GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

WEATHER FORECASTING USING DATA MNING

A Report for the Evaluation 3 of Project 2

Submitted by

ANKIT JAISWAL

(1613112006)

In partial fulfilment for the award of the degree

Of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION OF DATA ANALYTICS

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr. L.VETRIVENDAN SIR, M.Tech, Professor

April/May-2020

GALGOTIAS
UNIVERSITY

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report "**WEATHER FORECASTING USING DATA MINING**" is the bonafide work of "**ANKIT JAISWAL(1613112006)**" who carried out the project work under my supervision.

SIGNATURE OF THE HEAD	**SIGNATURE OF SUPERVISOR**
	MR. L. VETRIVENDAN SIR
SCHOOL OF COMPUTING	**PROFESSOR**
SCIENCE AND ENGINEERING	**SCHOOL OF COMPUTING**
	SCIENCE AND ENGINEERING

TABLE OF CONTENT

ABSTRACT

Changing Climatic conditions are leading to alternate weather patterns. Accurately predicting weather patterns is important as they have wide social and economic impact. This paper proposes and analyzes a predictive model which analyse a wide range of data points with the aim of predicting likelihood and pattern of location-specific rainfall at a high degree of confidence.

We extract knowledge from historical weather data collected from NOAA (National Oceanic Atmospheric Administration). From the collected weather data comprising of 15 attributes, only 5 attributes are most relevant to rainfall prediction. Data preprocessing and data transformation on raw weather data set is performed, so that it shall be possible to work on Decision tree regressor and the data mining, prediction model used for rainfall prediction. The model is trained using training dataset and tested on test data for accuracy. We have used comparative approach of Bayesian and K-NN models and found Bayesian approach to be more accurate.

Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. In this paper, we investigate the use of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed. This was carried out using Artificial Neural Network and Decision Tree algorithms and meteorological data collected. A data model for the meteorological data was developed and this was used to train the classifier algorithms. The performances of these algorithms were compared using standard performance metrics, and the algorithm which gave the best results used to generate classification rules for the mean weather variables. A predictive Neural Network model was also developed for the weather prediction program and the results compared with actual weather data for the predicted periods.

INTRODUCTION

Weather forecasting has been one of the most scientifically and technologically challenging problems around the world in the last century. This is due mainly to two factors: first, it's used for many human activities and secondly, due to the opportunism created by the various technological advances that are directly related to this concrete research field, like the evolution of computation and the improvement in measurement systems [3]. To make an accurate prediction is one of the major challenges facing meteorologist all over the world. Since ancient times, weather prediction has been one of the most interesting and fascinating domain. Scientists have tried to forecast meteorological characteristics using a number of methods, some of these methods being more accurate than others.

Weather prediction is a challenging task and that too for weather is even more complex, dynamic and mindboggling. Weather forecast postures right from the antiquated times as a major gigantic undertaking, because it depends on various parameters to predict the dependent variables like temperature, visibility, wind speed which are changing from weather calculation varies with the some specific location along with its atmospheric

attributes. Accurate forecasts can help to identify possible floods in future and to plan for better water management. Weather forecasts can be categorized as: Forecasts.Which is forecasts up to few hours, Short term forecasts which is mainly Rainfall forecasts is 1 to 3 days forecasts, Forecasts for 4 to 10 days are Medium range forecasts and Long term forecasts are for more than 10 days. Short range and Medium Range rainfall forecasts are important for flood forecasting and water resource management. There are many data mining techniques used for weather predictions. Naive Bayes approach and K-Nearest Neighbor has been used in this paper to forecast the Rainfall. This paper utilize 4 years (2011-2014) data[10] from the month May to October as training dataset. Dataset contains 15 attributes out of which 5 relevant attributes i.e. (Temp, Visibility, Dewpoint, Speed, Rainfall) according to factor analysis and linear regression techniques are considered for rainfall prediction. The test dataset results of Naïve Bayes approach and K-NN(KNearest Neighbor) approach are compared for better results. This paper is organized in three sections, Introduction, Literature survey of the weather prediction models, proposed data mining model with the results of the implementations, and conclusion.

Weather forecasting entails predicting how the present state of the atmosphere will change. Present weather conditions are obtained by ground observations, observations from ships and aircraft, radiosondes, Doppler radar, and satellites. This information is sent to meteorological centers where the data are collected, analyzed, and made into a variety of charts, maps, and graphs. Modern high-speed computers transfer the many thousands of observations onto surface and upper-air maps. Computers draw the lines on the maps with help from meteorologists, who correct for any errors. A final map is called an analysis. Computers not only draw the maps but predict how the maps will look sometime in the future. The forecasting of weather by computer is known as numerical weather prediction.

To predict the weather by numerical means, meteorologists have developed atmospheric models that approximate the atmosphere by using mathematical equations to describe how atmospheric temperature, pressure, and moisture will change over time. The equations are programmed into a computer and data on the present atmospheric conditions are fed into the computer. The computer solves the equations to determine how the different atmospheric variables will change over the next few minutes. The computer repeats this procedure again and again using the output from one cycle as the input for the next cycle. For some desired time in the future (12, 24, 36, 48, 72 or 120 hours), the computer prints its calculated information. It then analyzes the data, drawing the lines for the projected position of the various pressure systems. The final computer-drawn forecast chart is called a prognostic chart, or prog. A forecaster uses the progs as a guide to predicting the weather. There are many atmospheric models that represent the atmosphere, with each one interpreting the atmosphere in a slightly different way.

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a huge number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a particular application, such as pattern recognition or data classification, through a learning process. The artificial neuron is an information processing unit that is fundamental to the operation of a neural network.

A Decision Tree is a flow-chart-like tree structure. Each internal node denotes a test on an attribute. Each branch represents an outcome of the test. Leaf nodes represent class distribution. The decision tree structure provides an explicit set of “if-then” rules (rather than abstract mathematical equations), making the results easy to interpret. In the tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. In decision analysis, a decision tree can be used visually and explicitly to represent decisions and decision making. The concept of information gain is used to decide the splitting value at an internal node. The splitting value that would provide the most information gain is chosen. Formally, information gain is defined by entropy. In other to improve the accuracy and generalization of classification and regression trees, various techniques were introduced like boosting and pruning. Boosting is a technique for improving the accuracy of a predictive function by applying the function repeatedly in a series and combining the output of each function with weighting so that the total error of the prediction is minimized or growing a number of independent trees in parallel and combine them after all the trees have been developed.

List of Abbreviations

- IT = Information Technology
- SP = Service Provider
- ANN = Artificial Neural Network
- DT = Decision Tree
- RTR = Random Tree Regressor

LITERATURE SURVEY

Literature survey provides the required knowledge about the project and its background. It also helps in following the best practices in project development. Literature survey also helps in understanding the risk and feasibility of the project.

The author uses Naïve Bayes classification method. Weather historical data is collected from Indian Meteorological Department (IMD) Pune. From the collected weather data attributes which are most relevant to rainfall prediction are chosen. Data pre-processing and data transformation on raw weather data set is performed, so that it shall be possible to work on Bayesian, The model is trained using the training data set and has been tested for accuracy on available test data. The meteorological centers use high performance computing and supercomputing power to run weather prediction model. To address the issue of compute intensive rainfall prediction model, Author proposed and implemented data intensive model using data mining technique. The model works with good accuracy and takes moderate compute resources to predict the rainfall. Prediction was found to be working well with good accuracy.

The author investigates the use of data mining techniques in forecasting attributes like maximum temperature, minimum temperature. On available datasets the Decision Tree Algorithm is applied for deleting the inappropriate data.

Analysis and investigation was done using data mining techniques by examining changing patterns of weather parameters which includes maximum temperature, minimum temperature, wind speed and rainfall. After pre-processing of data and outlier analysis, K-means clustering algorithm and Decision Tree algorithm were applied. Two clusters were generated by using K-means Clustering algorithm with lowest and highest of mean parameters. Whereas in decision tree algorithm, a model was developed for modelling meteorological data and it was used to train an algorithm known as the classifier. While for the number of rules generated of the given tree was selected with minimum error of 25%. The results showed that for the given enough set data, these techniques can be used for weather analysis and climate change studies.

This paper describes empirical method technique belonging to clustering and classification and approach. ANNs are used to implement these techniques. The artificial neural networks analyse the data and learn from it for future predictions making them suitable for weather forecasting. Characteristics of neural networks can be used for the prediction of the weather processes. The input variables given are Temperature, Pressure, Relative humidity, Wind speed, perceptible water. The technique used for rainfall prediction is classification. In this technique, rainfall values are clustered using subtractive clustering and three classes or states are identified as low, medium and heavy.

The paper data used for the research work was obtained from meteorological tower of SRM University Chennai, India. Parameters like humidity, temperature, cloud cover, wind speed were used etc. Data Transformation and Data Pre-processing was performed. Different algorithms like Naïve Bayes and C4.5 (J48) Decision Tree algorithm was done simultaneously with dataset containing weather data collected over a period of 2 years. It was found that the performance of C4.5 (J48) decision tree algorithm was far better than that of Naïve Bayes.

Hardware Components:

- Processor – i3
- Hard Disk – 5 GB
- Memory – 1GB RAM

Software Requirements:

- Windows 10, Windows 7(ultimate, enterprise)
- Anaconda (Jupyter Notebook)

PROPOSED MODEL

I used factor analysis and linear regression techniques to find out the most relevant attribute needed for rainfall prediction. We ignored less relevant features in the dataset for model computation. We ignored variables like Station, Date as it had distinct value hence these attributes cannot be used for prediction. We also ignored Station pressure, Gust, Perception amount and Snow depth as these variables had similar duplicates values and factor reduction was not possible. Linear regression technique results shows that Mean Temp, Visibility, Dew point and Wind speed are the best predictors of Rainfall. Therefore we have included these attributes for prediction of rainfall.

The Bayesian Classifier is capable of calculating the most probable output depending on the input. The flow of the model is shown in Fig 2. It is possible to add new raw data at runtime and have a better probabilistic classifier. A Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.

METHEDOLOGY

1. DATA COLLECTION

The data used for this work was collected from Ibadan Synoptic Airport through the Nigerian Meteorological Agency, Oyo State office. The case data covered the period of 120 months, that is, January 2000 to December 2009. The following procedures were adopted at this stage of the research: Data Cleaning, Data Selection, Data Transformation and Data Mining.

2. DATA CLEANING

The data used for this work was collected from Ibadan Synoptic Airport through the Nigerian Meteorological Agency, Oyo State office. The case data covered the period of 120 months, that is, January 2000 to December 2009. The following procedures were adopted at this stage of the research: Data Cleaning, Data Selection, Data Transformation and Data Mining.

3. DATA SELECTION

In this stage, a consistent format for the data model was developed which took care of missing data, finding duplicated data, and weeding out of bad data. Finally, the cleaned data were transformed into a format suitable for data mining.

4. DATA TRANSFORMATION

This is also known as data consolidation. It is the stage in which the selected data is transformed into forms appropriate for data mining. The data file was saved in Commas Separated Value (CVS) file format and the datasets were normalized to reduce the effect of scaling on the data.

5. DATA MINING

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the meteorological datasets. The testing method adopted for this research was percentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge were identified.

Evaluation Metrics

In selecting the appropriate algorithms and parameters that best model the weather forecasting variable, the following performance metrics were used –

1. Correlation Coefficient:

This measures the statistical correlation between the predicted and actual values. This method is unique in that it does not change with a scale in values for the test cases. A higher number means a better model, with a 1 meaning a perfect statistical correlation and a 0 meaning there is no correlation at all.

2. Mean Squared Error:

Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value.

3. The Mean-squared Error

It is simply the square root of the mean-squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.

4. Experimental Design

C5 Decision Tree classifier algorithm which was implemented in See5 was used to analyze the meteorological data. The C5 algorithm was selected after comparison of results of tests carried out using CART and C4.5 algorithms. The ANN algorithms used were those capable of carrying out time series analysis namely: the Time Lagged Feedforward Network (TLFN) and Recurrent networks implemented in NeuroSolutions 6 (an ANN development and simulation software.

IMPLEMENTATION AND CODE

```
import numpy as np

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn import preprocessing

from sklearn.ensemble import RandomForestRegressor

from sklearn.pipeline import make_pipeline

from sklearn.model_selection import GridSearchCV

from sklearn.metrics import mean_squared_error, r2_score

from sklearn.externals import joblib

from sklearn.preprocessing import RobustScaler

import csv
import datetime

from sklearn.svm import SVR

import sklearn.svm as svm

from sklearn.linear_model import LinearRegression

dataset_url1 = 'https://opendata-download-
metobs.smhi.se/api/version/1.0/parameter/2/station/71420/period/corrected-archive/data.csv'

dataset_url2 = 'https://opendata-download-
metobs.smhi.se/api/version/1.0/parameter/2/station/71420/period/latest-months/data.csv'
data1 = pd.read_csv(dataset_url1, sep=';', skiprows=3607, names= [

'Fran Datum Tid (UTC)', 'till', 'day', 'temperature', 'Kvalitet', 'Tidsutsnitt:', 'Unnamed: 5'

])

data2 = pd.read_csv(dataset_url2, sep=';', skiprows=15, names= [

'Fran Datum Tid (UTC)', 'till', 'day', 'temperature', 'Kvalitet', 'Tidsutsnitt:', 'Unnamed: 5'

])

def train_data():

x = data1.drop('Kvalitet', axis = 1)


x = x.drop('Unnamed: 5', axis = 1)

x = x.drop('Fran Datum Tid (UTC)', axis = 1)

x = x.drop('Tidsutsnitt:', axis = 1)
```

```
y = x.temperature

X = x.drop('temperature', axis= 1)


x2 = data2.drop('Kvalitet', axis = 1)

x2 = x2.drop('Unnamed: 5', axis = 1)

# x2 = x2.drop('Till Datum Tid (UTC)', axis = 1)

x2 = x2.drop('Fran Datum Tid (UTC)', axis = 1)

x2 = x2.drop('Tidsutsnitt:', axis = 1)

y2 = x2.temperature

X2 = x2.drop('temperature', axis= 1)


new_dates = []

counter = 0
X = X.append(X2)

dates = X.day

for day in dates:

day = datetime.datetime.strptime(day, "%Y-%m-%d")

day2 = (day - datetime.datetime(1970,1,1)).total_seconds()

new_dates.append(day2)

X.day = new_dates

new_dates= []

for day in X.till:
```

```
day = datetime.datetime.strptime(day, "%Y-%m-%d %H:%M:%S")

day2 = (day - datetime.datetime(1970,1,1)).total_seconds()

new_dates.append(day2)

X.till = new_dates

y = y.append(y2)


X_train, X_test, y_train, y_test = train_test_split(X, y,

test_size=0.5,

random_state=123,
)

scaler = preprocessing.StandardScaler().fit(X_train)

X_train_scaled = scaler.transform(X_train)


pipeline = make_pipeline(preprocessing.StandardScaler()
,
RandomForestRegressor(n_estimators=100))


hyperparameters = { 'randomforestregressor__max_features' : ['auto', 'sqrt', 'log2'],

'randomforestregressor__max_depth': [None, 5, 3, 1], }


clf = LinearRegression()

clf.fit(X_train, y_train)

pred = clf.predict(X_test)
```

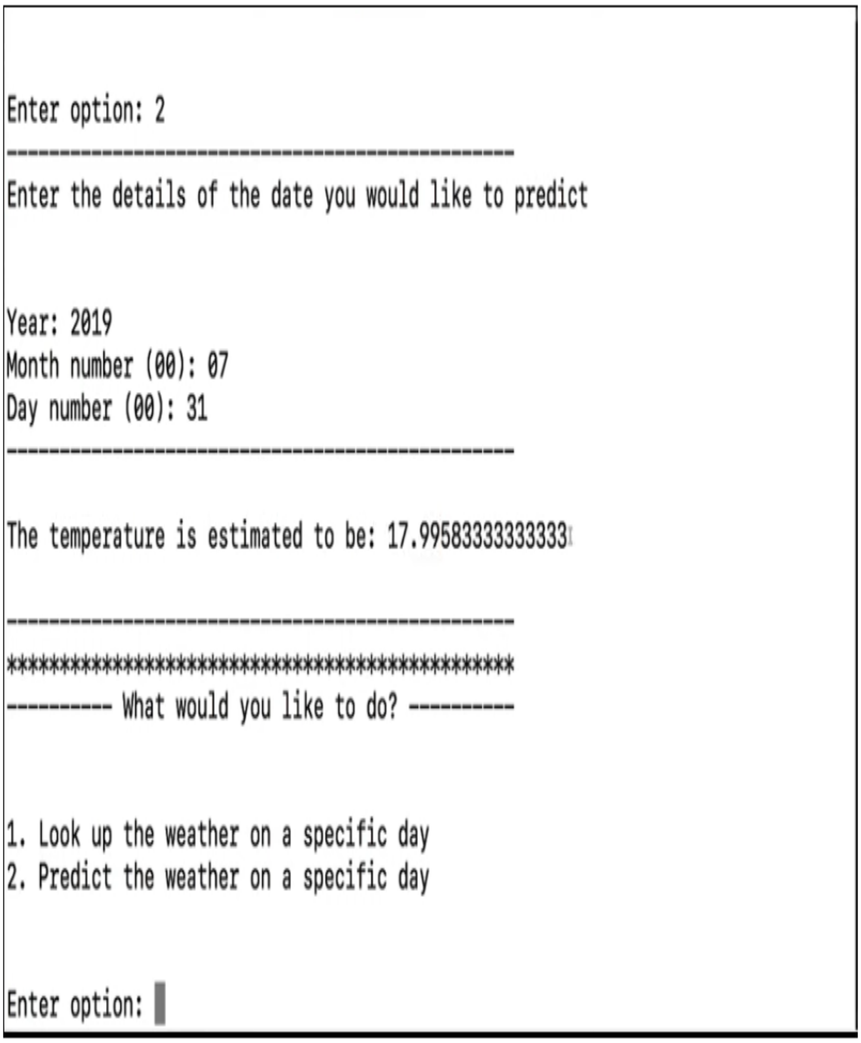
```
print r2_score(y_test, pred)
print mean_squared_error(y_test, pred)

joblib.dump(clf, 'weather_predictor.pkl')\

def get_the_weather(date):
weather = data1.day
temp = data1.temperature

for i in range(0, len(weather)):
day = datetime.datetime.strptime(weather[i], "%Y-%m-%d")
if (day == date):
return temp[i]


def predict_weather():
clf = joblib.load('weather_predictor.pkl')
print("-" * 48)
print("Enter the details of the date you would like to predict")
print("\n")
option = input("Year: ")
year = option
option = input("Month number (00): ")
month = option
```

```
option = input("Day number (00): ")
theday = option

day = str(year) + "-" + str(month) + "-" + str(theday)
day = datetime.datetime.strptime(day, "%Y-%m-%d")
date = (day - datetime.datetime(1970,1,1)).total_seconds()

day_x = str(year) + "-" + str(month) + "-" + str(theday+1)
day_x = datetime.datetime.strptime(day_x, "%Y-%m-%d")
date_x = (day_x - datetime.datetime(1970,1,1)).total_seconds()

X = [[date, date_x]]
print("\n")
print("-" * 48)
print("The temperature is predicted to be: " + str(clf.predict(X)[0]))
print("The temperature was actually: " + str(get_the_weather(day)))
print("-" * 48)
print("\n")

def run_menu():
print("*" *48)
print("-" *10 + " What would you like to do? " + "-" *10)
print("\n")

print("1. Look up the weather on a specific day")

print("2. Predict the weather on a specific day")

print("\n")

option = input("Enter option: ")

while True:

if option == 2 or option == 1 or option == 9:

Break

option = input("Enter option: ")

return option

def run_program(option):

if option == 1:

print("1")

elif option == 2:

predict_weather()


if __name__== "__main__":

train_data()

while True:
```

```
option = run_menu()

if option == 9:

Break

else:

run_program(option)
```

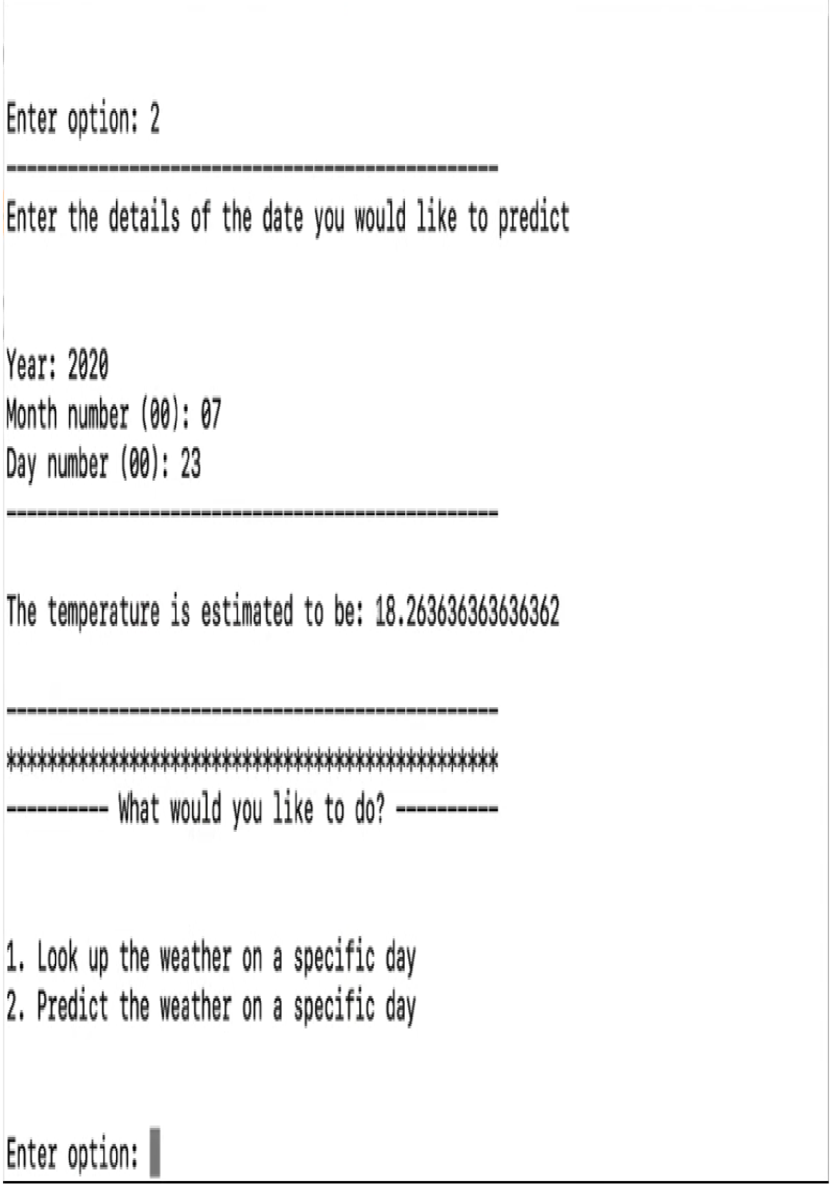
OUTPUT

1.



```
Enter option: 2
------------------------------------------------
Enter the details of the date you would like to predict


Year: 2019
Month number (00): 07
Day number (00): 31
------------------------------------------------


The temperature is estimated to be: 17.99583333333333


------------------------------------------------
************************************************
---------- What would you like to do? ----------


1. Look up the weather on a specific day
2. Predict the weather on a specific day


Enter option: 
```

2.



```
Enter option: 2
------------------------------------------------
Enter the details of the date you would like to predict


Year: 2020
Month number (00): 07
Day number (00): 23
------------------------------------------------


The temperature is estimated to be: 18.263636363636362


------------------------------------------------
************************************************
---------- What would you like to do? ----------


1. Look up the weather on a specific day
2. Predict the weather on a specific day


Enter option: 
```

CONCLUSION

In this report, decision tree and decision tree regressor process were implemented. The Bayesian prediction model can easily learn new classes. The accuracy will grow with the increase of learning data. The model returns good prediction results. The negative part of model is, when a predictor category is not present in the training data, the model assumes that a new record with that category has zero probability. This could be a major issue if this rare predictor value is important. On the hand KNN is effective when data is large but finding unknown patterns like forecasting the future trends. Value for temperature in decision tree regression technique is difficult to determine. In this report the Bayesian model proved to be more accurate then the K-NN Model.

Artificial Neural Networks can detect the relationships between the input variables and generate outputs based on the observed patterns inherent in the data without any need for programming or developing complex equations to model these relationships. Hence given enough data ANN’s can detect the relationships between weather parameter and use these to predict future weather conditions. Both TLFN neural networks and Recurrent network architectures were used to developed predictive ANN models for the prediction of future values of Wind speed, Evaporation, Radiation, Minimum Temperature, Maximum Temperature and Rainfall given the Month and Year.

REFERENCE

- https://opendata-download-metobs.smhi.se/api/version/1.0/parameter/2/station/71420/period/corrected-archive/data.csv

- https://opendata-download-metobs.smhi.se/api/version/1.0/parameter/2/station/71420/period/latest-months/data.csv

- http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6663175&queryText%3DWeather+Forecasting+Using+Data+Mining

- http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=4136992&queryText%3DWeather+Forecasting+Using+Data+Mining