Using Machine Learning to 'Nowcast' Precipitation in High Resolution." *Google AI Blog*, Google, 13 Jan. 2020, ai.googleblog.com/2020/01/using-machine-learning-to-nowcast.html. Accessed 18 Sept. 2021.

[4] Zhou, Zhi-Hua. *Ensemble Methods Foundations and Algorithms*. E-book ed., Taylor and Francis Group, 2012.

[5] Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Royal Statistical Society*, vol. 58, no. 1, 1996, pp. 267-88. *JSTOR*, www.jstor.org/stable/2346178?seq=1#metadata_info_tab_contents. Accessed 22 Sept. 2021.

[6] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long Short-Term Memory." *Neural Computation*, vol. 9, no. 8, 15 Nov. 1997, pp. 1735-80. *MIT Press Direct*, https://doi.org/10.1162/neco.1997.9.8.1735. Accessed 22 Sept. 2021.

[7] Lemmens, Aurélie, and Christophe Croux. "Bagging and Boosting Classification Trees to Predict Churn." *Journal of Marketing Research*, vol. 43, no. 2, 1 May 2006, pp. 276-86. *SAGE Journals Online*, https://doi.org/10.1509/jmkr.43.2.276. Accessed 22 Sept. 2021.

[8] Soui, Makram, et al. "NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient's symptoms." *National Library of Medicine*. *Pub Med*, https://doi.org/10.1007/s11071-021-06504-1. Accessed 22 Sept. 2021.

[9] Liu, Wenfei, et al. "Comparisions on KNN, SVM, BP and the CNN for Handwritten Digit Recognition." *IEEE Xplore*, Aug. 2020, pp. 25-27. *IEEE Xplore*, https://doi.org/10.1109/AEECA49918.2020.9213482. Accessed 22 Sept. 2021.

[10] Mirceva, G., et al. "Classifying Protein Structures by Using Protein Ray Based Descriptor, KNN and FuzzyKNN Classification Methods." *IEEE*, Nov. 2020. *IEEE XPLORE*, https://doi.org/10.23919/MIPRO48935.2020.9245442. Accessed 22 Sept. 2021.