# PREDICTION ANALYSIS FOR WEATHER DATA

18 33 002 - AKKSHAYA SRI. J

18 33 026 - MOHAMMED ISMAIL. A

18 33 046 - SANCHEZ INNOCENCIA D

18 33 052 -SUBASH. S

## I. ABSTRACT

This paper deals with the weather prediction, which is an attempt to predict the weather conditions at some future time. The parameters here in this dataset covers almost all the measurable factors that affect the weather and the temperature. Here, the maximum and minimum temperature are the main parameters of the dataset for the weather analysis. The data is collected from the Global Historical Climatology Network(GCHN) which helps in predicting the temperature values of a particular region. The data are pre-processed and we have built three machine learning algorithms such as Support Vector Regression, Decision Tree Regression and Random Forest Regression to see which model fits the best to our weather dataset. After building the models as a web application using streamlit , we have deployed them in the Heroku server.

**Keywords:** Multi target Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression,Streamlit, Heroku.

## II. INTRODUCTION

Weather forecasting has been one amongst the issues in the scientific and technological world since the last few centuries. Weather forecasting provides information about the future weather. Weather is one of the non-linear and dynamic parameters which varies day-to-day or even minute-to-minute. Weather forecasting remains the big challenge of its data intensive and the frenzied nature. G The data for this analysis is collected from the Global Historical Climatology Network(GCHN) , we have taken the dataset of the past 10 years i.e., from the

year 2011 - 2020. The weather condition is based on the minimum temperature, maximum temperature, dew point, station pressure, visibility, windspeed, sea level pressure, precipitation, GUST value in this analysis. Gust is a value which means a brief burst of wind. Apart from the other variables, we have also considered a FRSHTT variable which means the fog, rain, storm, hail, thunder and tornado in that particular place. The data is taken on a day-wise basis. After collecting the data, we have pre-processed the data. In this preprocessing process, we have dropped a few parameters which are not really required for the analysis. Using the preprocessed data, we developed a correlation heat map to see the relationship between the variables and most importantly, to identify the highly intercorrelated feature variables. After interpreting the heatmap, we removed the highly negatively correlated and positively correlated variables. After processing our data, we used the Support vector regression, decision tree regression and random forest regression to predict the minimum and maximum temperature of our dataset. We calculate the RMSE score and R-Squared value to choose which model fits the best for our dataset. This helps to choose the precise model for our dataset. Finally, we developed it as a web application using streamlit and deployed this in the heroku cloud server.

## III. LITERATURE REVIEW

Usman Malik in his work of Random Forest Algorithm with Python and Scikit-Learn explained about the working of Random Forest Algorithm and also explained about the advantages and disadvantages and further solved a problem using Python and Scikit-Learn.

Tom Sharp in his work of An Introduction to Support Vector Regression (SVR) explained about the usefulness of SVR compared to the other regression models using Boston Housing Price Example dataset.

Kiran Karkera in his work of Regression Models with Multiple Target Variables gave a good overview of the model approaches to multi-target regression, the problem transformations, and the comparison of ensemble methods.

Nagesh Singh Chauhan is his work of Decision Tree Algorithm, Explained gave an idea about the decision trees and how to build and optimize the decision tree classifier and the types and important Terminologies.

Colton Stapper in his work of An Analysis of Heroku and AWS for Growing Startups explained the information on the architecture of modern cloud-hosting platforms and gave an analysis of Platform-as-a-Service company: Heroku.

## IV. DATASET DESCRIPTION

Once, when we collected data from the Global Historical Climatology Network (GCHN) website, there were actually around 23 columns with many attributes such as station number, temperature, temperature count, dew point, dew point count, sea level pressure, sea level pressure count, station pressure, station pressure count, visibility, visibility count, wind speed, wind speed count, Maximum sustained wind speed, Maximum temperature, Minimum temperature, Precipitation amount, Snow depth, GUST, frshtt. The gust value here is the value that is counted when there is a brief burst of wind. The frshtt attribute is an attribute which includes fog, rain, storm, hail, thunder and tornado into the consideration. These are the attributes which we collected for the analysis. Since, we won't require the count of the particular attributes, we dropped those columns. Then, we also dropped the station number from the dataset because we are implementing this analysis only for one particular station. Hence, even that attribute is not really required for the analysis. Finally, we have 15 columns with 3637 row entries after dropping a few attributes.

## V. SOLUTION CORRELATION HEATMAP

When the dataset is large with many variables, then a quick way to analyse the relationship between the variables is to plot a correlation heat map for the variables. By this way, we can understand the relationship between the variables and this paves a way to identify the variables which are highly correlated.

From the visualization, we can see that the Gust and year are negatively highly correlated and also the temperature with max temp and min temperature are highly positively correlated . Hence, these variables can be removed from the datasets. Finally, we are removing the temp, mxspd and GUST variables from the dataset.

## VI SOLUTION STRATEGY

For weather data analysis, we have implemented three machine learning algorithms and we have built them altogether along with the data visualization as a web application using streamlit in python. Finally, we deployed this web application in a cloud server and here we used the Heroku cloud server to deploy this web application. The machine learning models we chose to build for our weather dataset are decision tree regression, support vector regression and random forest regression. We also calculated the RMSE score of each model to know which model fits the  best comparatively. The predicted value and actual value of each model are plotted and visualized for better interpretation. In all the three machine learning models, we have predicted the minimum and maximum temperature.

## RMSE

RMSE stands for Root Mean Square Error. It is the standard deviation of the predicted values. These predicted errors are also called residuals. Residuals are the distance measure of data points and the regression line. Root Mean Square Error calculates the dispersion of these residuals.Root mean square error is most commonly used in the climatology field, forecasting, and regression analysis to verify the experimental results.

## TECHNIQUES

## SUPPORT VECTOR REGRESSION

Support vector regression is a regression method which is implemented using the concept of support vector machines. In machine learning, support vector machines

technique is considered to be one of the supervised learning models with associated learning algorithms. It analyses the data for classification and regression analysis. Parallely, it also performs linear classification and this algorithm can also perform a non-linear classification efficiently. SVR implicitly maps their inputs into high-dimensional feature spaces and when data are unlabelled, supervised learning is not possible. Hence, an unsupervised learning approach is required to find natural clustering of the data to groups and then map these new data to these formed groups.

**DECISION TREE REGRESSION**

Decision tree is one of a machine learning algorithm used widely in every industry. This model builds the regression model or a classification model in the form of a tree structure. It breaks down the dataset into smaller and smaller subsets. These subsets together form the decision tree and are further developed. The final result is represented as a tree with decision nodes and leaf nodes. A decision node can have two or more branches and these branches represent the values for the attributes. Decision trees can handle both the data types i.e., numerical and categorical data.

**RANDOM FOREST REGRESSION**

Random forests can also be known as random decision forests. Random forest can manipulate the data and work upon it. Even if there are missing values in the dataset., it does not affect much in the random forest technique. Random forest regression is the best machine learning algorithm, if the dataset is huge. There are some chances of overfitting the data while training but the benefit of using random regression is that their accuracy rate will be most probably high comparatively. Random forest regression generates reasonable predictions across a huge amount of data.

**MULTI TARGET REGRESSION**

Usually, while predicting the values using machine learning algorithms, we predict only one variable and that is termed as the target variable. In the case of regression models, the target is real valued and those real values are our predicted value. To

predict those values, we use some machine learning algorithms. Now, this Multi target regression is a term which ensures that we can predict more than one variable. When there is more than one independent variable, then there can also be one or more dependent variables.

## DEPLOYMENT

### STREAMLIT

Streamlit is one of the python libraries which is used to build web applications using data and machine learning. Streamlit apps can be deployed instantly and it also helps in building powerful apps easily. Streamlit is one of the open source frameworks and this is very much useful to all the users in the data stream. It purely works in python. The best part is that we can build a highly interactive web application using stream lit.

### HEROKU

Heroku is a cloud application platform and it is a container based platform. Heroku enables the developers to build, run, monitor and operate the applications entirely in the cloud server itself. Heroku is also considered to be user - friendly. Heroku can handle both hardware and software, hence, the user can keep focusing on enhancing the applications alone. The advantage of working on heroku server is that there are free services for the first few small projects. Heroku supports several programming languages such as Java, Node.js, Scala, Clojure, Python and PHP.

## RESULTS AND DISCUSSIONS

### Random Forest Regression

Minimum Temperature and maximum temperature predicted values

RMSE score of random forest regression

**Decision Tree Regression**

Rmse Score of Decision Tree

**Support Vector Regression**

## APPLICATION

The link for the App that is deployed in Heroku is given below
https://mystreamlit1.herokuapp.com/

## Conclusion

We have made a comparison study between the three Machine Learning Predictive Algorithms by finding their accuracy. Using RMSE, we have come to a decision that the random forest regression model is a best fit model to our dataset compared to the other two models. We conclude this by comparing their accuracy score and random forest regression model tend to have higher accuracy. We have also plotted the graph of predicted and the exact value of the minimum and maximum temperature.Using these graphs, we can understand the trend and fluctuations in our predicted values.Finally,we have deployed the Machine Learning Model in heroku using Streamlit. Using Streamlit, we developed a Web Application and hosted it on the Heroku cloud server.