IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STATISTICAL ANALYSIS AND WEATHER PREDICTION USING MACHINE LEARNING MODEL

Mr. Manjunath M Katti , Sudarshan K, Charan B, Tontadharya M G , Vimala B,

Assistant Professor, CIT, Gubbi, 8th sem student,CIT, Gubbi, 8th sem student,CIT, Gubbi, 8th sem student,CIT, Gubbi, 8th sem student,CIT, Gubbi

Department of Civil Engineering,
Channabasaveshwara Institute of Technology, Tumkur, India

Abstract: Weather is a vital part of a person's life because it can tell us whether it will rain or be sunny. Weather forecasting is meteorologists attempt to predict weather conditions in the future, as well as weather conditions that may be predicted. Temperature, pressure, humidity, dew point, rainfall, precipitation, wind speed, and dataset size are all used to calculate the climatic state parameters. To begin, the data must be educated. We can use the data from the data collection to train the data. User can predict daily weather report using the particular dataset which is already imported like API. Here API is vital role for current data. Using this project user can get current weather report and only need the input of name of the city once enter the city name full report will be generated for current situation. We can analyze the predict temperature with original temperature and can predict future rain fall. We'll use the Decision Tree Classification Algorithm and the Random Forest Classification Algorithm to make this prediction. Python, NumPy, Jupiter Notebook, Sea born, and Panda will be used in this project. The project is split into three separate Jupiter Notebooks: one to collect the weather data, inspect it, and clean it; a second to further refine the features and fit the data to a Decision Tree model and a third to fit the data to a Random Forest model evaluate our output.

KEYWORDS: Machine learning, weather prediction, statistical analysis, rainfall, temperature.

1. INTRODUCTION

The statistical analysis and Weather prediction using machine learning model is a complex task that involves understanding and forecasting various atmospheric variables such as temperature, humidity, wind speed, and precipitation. Traditionally, meteorologists have relied on numerical weather prediction (NWP) models based on physical equations and simulations to forecast the weather. While these methods have been successful, they can sometimes fall short in certain scenarios or geographic regions due to the intricate nature of weather systems. Recently, machine learning (ML) models have gained popularity in weather prediction due to their ability to identify patterns in large datasets and make predictions based on those patterns. Machine learning algorithms can complement traditional NWP models and provide alternative methods for forecasting weather. Weather forecasting is the application of science and technology to predict atmospheric conditions at specific places and times. People have tried to predict the weather officially for thousands of years since the 19th century. After manual calculations, primarily based on changes in, current weather conditions, sky conditions or cloud cover, weather forecasts now rely on computer-based models, taking into account many atmospheric factors.

2. OBJECTIVES

- Develop models that accurately predict weather conditions such as temperature, precipitation, wind patterns, etc., with higher precision than traditional methods.
- Build models capable of providing early warnings for weather events like temperature, dew, or wind speed enabling timely preparedness and response measures.
- Utilize machine learning algorithms to recognize complex patterns and correlations within meteorological data, allowing for better insight into atmospheric processes.

3. METHODOLOGY

- Data Collection: Gather historical weather data from various sources (GitHub & Kaggle), including government weather stations and online APIs. Real-time weather data can be collected from online weather APIs, such as Open Weather Map

- Data Pre-processing: Clean and pre-process the collected data to handle missing values, outliers, and inconsistencies. Feature engineering to extract relevant information from raw data.

- User Interface: Develop a user-friendly interface that allows users to input location and time parameters for weather forecasts. Provide graphical representations of weather forecasts and historical weather data.

- Visualization: Utilize data visualization libraries like Matplotlib or seaborn to create interactive and informative weather charts and graphs.

- Data correlation: To analyze the correlation between variables in a dataset. It creates using Seaborn and Matplotlib provides a visual representation of the correlation matrix. making it easier to interpret the relationships between variables.

- Data Describe: To get a quick overview of your dataset, you can use the describe () function in the pandas library in Python. This function provides summary statistics of your numerical columns. Including count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values.

- Build a model using ML: The code for training a machine learning model can vary depending on the specific task, algorithm, and programming language you are using. Below, I'll provide a simple example using Python and the popular scikit-learn library for a basic classification task with a decision tree algorithm.

4. STUDY AREA

4.1 Model Implementation for Decision Tree Classifier.

```
# Splitting in independent and dependent
x = data.drop(columns=['Weather','Date/Time'])  #independent
y = data["Weather"] #dependent
```

```
x
```

	Temp_C	Dew Point Temp_C	relative humudity%	Wind Speed_km/h	Visibility_km	Pressure_kPa
0	-1.8	-3.9	86.0	4.0	22.75	101.24
1	-1.8	-3.7	87.0	4.0	22.75	101.24
2	-1.8	-3.4	89.0	7.0	22.75	101.26
3	-1.5	-3.2	88.0	6.0	22.75	101.27
4	-1.5	-3.3	88.0	7.0	22.75	101.23
...	...	...	...	...	...	...
8779	0.1	-2.7	81.0	30.0	22.75	100.13
8780	0.2	-2.4	83.0	24.0	22.75	100.03
8781	-0.5	-1.5	93.0	28.0	22.75	99.95
8782	-0.2	-1.8	89.0	28.0	22.75	99.91
8783	0.0	-2.1	86.0	30.0	22.75	99.89

8784 rows × 6 columns

```
y
```

```
0       2
1       2
2       4
3       4
4       2
       ..
8779    6
8780    6
8781    6
8782    6
8783    6
Name: Weather, Length: 8784, dtype: int32
```

Sklearn is a popular machine learn library for Python. It provides a range of algorithms for classification, regression, clustering, decomposition, dimensionality reduction, and more. Sklearn also includes tools for model selection, cross-validation, and evaluation. It is widely used in industry and academia for a variety of applications and is constantly being updated with new algorithms and features.

```
# split data into training and testing
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x, y, test_size=0.2, random_state=20)
```

The 'x train' variable in Python represents the feature matrix and labels of the data trained on the 'x train' data. In other words, it is a placeholder for the data that will be used to train a machine learning model. The practice of prepending an 'x' to variable names is commonly used the Python machine learning community to indicate that the variable holds data.

```
x_train
```

	Temp_C	Dew Point Temp_C	relative humudity%	Wind Speed_km/h	Visibility_km	Pressure_kPa
3988	13.0	9.0	77.0	15.0	24.10	102.19
7116	11.1	3.8	61.0	0.0	24.10	101.62
1451	-4.4	-6.8	83.0	28.0	22.75	100.66
8245	-2.3	-5.4	79.0	9.0	24.10	102.28
6441	16.3	8.6	60.0	32.0	26.35	100.93
...	...	...	...	...	...	...
5910	13.4	10.7	84.0	6.0	26.35	101.68
3915	20.2	14.3	69.0	17.0	25.00	101.00
7068	12.6	6.4	66.0	32.0	24.10	100.30
7391	4.3	-0.1	73.0	17.0	25.00	100.82
4367	22.3	15.6	66.0	4.0	25.00	100.07

7027 rows × 6 columns

The "y" in "y train" is likely a reference to the variable "y" commonly used in machine learning algorithms. In Python, you can train a machine learning model using various libraries.These libraries provide functions and tools for data preprocessing, model building, and optimization. To train a machine learning model,

you typically need to have a dataset and specify the model architecture, loss function, and optimization algorithm.

```
y_train
```

```
3988    0
7116    0
1451    6
8245    1
6441    0
       ..
5910    0
3915    1
7068    1
7391    1
4367    1
Name: Weather, Length: 7027, dtype: int32
```

Python is a popular programming language that is widely used for a variety of tasks, including machine learning, web development, and data analysis. X is a variable that can be used in Python to store a value, such as an integer, float, or string. To use a variable in Python, you first assign a value to it using the `=` operator.

```
x_test
```

	Temp_C	Dew Point Temp_C	relative humudity%	Wind Speed_km/h	Visibility_km	Pressure_kPa
5778	20.0	7.3	44.0	20.0	26.35	100.79
865	-0.8	-5.2	72.0	24.0	25.00	100.80
3566	19.1	7.8	48.0	28.0	26.35	100.98
1773	-0.1	-0.9	94.0	17.0	22.75	102.17
5559	23.5	11.3	46.0	17.0	26.35	100.77
...	...	...	...	...	...	...
3568	17.3	6.4	49.0	22.0	24.10	100.88
916	-14.0	-21.0	55.0	4.0	25.00	102.13
6254	24.6	12.6	47.0	15.0	24.10	100.95
509	-19.5	-23.4	71.0	6.0	22.75	102.84
6776	8.4	6.0	85.0	15.0	26.35	102.08

1757 rows × 6 columns

The "y" in Python represents a variable that stores a value. The "test" describes a function that will be tested. Testing in Python refers to verifying the functionality of a program or replacing incorrect code with new code. This process helps the programmer to ensure that the program runs without errors. Various types of testing can be done in Python, including unit testing, regression testing, and integration testing. A well-written test can help to maintain the codebase and ensure that the code functions as expected.

A decision tree classifier is a type of machine learning algorithm used for classification tasks. It works by recursively partitioning the input space based on which feature or attribute has the most impact on the target variable.

A data model fit test, also known as model validation or model selection, is a technique used to evaluate the compatibility of a data model with a set of data. It involves comparing the predictions or outputs of the model with the actual data to determine if there is a good fit.

```
# Model building
from sklearn.tree import DecisionTreeClassifier
```

Data prediction is the process of making future predictions based on previously collected data. Fit tests, also known as supposedly simple linear regression analysis (SSLRA), are used to evaluate the linearity of a regression model and assess how well the model fits the data. The test involves calculating the correlation coefficient between the predicted and observed values, and comparing it to a critical value or using a statistical test to determine if the correlation coefficient is statistically significant. If the correlation coefficient is greater than the critical value, it suggests that the model is a good fit for the data.

```
accuracy_score(data_pred,y_test)
```

```
0.6260671599317018
```

```
accuracy_score(data_pred,y_test) * 100
```

```
62.606715993170184
```

In Decision Tree Classifier Algorithm we got the accuracy of 62.060 % from the data sets

4.2 Model Implementation for Random forest Classifier

```
# Model building
from sklearn.ensemble import RandomForestClassifier
```

A decision tree classifier is a type of machine learning algorithm used for classification tasks. It works by recursively partitioning the input space based on which feature or attribute has the most impact on the target variable.

```
data_model_RF=RandomForestClassifier(n_estimators=100,random_state=23)
```

```
data_model_RF.fit(x_train,y_train)
```

```
RandomForestClassifier(random_state=23)
```

Random forest is a popular machine learning algorithm used for classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. The random forest tree classifier is a type of decision tree classifier that uses random feature selection and random tree splitting to make predictions. It is commonly used in image classification, text classification, and other classification problems. The random forest tree classifier is known for its high accuracy, robustness, and ability to handle large datasets.

Data prediction is the process of making future predictions based on previously collected data. Fit tests, also known as supposedly simple linear regression analysis (SSLRA), are used to evaluate the linearity of a regression model and assess how well the model fits the data. The test involves calculating the correlation coefficient between the predicted and observed values, and comparing it to a critical value or using a statistical test to determine if the correlation coefficient is statistically significant. If the correlation coefficient is greater than the critical value, it suggests that the model is a good fit for the data.

```
accuracy_score(data_pred_RF,y_test)
```

```
0.7046101309049516
```

```
accuracy_score(data_pred_RF,y_test) * 100
```

```
70.46101309049516
```

In Random Forest Classifier Algorithm we got the accuracy of 70.461 % from the data sets

5. RESULT AND DISCUSSION

In Decision Tree Classifier Algorithm we got the accuracy of 62.060 % from the data sets

```
from sklearn.metrics import *
```

```
accuracy_score(data_pred,y_test)
```

```
0.6260671599317018
```

```
accuracy_score(data_pred,y_test) * 100    # accuracy % is tooo low because of weather column as 50 label encoders
```

```
62.606715993170184
```

In Random Forest Classifier Algorithm we got the accuracy of 70.461 % from the data sets

```
from sklearn.metrics import *
```

```
accuracy_score(data_pred_RF,y_test)
```

```
0.7046101309049516
```

```
accuracy_score(data_pred_RF,y_test) * 100    # accuracy % is tooo low because of weather column as 50 label encoders
```

```
70.46101309049516
```

The Decision tree classifier shows less result compare to random forest classifier because Decision Trees have high variance, because their predictions can change significantly with small changes in the training data and Random Forests, being an ensemble method, reduce variance by averaging multiple decision trees. Each tree is trained on a different subset of the data (using bootstrapping), and features are randomly selected at each

split. This diversity among trees leads to a more balanced bias-variance tradeoff, resulting in better generalization. This instability makes them less reliable compared to ensemble methods.

6. WHAT IS THE INNOVATION IN THE PROJECTS?

Advanced Predictive Modeling Using machine learning (ML) techniques to analyze and predict weather patterns can significantly improve accuracy compared to traditional statistical methods. ML models can handle large datasets, identify complex patterns, and update predictions in real-time as new data becomes available.

Conclusion

The study on statistical analysis and weather prediction using machine learning models reveals significant advancements in forecasting accuracy over traditional methods. By leveraging models such as Random Forest, Decision Tree. The model which is prepared has an accuracy of is 70.461 and 62.060 percentages. The Decision tree classifier shows less result compare to random forest classifier because Decision Trees have high variance, because their predictions can change significantly with small changes in the training data and Random Forests, being an ensemble method, reduce variance by averaging multiple decision trees. Each tree is trained on a different subset of the data (using bootstrapping), and features are randomly selected at each split. This diversity among trees leads to a more balanced bias-variance tradeoff, resulting in better generalization. This instability makes them less reliable compared to ensemble methods. This model helps Future directions include integrating diverse datasets and developing hybrid models, with the potential to significantly benefit sectors like agriculture, disaster management, and renewable energy through improved weather forecasts.

Reference

- Anaconda software HTTPS://WWW.ANACONDA.COM ›
- Juputer notebook launch https://jupyter.org
- Data set From HTTPS://WWW.KAGGLE.COM/DATASETS/CLEOPHASCHEBII/Weather Data