
Predicting Rainfall Using Machine Learning Techniques

Peddaguttapalle Brijesh¹, M.Sudhakar²

¹M.Tech Student, Department of CSE, Golden Valley Integrated Campus, Madanapalli

²Associate Professor, Department of CSE, Golden Valley Integrated Campus, Madanapalli

Abstract

Rainfall prediction is one of the challenging and uncertain tasks which has a significant impact on human society. Timely and accurate predictions can help to proactively reduce human and financial loss. This study presents a set of experiments which involve the use of prevalent machine learning techniques to build models to predict whether it is going to rain tomorrow or not based on weather data for that particular day in major cities of Australia. This comparative study is conducted concentrating on three aspects: modeling inputs, modeling methods, and pre-processing techniques. The results provide a comparison of various evaluation metrics of these machine learning techniques and their reliability to predict the rainfall by analyzing the weather data.

Keywords: rainfall, financial loss, prediction, machine learning

INTRODUCTION

India's welfare is agriculture. The achievement of agriculture is dependent on rainfall. It also helps with water resources. Rainfall information in the past helps farmers better manage their crops, leading to economic growth in the country. Prediction of precipitation is beneficial to prevent flooding that saves people's lives and property. Fluctuation in the timing of precipitation and its amount makes forecasting of rainfall a problem for meteorological scientists. Forecasting is one of the utmost challenges for researchers from a variety of fields, such as weather data mining, environmental machine learning, functional hydrology, and numerical forecasting, to create a predictive model for accurate rainfall. In these problems, a common question is how to infer the past predictions and make use of future predictions. A variety of sub-processes are typically composed of the substantial process in rainfall. It is at times not promising to predict the precipitation correctly by on its global system. Climate forecasting stands out for all countries around the globe in all the benefits and services provided by the meteorological department. The job is very complicated because it needs specific numbers and all signals are intimated without any assurance. Accurate precipitation forecasting has been an important issue in hydrological science as early notice of stern weather can help avoid natural disaster injuries and damage if prompt and accurate forecasts are made. The theory of the modular model and the integrati2on of different models has recently gained more interest in rainfall forecasting to address this challenge. A huge range of rainfall prediction methodologies is available in India. In India, there are two primary methods of forecasting rainfall. Regression, Artificial Neural Network (ANN), Decision Tree algorithm, Fuzzy logic and team process of data handling are the majority frequently used computational methods used for weather forecasting The basic goal is to follow information rules and relationships while gaining intangible and potentially expensive knowledge. Artificial NN is a promising part of this wide field

Rainfall prediction remains a serious concern and has attracted the attention of governments, industries, risk management entities, as well as the scientific community. Rainfall is a climatic factor that affects many human activities like agricultural production, construction, power generation, forestry and tourism, among others [1]. To this extent, rainfall prediction is essential since this variable is the one with the highest correlation with adverse natural events such as

landslides, flooding, mass movements and avalanches. These incidents have affected society for years [2]. Therefore, having an appropriate approach for rainfall prediction makes it possible to take preventive and mitigation measures for these natural phenomena .

To solve this uncertainty, we used various machine learning techniques and models to make accurate and timely predictions. These paper aims to provide end to end machine learning life cycle right from Data preprocessing to implementing models to evaluating them. Data Preprocessing steps include imputing missing values, feature transformation, encoding categorical features, feature scaling and feature selection. We implemented models such as Logistic Regression, Decision Tree, K Nearest Neighbour, Rule-based and Ensembles. For evaluation purpose.

LITERATURE SURVEY

1.Climate Change and Human Health: Risks and Responses

The long-term good health of populations depends on the continued stability and functioning of the biosphere's ecological and physical systems, often referred to as life-support systems. We ignore this long-established historical truth at our peril: yet it is all too easy to overlook this dependency, particularly at a time when the human species is becoming increasingly urbanized and distanced from these natural systems. The world's climate system is an integral part of this complex of life-supporting processes, one of many large natural systems that are now coming under pressure from the increasing weight of human numbers and economic activities.

By inadvertently increasing the concentration of energy-trapping gases in the lower atmosphere, human actions have begun to amplify Earth's natural greenhouse effect. The primary challenge facing the world community is to achieve sufficient reduction in greenhouse gas emissions so as to avoid dangerous interference in the climate system. National governments, via the UN Framework Convention on Climate Change (UNFCCC), are committed in principle to seeking this outcome. In practice, it is proving difficult to find a politically acceptable course of action—often because of apprehensions about possible short-term economic consequences.

This volume seeks to describe the context and process of global climate change, its actual or likely impacts on health, and how human societies should respond, via both adaptation strategies to lessen impacts and collective action to reduce greenhouse gas emissions. As shown later, much of the resultant risk to human populations and the ecosystems upon which they depend comes from the projected extremely rapid rate of change in climatic conditions. Indeed, the prospect of such change has stimulated a great deal of new scientific research over the past decade, much of which is elucidating the complex ecological disturbances that can impact on human well-being and health—as in the following example.

The US Global Change Research Program (Alaska Regional Assessment Group) recently documented how the various effects of climate change on aquatic ecosystems can interact and ripple through trophic levels in unpredictable ways. For example, warming in the Arctic region has reduced the amount of sea ice, impairing survival rates for walrus and seal pups that spend part of their life cycle on the ice. With fewer seal pups, sea otters have become the alternative food source for whales. Sea otters feed on sea urchins, and with fewer sea otters sea urchin populations are expanding and consuming more of the kelp that provides breeding grounds for fish. Fewer fish exacerbate the declines in walrus and seal populations. Overall, there is less food available for the Yupik Eskimos of the Arctic who rely on all of these species.

Global climate change is thus a significant addition to the spectrum of environmental health hazards faced by humankind. The global scale makes for unfamiliarity—although most of its health impacts comprise increases (or decreases) in familiar effects of climatic variation on human biology and health. Traditional environmental health concerns long have been focused on toxicological or microbiological risks to health from local environmental exposures. However, in the early years of the twenty-first century, as the burgeoning human impact on the environment continues to alter the planet's geological, biological and ecological systems, a range of larger-scale environmental hazards to human health has emerged. In addition to global climate change, these include: the health

risks posed by stratospheric ozone depletion; loss of biodiversity; stresses on terrestrial and ocean food-producing systems; changes in hydrological systems and the supplies of freshwater; and the global dissemination of persistent organic pollutants.

Climate change and stratospheric ozone depletion are the best known of these various global environmental changes. Human societies, however, have had long experience of the vicissitudes of climate: climatic cycles have left great imprints and scars on the history of humankind. Civilisations such as those of ancient Egypt, Mesopotamia, the Mayans, the Vikings in Greenland and European populations during the four centuries of Little Ice Age, all have both benefited and suffered from nature's great climatic cycles. Historical analyses also reveal widespread disasters, social disruption and disease outbreaks in response to the more acute, inter-annual, quasi-periodic ENSO (El Niño Southern Oscillation) cycle (1). The depletion of soil fertility and freshwater supplies, and the mismanagement of water catchment basins via excessive deforestation, also have contributed to the decline of various regional populations over the millennia.

2. Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries

The significance of the prevention of natural disasters is made evident by the commemoration of the International Decade for Natural Disaster Reduction (IDNDR). This paper focuses on the role of geomorphology in the prevention of natural disasters in developing countries, where their impact has devastating consequences. Concepts such as natural hazards, natural disasters and vulnerability have a broad range of definitions; however, the most significant elements are associated with the vulnerability concept. The latter is further explored and considered as a key factor in understanding the occurrence of natural disasters, and consequently, in developing and applying adequate strategies for prevention. Terms such as natural and human vulnerabilities are introduced and explained as target aspects to be taken into account in the reduction of vulnerability and for prevention and mitigation of natural disasters. The importance of the incorporation not only of geomorphological research, but also of geomorphologists in risk assessment and management programs in the poorest countries is emphasized.

3. Atmospheric and climatic hazards: Improved monitoring and prediction for disaster mitigation

The last few years have seen enormous damage and loss of life from climate and weather phenomena. The most damaging events have included the severe 1997/98 El Niño (with its near-global impacts), Hurricane Mitch, and floods in China in mid-1998. What have we learnt regarding the causes, variability, and predictability, of these phenomena? Can we predict the occurrence of these extreme events, and thereby mitigate their damage? This paper reviews what we have learnt in the last decade or so regarding the predictability of these climate and weather extremes. The view starts with the largest (El Niño) scales, and works towards the scale of individual thunderstorms. It focuses on the practical outcomes of our improved knowledge with regard to decreasing the impact of natural disasters, rather than describing in detail the scientific knowledge underlying these outcomes. The paper concludes with a discussion of some of the factors that still restrict our ability to mitigate the deleterious effects of atmospheric and climatic hazards.

4. Exploratory Data Analysis: the Best way to Start a Data Science Project

Exploratory Data Analysis is a set of techniques that were developed by Tukey, John Wilder in 1970. The philosophy behind this approach was to examine the data before building a model. John Tukey encouraged statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. Today Data scientists and analysts spend most of their time in Data Wrangling and Exploratory Data Analysis also known as EDA. But what is this EDA and why it is so important? This article explains what is EDA and how to apply EDA techniques to a dataset.

EXISTING SYSTEM

Rainfall prediction is important as heavy rainfall can lead to many disasters. The prediction helps people to take preventive measures and moreover the prediction should be accurate. There are two types of prediction short term rainfall prediction and long term rainfall. Prediction mostly short term prediction can give us the accurate result. The main challenge is to build a model for long term rainfall prediction. Heavy precipitation prediction could be a major drawback for earth science department because it is closely associated with the economy and lifetime of human.

Disadvantages

We can just do it by having the historical data analysis of rainfall and can predict the rainfall for future seasons. We can apply many techniques like classification, regression according to the requirements and also we can calculate the error between the actual and prediction and also the accuracy. Different techniques produce different accuracies so it is important to choose the right algorithm and model it according to the requirements

PROPOSED SYSTEM

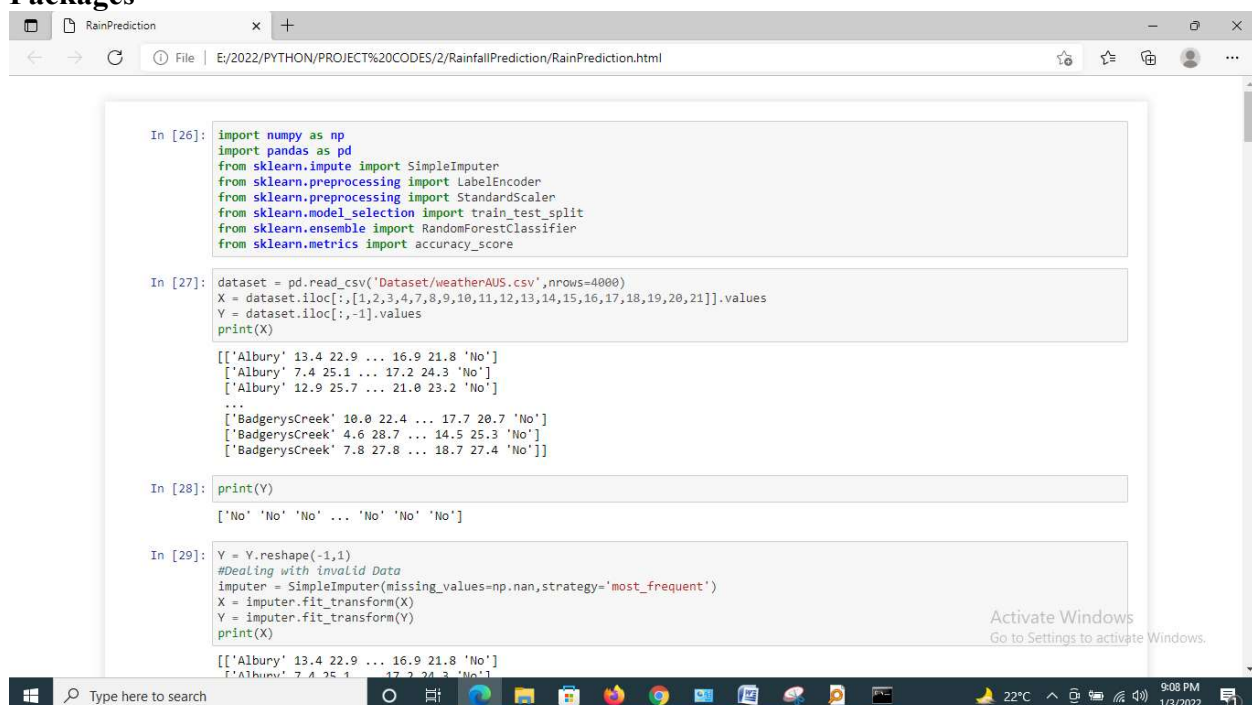
It's a cause for natural disasters like flood and drought that square measure encountered by individuals across the world each year. Accuracy of rainfall statement has nice importance for countries like India whose economy is basically dependent on agriculture. The dynamic nature of atmosphere, applied mathematics techniques fail to provide sensible accuracy for precipitation statement. The prediction of precipitation using machine learning techniques may use regression. Intention of this project is to offer non-experts easy access to the techniques, approaches utilized in the sector of precipitation prediction and provide a comparative study among the various machine learning techniques.

Advantages

1. It is a powerful technique for testing relationship between one dependent variable and many independent variables.
2. It allows researchers to control extraneous factors.
3. Regression assesses the cumulative effect of multiple factors.
4. It also helps to attain the measure of error using the regression line as a base for estimations.

RESULTS

Packages



```
In [26]: import numpy as np
import pandas as pd
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

In [27]: dataset = pd.read_csv('Dataset/weatherAUS.csv', nrows=4000)
X = dataset.iloc[:, [1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]].values
Y = dataset.iloc[:, -1].values
print(X)

[['Albury' 13.4 22.9 ... 16.9 21.8 'No']
['Albury' 7.4 25.1 ... 17.2 24.3 'No']
['Albury' 12.9 25.7 ... 21.0 23.2 'No']
...
['BadgenysCreek' 10.0 22.4 ... 17.7 20.7 'No']
['BadgenysCreek' 4.6 28.7 ... 14.5 25.3 'No']
['BadgenysCreek' 7.8 27.8 ... 18.7 27.4 'No']]

In [28]: print(Y)

['No' 'No' 'No' ... 'No' 'No' 'No']

In [29]: Y = Y.reshape(-1,1)
#Dealing with Invalid Data
imputer = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
X = imputer.fit_transform(X)
Y = imputer.fit_transform(Y)
print(X)

[['Albury' 13.4 22.9 ... 16.9 21.8 'No']
['Albury' 7.4 25.1 ... 17.2 24.3 'No']
```



Analysis

```
In [30]: #Encoding Dataset
le1 = LabelEncoder()
X[:,0] = le1.fit_transform(X[:,0])
le2 = LabelEncoder()
X[:,4] = le2.fit_transform(X[:,4])
le3 = LabelEncoder()
X[:,6] = le3.fit_transform(X[:,6])
le4 = LabelEncoder()
X[:,7] = le4.fit_transform(X[:,7])
le5 = LabelEncoder()
X[:,1] = le5.fit_transform(X[:,1])
le6 = LabelEncoder()
Y[:,1] = le6.fit_transform(Y[:,1])
print(X)

[[0 13.4 22.9 ... 16.9 21.8 0]
 [0  7.4 25.1 ... 17.2 24.3 0]
 [0 12.9 25.7 ... 21.0 23.2 0]
 ...
 [1 10.0 22.4 ... 17.7 20.7 0]
 [1  4.6 20.7 ... 14.5 25.3 0]
 [1  7.0 27.8 ... 18.7 27.4 0]]

In [31]: print(Y)
[[0]
 [0]
 [0]
 ...
 [0]
 [0]]

In [32]: Y = np.array(Y,dtype=float)
print(Y)
[[0.]
 [0.]
 [0.]
 ...
 [0.]
 [0.]]
```

Training

```
In [32]: Y = np.array(Y,dtype=float)
print(Y)
[[0.]
 [0.]
 [0.]
 ...
 [0.]
 [0.]]

In [33]: #Feature Scaling
sc = StandardScaler()
X = sc.fit_transform(X)
print(X)

[[-5.61951487e-01  5.87241678e-01  1.54909661e-03 ...  3.42299439e-01
  3.88963551e-02 -4.99609344e-01]
 [-5.61951487e-01 -4.85310961e-01  2.96615118e-01 ...  3.91129021e-01
  3.84764709e-01 -4.99609344e-01]
 [-5.61951487e-01  5.84487292e-01  3.77087669e-01 ...  1.00963706e+00
  2.32582633e-01 -4.99609344e-01]
 ...
 [ 1.77951304e+00  2.45118493e-02 -6.55113628e-02 ...  4.72511657e-01
 -1.13285721e-01 -4.99609344e-01]
 [ 1.77951304e+00 -8.69235526e-01  7.79458426e-01 ... -4.83372170e-02
  5.23112851e-01 -4.99609344e-01]
 [ 1.77951304e+00 -3.39607452e-01  6.58741599e-01 ...  6.35276931e-01
  8.13641468e-01 -4.99609344e-01]]

In [34]: #Splitting Dataset into Training set and Test set
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.2,random_state=0)
print(X_train)

[[-0.56195140  0.42173291  0.44414813 ...  0.30974638  0.45393838
 -0.49960934]
 [-0.56195140 -1.13404956 -1.1653029 ... -1.18769413 -1.04021291]]
```

Algorithms

```
In [138]: #Training Model
classifier = RandomForestClassifier(n_estimators=100,random_state=0)
classifier.fit(X_train,Y_train)
print(classifier.score(X_train,Y_train))
y_pred = le6.inverse_transform(np.array(classifier.predict(X_test),dtype=int))
Y_test1 = le6.inverse_transform(np.array(Y_test,dtype=int))
print(y_pred)

c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:3: DataConversionWarning: A column n-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using.ravel().
This is separate from the ipykernel package so we can avoid doing imports until

0.9996075
["No" "No" "No" "No" "No" "No" "Yes" "No" "Yes" "No" "No"
 "No" "No" "No" "No" "No" "Yes" "No" "No" "No" "No" "No"
 "No" "No" "No" "No" "No" "No" "No" "No" "Yes" "No" "No"
 "No" "No" "No" "Yes" "No" "No" "No" "No" "No" "No" "No"
 "No" "No" "No" "Yes" "No" "No" "Yes" "No" "No" "No" "No"
 "Yes" "No" "No" "No" "No" "No" "No" "No" "Yes" "No" "No"
 "Yes" "No" "No" "No" "Yes" "No" "No" "Yes" "No" "No" "Yes"
 "No" "No" "No" "No" "Yes" "No" "No" "No" "No" "No" "Yes"
 "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 "Yes" "No" "No" "No" "Yes" "No" "No" "No" "No" "No" "No"
 "No" "No" "No" "No" "Yes" "No" "No" "No" "No" "No" "No"
 "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "No" "No" "No" "Yes"
 "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "No" "No" "No" "Yes"
 "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "No" "No" "No" "Yes"
 "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 "No" "No" "No" "Yes" "No" "No" "Yes" "No" "No" "No" "No"
 "No" "No" "No" "No" "No" "No" "No" "Yes" "No" "No" "No"
 "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 "No" "No" "No" "No" "No" "Yes" "No" "No" "No" "No" "No"]
```

Random Forest

```
File | E:/2022/PYTHON/PROJECT%20CODES/2/RainfallPrediction/RainPrediction.html
In [141]: rf_accuracy = accuracy_score(Y_test1,y_pred)
print("\nRandom Forest Accuracy: "+str(rf_accuracy))
Random Forest Accuracy: 0.87375

In [142]: from sklearn.ensemble import BaggingClassifier
dt = BaggingClassifier(n_estimators=250,max_features=12)
dt.fit(X_train,Y_train)
print(dt.score(X_train,Y_train))
y_pred = le6.inverse_transform(np.array(dt.predict(X_test),dtype=int))
Y_test2 = le6.inverse_transform(np.array(Y_test,dtype=int))
y_pred = y_pred.reshape(-1,1)
Y_test2 = Y_test2.reshape(-1,1)
df = np.concatenate((Y_test2,y_pred),axis=1)
dataframe = pd.DataFrame(df,columns=['Rain on Tomorrow','Prediction of Rain'])
print(dataframe)
dt_accuracy = accuracy_score(Y_test1,y_pred)
print("\nBagging Classifier Accuracy: "+str(dt_accuracy))
#print(y_pred)
#print(Y_test)

c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\sklearn\ensemble\_bagging.py:645: DataConversionWarnin
g: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using
ravel().
  y = column_or_1d(y, warn=True)

1.0
Rain on Tomorrow Prediction of Rain
0 No No
1 No No
2 No No
3 No No
4 No No
... ..
795 No No
796 No Yes
797 No No
798 Yes Yes
799 No No

[800 rows x 2 columns]

Bagging Classifier Accuracy: 0.8775

c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\sklearn\preprocessing\_label.py:289: DataConversionWarnin
g: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example usin
g ravel().
  y = column_or_1d(y, warn=True)
```

Bagging Classifier

```
File | E:/2022/PYTHON/PROJECT%20CODES/2/RainfallPrediction/RainPrediction.html
dt.fit(X_train,Y_train)
print(dt.score(X_train,Y_train))
y_pred = le6.inverse_transform(np.array(dt.predict(X_test),dtype=int))
Y_test2 = le6.inverse_transform(np.array(Y_test,dtype=int))
y_pred = y_pred.reshape(-1,1)
Y_test2 = Y_test2.reshape(-1,1)
df = np.concatenate((Y_test2,y_pred),axis=1)
dataframe = pd.DataFrame(df,columns=['Rain on Tomorrow','Prediction of Rain'])
print(dataframe)
dt_accuracy = accuracy_score(Y_test1,y_pred)
print("\nBagging Classifier Accuracy: "+str(dt_accuracy))
#print(y_pred)
#print(Y_test)

c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\sklearn\ensemble\_bagging.py:645: DataConversionWarnin
g: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example usin
g ravel().
  y = column_or_1d(y, warn=True)

1.0
Rain on Tomorrow Prediction of Rain
0 No No
1 No No
2 No No
3 No No
4 No No
... ..
795 No No
796 No Yes
797 No No
798 Yes Yes
799 Yes Yes

[800 rows x 2 columns]

Bagging Classifier Accuracy: 0.8775

c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\sklearn\preprocessing\_label.py:289: DataConversionWarnin
g: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example usin
g ravel().
  y = column_or_1d(y, warn=True)
```

Gradient Boosting

```
File | E:/2022/PYTHON/PROJECT%20CODES/2/RainfallPrediction/RainPrediction.html
In [143]: from sklearn.ensemble import GradientBoostingClassifier
dt = GradientBoostingClassifier(n_estimators=170,max_depth=1)
dt.fit(X_train,Y_train)
print(dt.score(X_train,Y_train))
y_pred = le6.inverse_transform(np.array(dt.predict(X_test),dtype=int))
Y_test3 = le6.inverse_transform(np.array(Y_test,dtype=int))
y_pred = y_pred.reshape(-1,1)
Y_test3 = Y_test3.reshape(-1,1)
df = np.concatenate((Y_test3,y_pred),axis=1)
dataframe = pd.DataFrame(df,columns=['Rain on Tomorrow','Prediction of Rain'])
print(dataframe)
dt_accuracy = accuracy_score(Y_test1,y_pred)
print("\nGradient Boosting Accuracy: "+str(dt_accuracy))
#print(y_pred)
#print(Y_test)

c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\sklearn\ensemble\_gb.py:1454: DataConversionWarning: A
column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using
ravel().
  y = column_or_1d(y, warn=True)

0.878125
Rain on Tomorrow Prediction of Rain
0 No No
1 No No
2 No No
3 No No
4 No No
... ..
795 No No
796 No Yes
797 No No
798 Yes Yes
799 Yes Yes

[800 rows x 2 columns]

Gradient Boosting Accuracy: 0.88
```

Xgboost

```
Y_test4 = le6.inverse_transform(np.array(Y_test,dtype=int))
#print(Y_pred)
#print(Y_test)
y_pred = y_pred.reshape(-1,1)
Y_test4 = Y_test4.reshape(-1,1)
df = np.concatenate((Y_test4,y_pred),axis=1)
dataframe = pd.DataFrame(df,columns=['Rain on Tommorrow','Prediction of Rain'])
print(dataframe)
dt_accuracy = accuracy_score(Y_test1,y_pred)
print("\nXGBoost Accuracy: "+str(dt_accuracy))
```

```
c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\sklearn\preprocessing\_label.py:235: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using.ravel().
  y = column_or_1d(y, warn=True)
c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\sklearn\preprocessing\_label.py:268: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using.ravel().
  y = column_or_1d(y, warn=True)
```

```
1.0
   Rain on Tommorrow Prediction of Rain
0                No                No
1                No                No
2                No                No
3                No                No
4                No                No
..                ...                ...
795              No                No
796              No                No
797              No                No
798              Yes                Yes
799              Yes                No
```

```
[800 rows x 2 columns]
XGBoost Accuracy: 0.875
```

Conclusion

In this paper, we explored and applied several preprocessing steps and learned their impact on the overall performance of our classifiers. We also carried a comparative study of all the classifiers with different input data and observed how the input data can affect the model predictions.

We can conclude that Australian weather is uncertain and there is no such correlation among rainfall and the respective region and time. We figured certain patterns and relationships among data which helped in determining important features. Refer to the appendix section. As we have a huge amount of data, we can apply Deep Learning models such as Multilayer Perceptron, Convolutional Neural Network, and others.

It would be great to perform a comparative study between the Machine learning classifiers and Deep learning models.

References

1. World Health Organization: Climate Change and Human Health: Risks and Responses. World Health Organization, January 2003
2. Alcantara-Ayala, I.: Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. *Geomorphology* 47(24), 107124 (2002)
3. Nicholls, N.: Atmospheric and climatic hazards: Improved monitoring and prediction for disaster mitigation. *Natural Hazards* 23(23), 137155 (2001)
4. [Online] InDataLabs, Exploratory Data Analysis: the Best way to Start a Data Science Project. Available: <https://medium.com/@InDataLabs/why-start-a-data-science-project-with-exploratory-data-analysis-f90c0efcbe49>
5. [Online] Pandas Documentation. Available: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html
6. [Online] Scikit-Learn Documentation Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.FeatureHasher.html
7. [Online] Scikit-Learn Documentation Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>



8. [Online] Scikit Learn Documentation Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
9. [Online] Raheel Shaikh, Feature Selection Techniques in Machine Learning with Python Available: <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>
10. [Online] Imbalanced-learn Documentation Available: <https://imbalanced-learn.readthedocs.io/en/stable/introduction.html>
11. V. Veeralakshmi and D. Ramyachitra, Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset. Issues, vol 1, p. 79-85.
12. [Online] Aditya Mishra, Metrics to Evaluate your Machine Learning Algorithm Available: <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>