Machine Learning Models for Weather Prediction

Neethika Hariharasakthy
Department of Electrical and Infromation Engineering.
University of Ruhuna.
Galle, SriLanka.
neethikahariharasakthy2000@gmail.com

Praveenan Jeevarethinam
Department of Electrical and Infromation Engineering.
University of Ruhuna.
Galle, SriLanka.
praveenanjvp@gmail.com

***Abstract*—This project investigates one of the crucial application of machine learning algorithms, specifically Logistic Regression Algorithm and Decision Tree Algorithm in the domain of Weather Prediction. The principal goal of this project is predicting weather conditions based on the input data which means that based on the other types of weather data. The choice of the Logistic Regression & Decision Tree algorithm is made based on its attitude for multiple category classification tasks and interpretability. By thinking the importance of weather data prediction in the field of Agriculture, Transportation, Renewable Energy , Disaster Preparedness, Tourism etc ,this type of weather dataset is selected for our machine learning model analysis project. In the existing manual process for weather prediction has many challenges because of the vast data volume which may affect the data quality & forecasting accuracy. So that our project is considered as successful and have the ability to predict the weather based on the given unseen data.**

Keywords—Machine Learning(ML), Decision tree(DT),Logistic Regression(LR)

I. INTRODUCTION

Weather data is critical to our everyday life and different businesses, including agriculture, transportation, and disaster preparedness. Understanding and successfully utilizing weather data may assist individuals and companies in making informed decisions, increasing safety, and optimizing operations [1]. So that this project employs machine learning methods, notably Logistic Regression and Decision Tree, to improve weather forecast accuracy. By this we can develop a model that can forecast weather conditions based on a variety of input variables to overcome the struggles in existing system process.

Accurate weather prediction is extremely important since it has a direct impact on crucial industries. In the sector of agriculture the farmers plan the schedule of irrigation ,planting, harvesting based on the weather predictions. At the same time weather data helps airlines, shipping businesses, and road networks optimize routes, avoid dangerous situations, and assure passenger safety. Wind and solar energy generating also rely on meteorological data to forecast energy output and efficiently manage power networks. In the daily life weather predictions help tour operators and travellers plan outdoor activities and get the most out of their travels. Eventually as a Sri Lankan ,weather data is critical for early warning systems and emergency response plans during hurricanes, floods, and other severe weather occurrences [1].

So the main aim is by using the capabilities of Logistic Regression and Decision Tree algorithms, we want to overcome the issues associated with human weather forecast procedures, providing a more efficient and trustworthy solution. So that when we analysing the advantages & disadvantages of the both algorithms we have got an idea about the adaptability level of these algorithms related with our dataset .Logistic Regression is simple, understandable, and effective binary or multi-class categorization & suitable for probabilistic predictions. But the issue is it assumes a linear relationship between features and may not capture complex interactions in the data. On the other hand decision tree can handle the non-linear relationships & high interpretation data. So that by using this algorithm we can be able to capture complex decision boundaries. In contrast decision tree algorithm is sensitive to overfitting in the situation of noise presence & redundant prediction. So eventually ,because of our large number of category in labels the logistic regression could not be able to predict with high accuracy .

II. METHODOLOGY

A. Data

Our dataset is related with the field of weather prediction which plays crucial role from daily lives up to various huge industries. [1] Here in our case this weather dataset is little bit complex because of the multi category classification instead of binary classification. In this way our dataset contains 8784 data and 7 features .Here our features are 'Date/Time', 'Temp_C', 'Dew Point Temp_C', 'Rel Hum_%', 'Wind Speed_km/h', 'Visibility_km', 'Press_kPa' and label is 'Weather'. But here ‘Date/Time’ is object data type which means text , 'Temp_C', 'Dew Point Temp_C' 'Visibility_km' and 'Press_kPa' are float data type and 'Rel Hum_%' and 'Wind Speed_km/h' are integer data type and the label ‘Weather’ is also object data type which means text type. In the label there are 50 unique categories which are listed down below with the count of each category in the dataset :

Mainly Clear ,Mostly Cloudy ,Cloudy , Clear, Snow, Rain, Rain Showers, Fog, Rain,Fog, Drizzle,Fog, Snow Showers, Drizzle, Snow,Fog Snow,Blowing Snow, Rain,Snow, Thunderstorms,Rain Showers, Haze, Drizzle,Snow,Fog, Freezing Rain, Freezing Drizzle,Snow, Freezing Drizzle, Snow,Ice Pellets, Freezing Drizzle,Fog, Snow,Haze, Freezing Fog, Snow Showers,Fog, Moderate Snow, Rain,Snow,Ice Pellets, Freezing Rain,Fog, Freezing Drizzle,Haze, Rain,Haze, Thunderstorms,Rain, & Thunderstorms,Rain Showers,Fog

This dataset is taken from the Kaggle website which contains variety of datasets from the various fields in real world [1].

B. Pre-processing

In the pre-processing ,we have checked the null or missing & duplicate values in our data set. But there are no null values or duplicate values in our data. Then in the feature selection , we have dropped the Date/Time column because it is not necessary to our data analysing process. Then we have found the correlation among the features by using correlation matrix which is shown in Fig. 1.

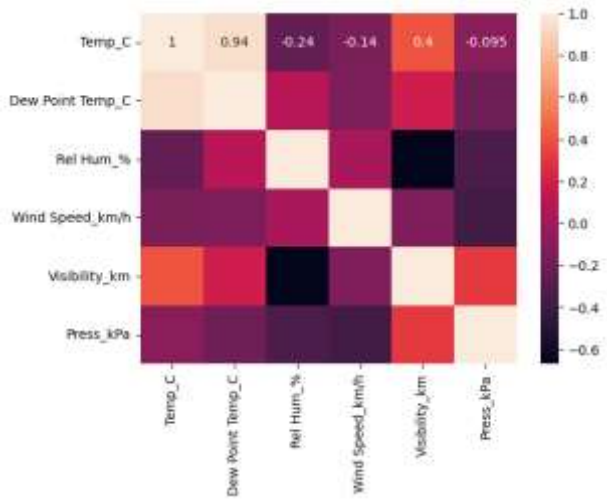

Fig. 1. Heat Map of Features

Then to handle the outliers in the features ,we have used the capping method which involves setting a predetermined threshold beyond which extreme values in features are replaced with the threshold value to mitigate the impact of outliers [2]. In addition to this we have scaled features to convert the numerical values of several features in a dataset to a predefined range, ensuring that no single feature has disproportionate effect on a model due to its larger size. So the above steps we have followed commonly in both algorithms. But especially for each algorithm we have followed some specific pre-processing techniques which will be described below.

1) Decision tree algorithm

*a) Converting the Weather Categories into Standard Categories:*Here in order to increase the accuracy level of DT model ,we have tried some different ways.

- Removing low frequency categories from the weather column & making them as 12 categories
- Removing low frequency categories & rearranged in to 6 categories
- Without removing any categories ,using all 50 categories

b) Encoding the labels: Label encoding method is used for encoding the final all 50 unique categories.

2) Logistic Regression algorithm

*a) Converting the Weather Categories into Standard Categories:*Here inorder to increase the accuracy of LR model, we have tried some different ways.

- Removing low frequency categories from the weather column & making them as 12 categories
- Removing low frequency categories & rearranged in to 6 categories
- Without removing any categories ,recategorized in to 7 categories

b) Sample Selection : 600 samples are selected from each 7 category .So that finally we got 4 categories.

c) Encoding the labels: Label encoding method is used for encoding the final 4 unique categories.

Eventually we have splitted our data set into training & testing with test size is 30% of the data to training & evaluating the model.

C. Algorithms

In our weather prediction project, we have used Decision tree algorithm and Logistic regression algorithm which allowed us to classify the multi-categorical nature of our weather dataset and make accurate predictions for different weather states.

1) Decision tree algorithm(DT)

DT algorithm is used for both classification and regression tasks, it is used in the context of weather prediction to provide predictions based on various weather-related features. It is constructed by recursively splitting the dataset based on the feature that provides the most information about the target variable (e.g., rain, snow, sunny, cloudy).

Decision trees can handle missing values itself during the training process and generally robust to outliers. Decision tree is capable of handling both numerical and categorical data. For categorize our 50 unique weather outcomes, the algorithm uses technique label-encoding to represent them as numerical values for decision making.

$$MSE = \Sigma(yi - \bar{y})^2/n \quad (1)$$

$$VR = Var_{parent} - \Sigma(NiVari/N) \quad (2)$$

Mean Squared Error (1) and Variance Reduction (2) equations guide the decision tree algorithm in selecting the most suitable features and thresholds to split the data at each node.

2) Logistic Regression algorithm

Logistic regression is a greatest classification algorithm that models the probability of a binary outcome. Here the logistic regression was applied to classify multi-categorical weather data, assigning probabilities to various weather conditions. A key component of logistic regression was the logistic function (sigmoid function), it made the classification of weather categories easier by converting the linear combination of input information into probabilities.

The logistic regression model was trained using gradient descent to optimize its parameters and other iterative techniques to change weights and biases. The goal of this training phase was to minimize the difference between predicted and actual outcomes. Feature scaling and

normalization were applied to ensure balanced contributions from Various weather features, such as temperature, humidity, and wind speed etc.

Logistic Function (Sigmoid Function)

The logistic regression model uses the sigmoid function (3) to transform the linear combination of input features.

$$\sigma(z) = 1/(1+e)^{-z} \quad (3)$$

$$z = b + w_1 x_1 + w_2 x_2 + w_3 x_3 + \text{-------} + w_n x_n \quad (4)$$

In (4), z is the linear combination of input features and their associated weights.

Logistic Regression Cost Function

The cost function (5) is used to measure the difference between the predicted value and the actual value.

$$z = -y \log[h\theta(x)] - (1-y)\log[1- h\theta(x)] \quad (5)$$

Gradient Decent For Logistic Regression

Gradient Descent (6) is used in logistic regression to find the minimum of the cost function.

$$\theta j := \theta j - \alpha \, \partial J(\theta)/ \partial \theta j \quad (6)$$

In order to make use of the advantages of each method for accurate and reliable classification diverse weather conditions., we used both the Decision Tree and Logistic Regression algorithms in our weather prediction project.

D. Implementation

The algorithm code utilized in this implementation was obtained from Kaggle website for decision tree [3] and from a you tube video for logistic regression [4].

Modifications to the decision tree Algorithm code:

- We have added the k-fold cross validation method for increase the performance of the decision tree algorithm.
- K-fold cross-validation was employed with the following settings in our implementation:

  Number of Folds (k): 5
- And also, we have used the confusion matrix that is used to evaluate the performance of our model.
- By using an overfitting graph we have visualized the model's performance on both training and validation datasets.

Modifications to the Logistic Regression Algorithm code:

- We have used the capping method to handling the outliers in all the features in our weather data set.
- Here also we have used the k-fold cross validation method for increase the performance of the Logistic Regression algorithm.

  Number of Folds (k): 10
- We have used an overfitting graph to visualized the model's performance on both training and validation datasets.
- Finally, we have added the Classification report to analyze the model performance using Accuracy (7), Precision (8), Recall (9) and F Score (10).

NOCP : Number of correct predictions

TNOP : Total number of predictions

TP : True Positive

FP : False Positive

FN : False Negative

$$Accuracy(A) = NOCP/TNOP \quad (7)$$

$$Precision(P) = TP/(TP+FP) \quad (8)$$

$$Recall(R) = TP/(TP+TN) \quad (9)$$

$$F\ Score = (2\times P\times R)/(P+R) \quad (10)$$

Test data set values and Train data set values are the parameters that we have used for running our both Decision Tree algorithm and Logistic Regression algorithm. Here 70% of fixed data used for training and 30% fixed data for testing.

III. RESULTS

After the implementation ,we have got some results as accuracies of each models in each & every steps that we have followed.

A. Decision tree algorithm

- Removing low frequency categories leads low accuracy as 0.72
- Removing low frequency categories & rearranged in to 6 categories leads to reducing the accuracy as 0.45.
- Without removing any categories ,using all 50 categories leads to achieve the final high accuracy as 0.89.

There is no overfitting in this model so that the training accuracy & testing accuracy are achieved as below.

- Train accuracy: 0.8964319630986297
- Test accuracy : 0.8913896802785691

B. Logistic Regression Algorithm

- Removing low frequency categories from the weather column & making them as 12 categories leads to reducing the accuracy as 0.39.
- Removing low frequency categories & rearranged in to 6 categories leads to reducing the accuracy as 0.45.
- Without removing any categories ,recategorized in to 7 categories & then finally in data balancing comes to 4 categories leads to achieve the final high accuracy as 0.63.

There is no overfitting in this model so that the training accuracy & testing accuracy are achieved as below.

- Train accuracy: 0.642640823743186
- Test accuracy : 0.6313559322033898

IV. Discussion & Conclusion

A. Decision Tree Algorithm

1) *Impact of Category Removal:*
   The experiment indicated that deleting low-frequency categories significantly affected the accuracy of the Decision Tree model. The accuracy reduced to 0.72, indicating that the low-frequency categories include useful information for the model.
2) *Effect of Category Rearrangement:*
   When low-frequency categories were not only deleted, but also restructured into six categories, accuracy dropped even further (0.45). This suggests that the initial distribution and order of categories aids the model's understanding of trends in the weather.
3) *Performance with All Categories:*
   Unexpectedly, when all 50 categories were preserved without elimination, the model had a high accuracy of 0.89. This implies that the Decision Tree algorithm can handle a wide range of weather categories, and that removing any one category has a negative impact on its performance.
4) *No Overfitting:*
   The absence of overfitting, as indicated by equal training and testing accuracies (0.896 and 0.891, respectively), supports the Decision Tree model's generalizability.

B. Logistic Regression Algorithm

1) *Effect of Category Removal:*
   Similarly to the Decision Tree, deleting low-frequency categories from the weather column reduced accuracy to 0.39. This shows that these unusual categories add useful information to the Logistic Regression model.
2) *Impact of Category Rearrangement:*
   When low-frequency categories were restructured into six categories, accuracy was reduced by 0.45. This underscores the relevance of the initial category distribution on the model's success.
3) *Data Balancing and Recategorization:*
   Notably, reaching a high accuracy of 0.63 necessitated a deliberate strategy that included recategorizing into seven categories and then balancing the data to four categories. This demonstrates Logistic Regression's sensitivity to category makeup and balance.
4) *No Overfitting:*
   Logistic Regression, like the Decision Tree, showed no overfitting, with training and testing accuracies (0.642 and 0.631) that were almost identical.

C. Ethical Aspects

1) *Bias in Category Removal:*
   The decision to exclude low-frequency categories may induce bias, particularly if these categories are related to unique weather conditions that affect distinct groups or areas. Ethical concerns should address the possibility of uneven representation and unexpected effects.
2) *Data Balance and Fairness:*
   The technique of data balancing for Logistic Regression poses ethical concerns regarding fairness. Ensuring that categories are balanced without mistakenly favoring particular groups is critical for avoiding algorithmic bias and making fair forecasts.

In conclusion, the use of Decision Tree and Logistic Regression algorithms demonstrated the importance of category distribution and composition in weather prediction. Both algorithms were sensitive to category removal and rearrangement, although the Decision Tree model performed better when all categories were retained.

References

[1] H. Realm, "How to Detect and Remove Outliers in the Data | Python," 2023. [Online]. Available: https://youtu.be/Cw2IvmWRcXs?si=1vVyKcW3CB3H_aIu. [Accessed 1 January 2024].

[2] B. BISWAS, "Weather Data," June 2023. [Online]. Available: https://www.kaggle.com/datasets/bhanupratapbiswas/weather-data. [Accessed 1 January 2024].

[3] S. MENON, "Weather Data Classification(.96 Accuracy)," June 2023. [Online]. Available: https://www.kaggle.com/code/swathiunnikrishnan/weather-data-classification-96-accuracy.

[4] T. D. f. Lab, "Machine Learning Weather Forecasting [End to End Project]," [Online]. Available: https://youtu.be/70x6TDNP5HY?si=_CfbQUAcAJsSIonB. [Accessed 1 January 2024].