Weather Prediction

Team name: BlackSharks

Team members:

1. Sudheer Kumar Reddy Reddymalla
   sredd1@unh.newhaven.edu
   UNH ID: 00719922
2. Durga Guna Sekhar Reddy Savanam
   dsava2@unh.newhaven.edu
   UNH ID:00755099
3. Vamsi Krishna Regalla
   vrega1@unh.newhaven.edu
   UNH ID: 00730353

Abstract:

Most clouds in the tropical environment are produced by convective instabilities, which can arise within an hour. Forecasting the weather in any region can be difficult. Convective precipitation is a type of short-lived, quickly fluctuating precipitation that covers a specific area. Floods and landslides can result from intense convective precipitation, resulting in costs for the economy and society. Here we are going to forecast the weather based on information previously gathered through data mining methods.

Introduction:

Weather prediction is the use of science and technology to make predictions about the state of the atmosphere in a specific area. This covers the temperature, snow, rain, and clouds. A unique type of short-range prediction is used to issue weather alerts to preserve human life. Depending on the area, governments all around the world issue weather warnings for a variety of major weather occurrences, including tropical storms and tropical cyclones. Long-range or short-range forecasts are both possible. It is both an exciting and difficult endeavor. This report gives readers a fundamental grasp of the intent and scope of weather forecasts, as well as the fundamental ideas and common prediction models.

For a variety of applications, weather forecasting is crucial. Some of these include monitoring the climate, spotting droughts, forecasting severe weather, and the agriculture industry and tourism industry. There is a long historical record of times when weather conditions have changed the outcome of military operations. The dynamic nature of the atmosphere makes it difficult to predict weather conditions with any degree of accuracy. Some variables may be used to depict the weather at any time. One discovered that the most important of those variables was chosen to be included in the prediction procedure. Weather forecasting could be done in hours, or days or it may even take more than expected in certain areas with certain weather conditions.

Research Question:

To predict the weather patterns using data mining techniques.

Dataset:

The dataset "Weather Prediction" we have chosen is from Kaggle Website (https://www.kaggle.com/datasets/ananthr1/weather-prediction). This dataset contains 1461 tuples and 6 attributes. Here, we have been taken Seattle city weather dataset.

Data Defining:
Based on some factors we are, going to predict the weather.

Input Variables:
date: starting from 01-01-2012 to 31-12-2015.
precipitation: All forms in which water falls on the land surface and open water bodies as rain, sleet, snow, hail, or drizzle.
temp_max: Maximum temperature of a particular day.
temp_min: Minimum temperature of a particular day.
wind: Wind speed (Km/hr.)
Target Variable:
weather: 'drizzle', 'rain', 'sun', 'snow', 'fog'

Related Work:

1. Title: Evaluation of High-Resolution Tropical Weather Forecasts Using Satellite Passive Millimeter-Wave Observations
   Authors: Chinnawat Surussavadee
   Affiliation: Massachusetts Institute of Technology, Cambridge, MA, USA.
   Publication Date: 28 June 2013
   Publisher: IEEE
2. Title: Machine Learning Based Weather Prediction Model for Short Term Weather Prediction in Sri Lanka
   Authors: K.M.S.A. Hennayake; Randima Dinalankara; Dulini Yasara Mudunkotuwa
   Affiliation: Department of Computer Engineering, Faculty of Engineering, University of Sri Jayewardenepura, Sri Lanka
   Publication Date: 10 November 2021
   Publisher: IEEE
3. Title: Forecasting Weekly Rainfall Using Data Mining Technologies
   Authors: T. Dananjali, S. Wijesinghe, J. Ekanayake
   Affiliation: Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
   Publication Date: 11 June 2021
   Publisher: IEEE
4. Title: Rainfall prediction based on 100 years of meteorological data
   Authors: Sandeep Kumar Mohapatra; Anamika Upadhyay; Channabasava Gola
   Affiliation: Aricent Technologies Private Limited, Bangalore, India
   Publication Date: 08 February 2018
   Publisher: IEEE

5. Title: Data Mining as a tool for hot day prediction during summer monsoon
   Authors: B.Prashanthi, S.Meganathan, R.Bala Krishnan, R.Varahasamy, S.Swaminathan

Affilation: Sastra University, Thirumalaisamudram, Thanjavur, Tamil Nadu, India
Publication date: 04 August 2016
Publisher: IEEE

Proposed Method:

The proposed methodology consists of several steps including acquiring weather observation data, data preprocessing, learning model, make predictions on test set and finally evaluate the predictions made by the model.

Preprocessing:

Data pre-processing is the technique of transforming a raw, real-world dataset into a state that can be easily interpreted by the data mining techniques. The collected weather dataset is preprocessed by following steps such as data cleaning, feature selection, inspecting the dataset and visualization, data normalization, and dividing the data set into training and test sets. In the data cleaning step, the missing dates are imputed in the dataset and then missing values and null values are imputed using the next observation carried backward method. This method is used because of the persistence nature of the weather variables. In feature selection, a correlation matrix is created to analyze the correlation between weather variables in the dataset. In the data inspect and visualization step, the weather dataset was visualized using various plots.

Naïve Bayes:

A probabilistic classifier is the Naive Bayes algorithm for classification. It is based on probability models that make substantial assumptions about independence. The independence presumptions frequently do not affect reality. They are therefore viewed as being naive.
By using Naive bayes model, correctly classified instances are 1269 which is 86.8%.

K-Star algorithm:

K-star is an instance-based classifier, meaning that it bases the classification of a test instance on the classification of training instances that are like it as specified by some similarity function. Use of a modelling and simulation distance function sets it apart from other instance-based learners.
By using K-star algorithm, correctly classified instances are 100%.

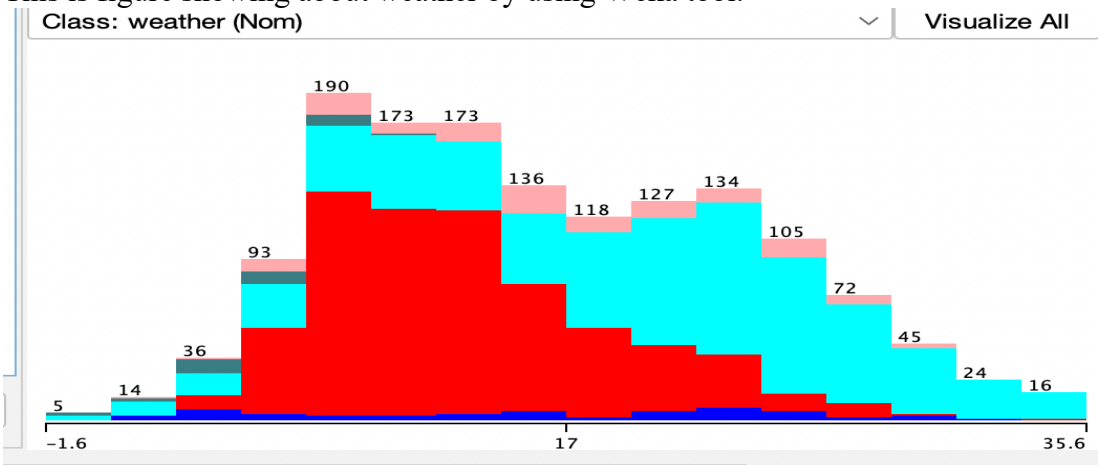
SMO model:

It is a method for resolving the quadratic programming (QP) problem that appears during training support vector machines (SVM). Support vector machines are frequently trained using SMO, which is implemented by the well-liked LIBSVM tool. As a result, that earlier SVM training techniques were far more complicated and needed pricey third-party QP solvers.
In this model, correctly classified instances are 99%.

Random Forest:

Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. It creates decision trees from several samples, using the majority vote for classification and the average in the case of regression.In this random forest model, correctly classified instances are 100%.

Visualization:
This is figure showing about weather by using Weka tool.



Parameters and hyperparameters:
Random Forest:

- Batch size: 300
- maxDepth: 0
- minDepth: 1.0
- minVarianceProp: 0.001
- numFolds: 0
- seed: 1

Decision Table:

- Batch size: 300
- evaluationMeasure: accuracy
- numDecimalPlaces: 2
- search: BestFirst
- debug: False

Naïve bayes:

- Parameter batch size- 300
- useKernekEstimator: False
- numDecimalPlaces: 2

SMO model:

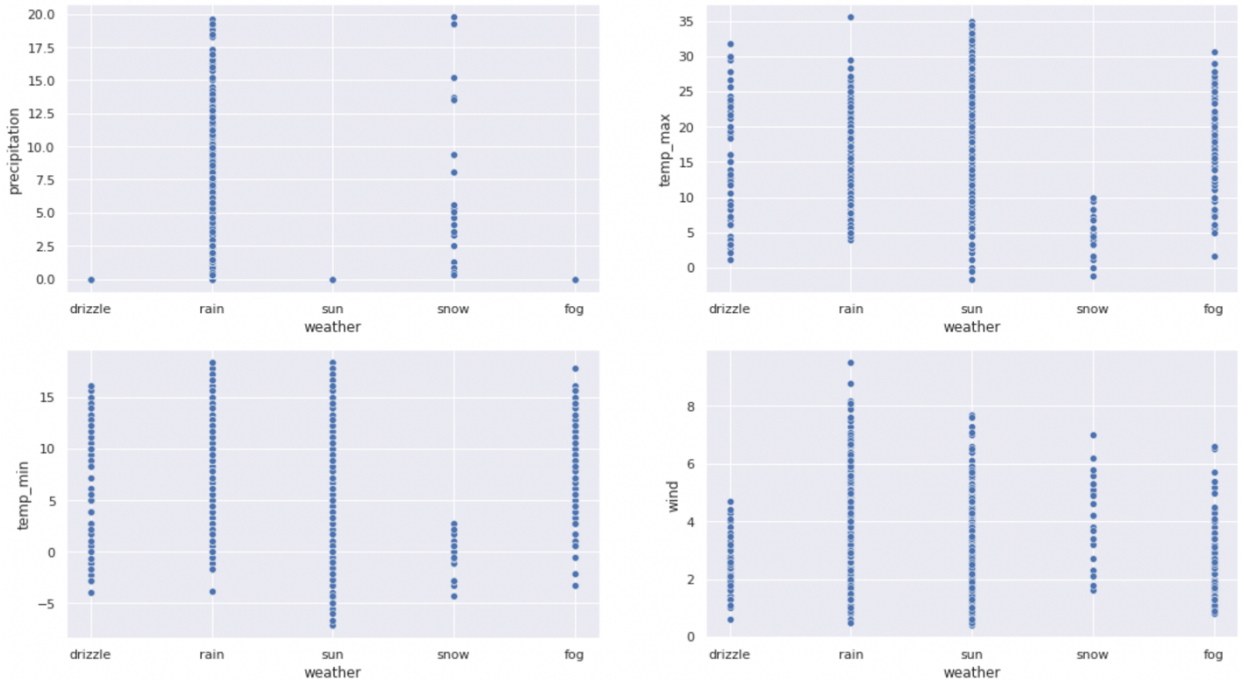
- Parameter batch size- 300
- useKernekEstimator: PolyKernel
- numDecimalPlaces: 2
- tolerance: 0.001
- filter type: Normalize training data

Experimental Results:

The result demonstrates the weather prediction in Seattle city with the help of these classifications. Out of all these classifications, K-star has 100% correct instances with zero missing data and high accuracy.

	Correct instances	True positive	False Positive	Precision	Recall
Naïve Bayes Classification	86.8 %	0.869	0.093	0.885	0.869
SMO model	99%	0.995	0.004	0.995	0.995
K-Star algorithm	100%	1.0	0	1.0	1.0
Random Forest	100%	0.9	0.1	1.0	1.0
Decision Table	85.2%	0.852	0.114	0.88	0.852

These the following graphs that predicting the range of values for wind, precipitation, maximum and minimum temperatures by using K-star and Random Forest.



Discussion:

In this study, the maximum temperature, minimum temperature, and wind speed are classified according to the month and year using a forecasting model's algorithm. With the use of data mining techniques to solve forecasting issues, particularly those involving the forecasting of drizzle, rain, snow, sun, and fog. The technology also incorporates the data mining prediction algorithm.

Conclusion and future work:

In this paper, we experimented with weather prediction using different weather attributes as inputs by implementing a K-star and decision tree model for Seattle city gave us good results when compared with other data mining techniques. Weather is a transient phenomenon, every second the parameters affecting weather change, we have tried to predict the weather a day in advance using the past years of data in a particular area. For training the model we used a fixed sampling of the input dataset to obtain accurate results.

In future, we are going to predict weather accurately over a long period of time like more than a week in advance. Will try to predict using larger dataset for better results. To develop an ANN model to predict in different locations.

GitHub link:

https://github.com/sudheerredde/BLACKSHARKS.git

References:

[1]. Evaluation of High-Resolution Tropical Weather Forecasts Using Satellite Passive Millimeter-Wave Observations, 14131292, IEEE

[2]. Machine Learning Based Weather Prediction Model for Short Term Weather Prediction in Sri Lanka, 21360419, IEEE

[3]. Forecasting Weekly Rainfall Using Data Mining Technologies, 20678724, IEEE

[4]. Rainfall prediction based on 100 years of meteorological data, 17545189, IEEE

Proof Reading with writing center: