Unveiling the Future: Machine Learning in Weather Forecasting

WEATHER FORECASTING SYSTEM
(UML501)
Fifth-Semester

Submitted by:

Pareesh Sharma 102116092

Vimlendu Sharma 102166004

BE Third Year, CSE

Submitted To:

Dr. Archana Singh

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala

November 2023

TABLE OF CONTENTS:

INTRODUCTION:

The Weather Dataset of different cities of Australia from kaggle collaborated by Arunava Kr. Chakraborty (Owner) to test different weather conditions in different cities of Australia and use it to predict whether it will be rain tomorrow or not.
We use this dataset and make a project called Weather Forecasting System.
Our Project delves into the realm of Machine Learning to harness the power of algorithms, data analytics to enhance the precision and reliability of Weather Forecast.
As we navigate through this presentation , we will unravel the intricacies of our machine learning model , its training process and the significant strides we made in achieving a paradigm shift in weather forecasting.
From Predicting rain patterns to anticipating temperature fluctuations, our machine learning system is designed to decipher complex atmospheric data , providing forecast that not only surpass traditional methods in accuracy but also adapt dynamically to evolving weather conditions .
Weather Forecasting plays a crucial role in various sectors , including agriculture , transportation and disaster management. Traditional Methods of weather prediction rely on historical data and meteorological models.
However , the increasing complexity of weather patterns and need for accurate predictions call for advanced approaches.

This Project aims to develop a Weather Forecasting System using Machine Learning techniques to enhance the precision and reliability of weather predictions.

PROBLEM STATEMENT:

In a world where climate patterns are becoming increasingly unpredictable, the need for accurate and timely weather forecasts has never been more critical.

The Conventional method, while commendable often face challenges in keeping pace with rapid atmospheric transformations.

The system should be capable of processing large volumes of diverse data sources, historical weather data and real time sensor readings.

KEY CHALLENGES:

1. **Data Integration:** Integrate and preprocess diverse data sources including historical data and real time sensor readings to create comprehensive data set for training and testing machine learning models.

2. **Feature Selection:** Identify and select relevant features that significantly influence weather patterns to improve the effieciency of machine learning model.
3. **Model Complexity:** Choose and implement appropriate machine learning algorithms that can handle the complexity of weather patterns.
4. **Accuracy Improvement:** Investigate methods to continually improve the accuracy of predictions by refining models with new data.

Key Components:

- **Dataset:** The Weather Dataset (of Australia) comprises of about 10 years of daily weather observations from many locations across Australia. It contains attributes like- Location, Mintemp, Rainfall , RainTomorrow, RainToday etc.

```
  row ID Location  MinTemp  MaxTemp  Rainfall  Evaporation  Sunshine  \
0   Row0   Albury     13.4     22.9       0.6          NaN       NaN
1   Row1   Albury      7.4     25.1       0.0          NaN       NaN
2   Row2   Albury     17.5     32.3       1.0          NaN       NaN
3   Row3   Albury     14.6     29.7       0.2          NaN       NaN
4   Row4   Albury      7.7     26.7       0.0          NaN       NaN

  WindGustDir  WindGustSpeed WindDir9am  ... Humidity9am  Humidity3pm  \
0           W           44.0          W  ...        71.0         22.0
1         WNW           44.0        NNW  ...        44.0         25.0
2           W           41.0        ENE  ...        82.0         33.0
3         WNW           56.0          W  ...        55.0         23.0
4           W           35.0        SSE  ...        48.0         19.0

   Pressure9am  Pressure3pm  Cloud9am  Cloud3pm  Temp9am  Temp3pm  RainToday  \
0       1007.7       1007.1       8.0       NaN     16.9     21.8         No
1       1010.6       1007.8       NaN       NaN     17.2     24.3         No
2       1010.8       1006.0       7.0       8.0     17.8     29.7         No
3       1009.2       1005.4       NaN       NaN     20.6     28.9         No
4       1013.4       1010.1       NaN       NaN     16.3     25.5         No

   RainTomorrow
0             0
1             0
2             0
3             0
```

-

- **Data Preprocessing:** Prior to model training, the weather training dataset undergo preprocessing steps such as normalization, resizing, and noise removal. These steps aim to enhance the quality and uniformity of input data, facilitating effective learning by various classifiers.
- **Classifiers:** The Machine learning and Deep learning classifiers such as : logistic regression, decision trees, random forest, KNN, ANN and LSTM are used to classify the accuracy of the dataset and out which the Random Forest Classifier is the main classifier which classifies our model in the most efficient way.

- **Training and Testing:** The model is trained on the labeled training dataset, optimizing its parameters to classify useful attributes for weather prediction accurately. The test set is used to assess the model's generalization performance and identify potential overfitting.

Expected Outcomes:

- Accurate Prediction: The successful completion of this project will result in a Weather Forecasting that leverages Machine Learning and Deep Learning Techniques to provide accurate and timely weather predictions.
- Robustness to Noise: The model should demonstrate robustness to variations and noise within the dataset, considering the potential challenges associated with low-quality data.

DATASET:

The success of any machine learning project, particularly one involving through various Machine Learning Algorithms, depends on the quality and diversity of the dataset used for training and evaluation. In this project, a meticulously curated dataset of weather observations will be employed to develop and validate the Predicting model for Weather Forecasting.

The dataset has been taken from Kaggle. It comprises a diverse collection of about 10 years of daily weather observations from different locations of Australia. Observations are drawn from numerous weather stations.

The daily observations are available from-

Link:

http://www.bom.gov.au/climate/data

Definitions adapted from:

Link:

http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

Data Source:

Link:

http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data

Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

METHODOLOGY:

1. Installing of necessary libraries such as numpy, pandas and keras etc.
2. Loading of Weather Dataset.
3. And then the process of Preprocessing starts.

- **PROPROCESSING ON DATA:**

1. Get dimensions of dataset.
2. Removal of null values in dataset by Dropna inbuilt function.
3. Removal of outliers by use of z-score.
4. Conversion of categorical data into numerical values by use of One Hot Encoding
5. ONE HOT ENCODING: Technique of Machine Learning use to encode categorical values into binary values.The term one-hot comes from the fact that only element in data is 'hot' and set to 1 while other elements except it are set to 0.
6. Standardisation of data by Min_Max_Scaler which helps to normalise the data values in range among [-1,0,1].

7. **FEATURE SELECTION IN DATA:**

8. Selection of most efficient variables from dataset that will surely affect the output variable – "RainTomorrow".

9. **DATA MODELLING :**

    1. **BY Logistic Regression:** It uses a threshold term to divide the data in training and testing phase.
    2. **BY RANDOM CLASSIFIER:** It is best known classifier use for classification and prediction of data .
    3. **BY DECISION TREE CLASSIFIER:** It is used to classify the data to some good extent.

4. **SUPPORT VECTOR MACHINE:** It is supervised machine learning algorithm that can be used for classification and regression tasks.
5. **BY DEEP LEARNING :**
6. **LSTM**
7. **ARTIFICIAL NEURAL NETWORK**

JUPYTER NOTEBOOK (CODE) SNAP SHOTS:

```
In [3]: !pip install numpy
        !pip install pandas

Requirement already satisfied: numpy in c:\users\dell\anaconda3\lib\site-packages (1.23.5)
Requirement already satisfied: pandas in c:\users\dell\anaconda3\lib\site-packages (1.5.3)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\dell\anaconda3\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\dell\anaconda3\lib\site-packages (from pandas) (2022.7)
Requirement already satisfied: numpy>=1.21.0 in c:\users\dell\anaconda3\lib\site-packages (from pandas) (1.23.5)
Requirement already satisfied: six>=1.5 in c:\users\dell\anaconda3\lib\site-packages (from python-dateutil>=2.8.1->pandas) (1.1
6.0)

In [4]: import pandas as pd #data preprocessing,csv file i/0
        import numpy as np#linear algebra

In [5]: df=pd.read_csv('Weather Training Data.csv')

In [6]: print('Size of weather data frame is:',df.shape)

Size of weather data frame is: (99516, 23)
```

```
In [7]: print(df[0:5])

  row ID Location  MinTemp  MaxTemp  Rainfall  Evaporation  Sunshine  \
0   Row0   Albury     13.4     22.9       0.6          NaN       NaN
1   Row1   Albury      7.4     25.1       0.0          NaN       NaN
2   Row2   Albury     17.5     32.3       1.0          NaN       NaN
3   Row3   Albury     14.6     29.7       0.2          NaN       NaN
4   Row4   Albury      7.7     26.7       0.0          NaN       NaN

  WindGustDir  WindGustSpeed WindDir9am  ... Humidity9am  Humidity3pm  \
0           W           44.0          W  ...        71.0         22.0
1         WNW           44.0        NNW  ...        44.0         25.0
2           W           41.0        ENE  ...        82.0         33.0
3         WNW           56.0          W  ...        55.0         23.0
4           W           35.0        SSE  ...        48.0         19.0

   Pressure9am  Pressure3pm  Cloud9am  Cloud3pm  Temp9am  Temp3pm  RainToday  \
0       1007.7       1007.1       8.0       NaN     16.9     21.8         No
1       1010.6       1007.8       NaN       NaN     17.2     24.3         No
2       1010.8       1006.0       7.0       8.0     17.8     29.7         No
3       1009.2       1005.4       NaN       NaN     20.6     28.9         No
4       1013.4       1010.1       NaN       NaN     16.3     25.5         No
```

In [8]:
```
#checking null values
#data_preprocessing
print(df.count().sort_values())
```

```
Sunshine         52199
Evaporation      56985
Cloud3pm         59514
Cloud9am         61944
Pressure9am      89768
Pressure3pm      89780
WindDir9am       92510
WindGustDir      92995
WindGustSpeed    93036
WindDir3pm       96868
Humidity3pm      97010
Temp3pm          97612
WindSpeed3pm     97681
Humidity9am      98283
Rainfall         98537
RainToday        98537
WindSpeed9am     98581
Temp9am          98902
MinTemp          99073
MaxTemp          99286
row ID           99516
Location         99516
RainTomorrow     99516
dtype: int64
```

In [9]:
```
#removing_unwanted_variables
df=df.drop(columns=['Sunshine','Evaporation','Cloud3pm','Cloud9am','Location'],axis=1)
print(df.shape)
```


```
RainTomorrow     99516
dtype: int64
```

In [9]:
```
#removing_unwanted_variables
df=df.drop(columns=['Sunshine','Evaporation','Cloud3pm','Cloud9am','Location'],axis=1)
print(df.shape)
```


```
(99516, 18)
```

In [10]:
```
#get_rid_of_null_values
df=df.dropna(how='any')
print(df.shape)
```

```
(79140, 18)
```

In [11]:
```
#remove_outliers_in_data_using_z_score
from scipy import stats
z=np.abs(stats.zscore(df._get_numeric_data()))
print(z)
df=df[(z<3).all(axis=1)]
print(df.shape)
```

```
        MinTemp   MaxTemp  Rainfall  WindGustSpeed  WindSpeed9am  \
0      0.119802  0.105934  0.207977       0.240913      0.576967
1      0.842097  0.209274  0.278039       0.240913      1.339686
2      0.777100  1.240864  0.161269       0.015814      0.980314
3      0.312182  0.868346  0.254685       1.141306      0.457176
4      0.794002  0.438516  0.278039       0.434382      1.100105
...         ...       ...       ...            ...           ...
99511  0.745907  0.421143  0.278039       0.015814      0.457176
99512  1.467331  0.263539  0.278039       0.734513      0.021987
99513  1.579553  0.034296  0.278039       0.734513      0.261569
99514  1.451299  0.237929  0.278039       1.409809      0.261569
99515  1.162730  0.467171  0.278039       0.284317      0.740732

       WindSpeed3pm  Humidity9am  Humidity3pm  Pressure9am  Pressure3pm  \
0          0.523188     0.188186     1.381221     1.385198     1.143726
1          0.290050     1.237542     1.236549     0.973301     1.043301
2          0.056912     0.769038     0.850755     0.944895     1.301538
3          0.523188     0.656690     1.332997     1.172148     1.387617
4          0.292795     1.026323     1.525894     0.575608     0.713331
...             ...          ...          ...          ...          ...
99511      0.756326     0.603885     0.898979     1.512280     1.323876
99512      0.759071     0.445471     1.140100     1.029367     0.879133
99513      0.992209     0.867909     1.284773     1.015164     0.750015
```

In [12]:

```python
#for-categorical_columns_change_yes_no_to_1/0_for_rain_today_and_rain_tommorrow
df['RainToday'].replace({'No':0,'Yes':1},inplace=True)
df['RainTomorrow'].replace({'No':0,'Yes':1},inplace=True)
print(df.shape)
print(df)
```

```
(75601, 18)
          row ID  MinTemp  MaxTemp  Rainfall WindGustDir  WindGustSpeed  \
0           Row0     13.4     22.9       0.6           W           44.0
1           Row1      7.4     25.1       0.0         WNW           44.0
2           Row2     17.5     32.3       1.0           W           41.0
3           Row3     14.6     29.7       0.2         WNW           56.0
4           Row4      7.7     26.7       0.0           W           35.0
...          ...      ...      ...       ...         ...            ...
99511  Row101816      8.0     20.7       0.0         ESE           41.0
99512  Row101817      3.5     21.8       0.0           E           31.0
99513  Row101818      2.8     23.4       0.0           E           31.0
99514  Row101819      3.6     25.3       0.0         NNW           22.0
99515  Row101820      5.4     26.9       0.0           N           37.0

      WindDir9am WindDir3pm  WindSpeed9am  WindSpeed3pm  Humidity9am  \
0              W        WNW          20.0          24.0         71.0
1            NNW        WSW           4.0          22.0         44.0
2            ENE         NW           7.0          20.0         82.0
3              W          W          19.0          24.0         55.0
4            SSE          W           6.0          17.0         48.0
...          ...        ...           ...           ...          ...
99511         SE          E          19.0          26.0         56.0
99512        ESE          E          15.0          13.0         59.0
99513         SE        ENE          13.0          11.0         51.0
99514         SE          N          13.0           9.0         56.0
99515         SE        WNW           9.0           9.0         53.0

       Humidity3pm  Pressure9am  Pressure3pm  Temp9am  Temp3pm  RainToday  \
0             22.0       1007.7       1007.1     16.9     21.8          0
1             25.0       1010.6       1007.8     17.2     24.3          0
2             33.0       1010.8       1006.0     17.8     29.7          0
3             23.0       1009.2       1005.4     20.6     28.9          0
4             19.0       1013.4       1010.1     16.3     25.5          0
...            ...          ...          ...      ...      ...        ...
99511         32.0       1028.1       1024.3     11.6     20.0          0
99512         27.0       1024.7       1021.2      9.4     20.9          0
99513         24.0       1024.6       1020.3     10.1     22.4          0
99514         21.0       1023.5       1019.1     10.9     24.5          0
```

In [15]:

```python
#Preprocessing_is_complete
#Expolatory_data_analysis
#feature_selection
#selectKBest_function is used to select some selective variables
from sklearn.feature_selection import SelectKBest ,chi2
```

In [16]:

```python
#it will select the most significant predictor variable
x=df.loc[:,df.columns!='RainTomorrow']
y=df[['RainTomorrow']]
selector=SelectKBest(chi2,k=3)
selector.fit(x,y)
x_new=selector.transform(x)
print(x.columns[selector.get_support(indices=True)])#top 3 columns
```

```
Index(['Rainfall', 'Humidity3pm', 'RainToday'], dtype='object')
```

In [17]:

```python
#get_hold_of_important_features_and_assign_them_as_x
df=df[['Humidity3pm','Rainfall','RainToday','RainTomorrow']]
x=df[['Humidity3pm']]#Let's_use_only_one_feature
y=df[['RainTomorrow']]
```

In [18]:

```python
#data_modelling
#use_classification_logisitic_regression
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
import time
```

In [19]:

```python
#calculate_accuracy_and_time_taken_by_classifier
t0=time.time()
```

In [20]:

```python
#data_splicing-splitting_data_in_testing_and_training_data
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25)#testing-data=25%and_remaining_training_data=75%
clf_logreg=LogisticRegression(random_state=0)#creation_of_instance_for_logistic_regression
#fit/build the model_using_training_dataset
clf_logreg.fit(x_train,y_train)
```

In [37]:
```
model = Sequential()
```

```
WARNING:tensorflow:From C:\Users\DeLL\anaconda3\Lib\site-packages\keras\src\backend.py:873: The name tf.get_default_graph is de
precated. Please use tf.compat.v1.get_default_graph instead.
```

In [38]:
```
import numpy as np
import matplotlib.pyplot as plt
```

In [39]:
```
score = np.float64(0.85)

# Check the type of the variable
if isinstance(score, dict):
    names = list(score.keys())
    values = list(score.values())

    plt.figure(figsize=(10, 5))
    plt.bar(names, values)
    plt.xlabel('Classifiers')
    plt.ylabel('Accuracy Score')
    plt.title('Accuracy Score of Different Classifiers')
    plt.ylim(0, 1.0)  # Set y-axis limit to 0-1 for accuracy scores
    plt.show()
else:
    print("The variable is not a dictionary.")
```

```
The variable is not a dictionary.
```

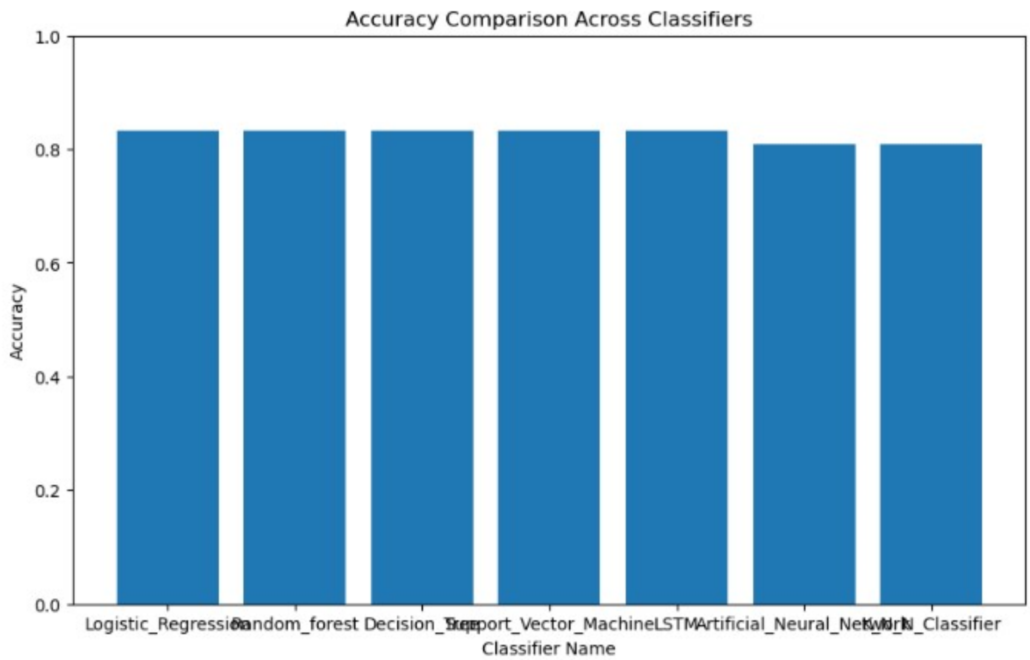
In [40]:
```
# Deep Learning Model (LSTM)
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(1, 1)))  # Adjust input shape based on your data
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')
```

```
WARNING:tensorflow:From C:\Users\DeLL\anaconda3\Lib\site-packages\keras\src\optimizers\__init__.py:309: The name tf.train.Optim
izer is deprecated. Please use tf.compat.v1.train.Optimizer instead.
```

CLASSIFIERS ACCURACY TABLE:

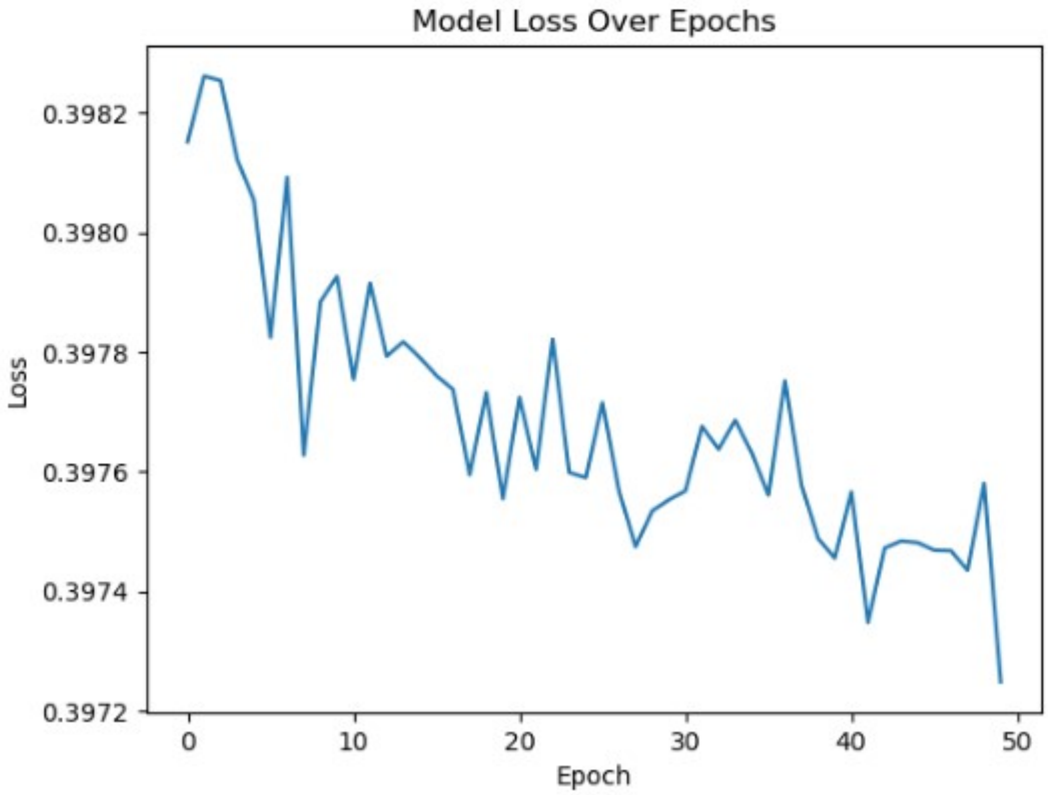
	Classifier_Name	Accuracy
1	Logistic_Regression	[0.8331305221945928]
2	Random_forest	[0.8328659859266705]
3	Decision_Tree	[0.8326014496587482]
4	Support_Vector_Machine	[0.8335008729696841]
5	LSTM	[0.8323369133908258]
6	Artificial_Neural_Network	[0.8091106290672451]
7	K_N_N_Classifier	[0.8091106290672451]

FOR ACCURACY PLOT:



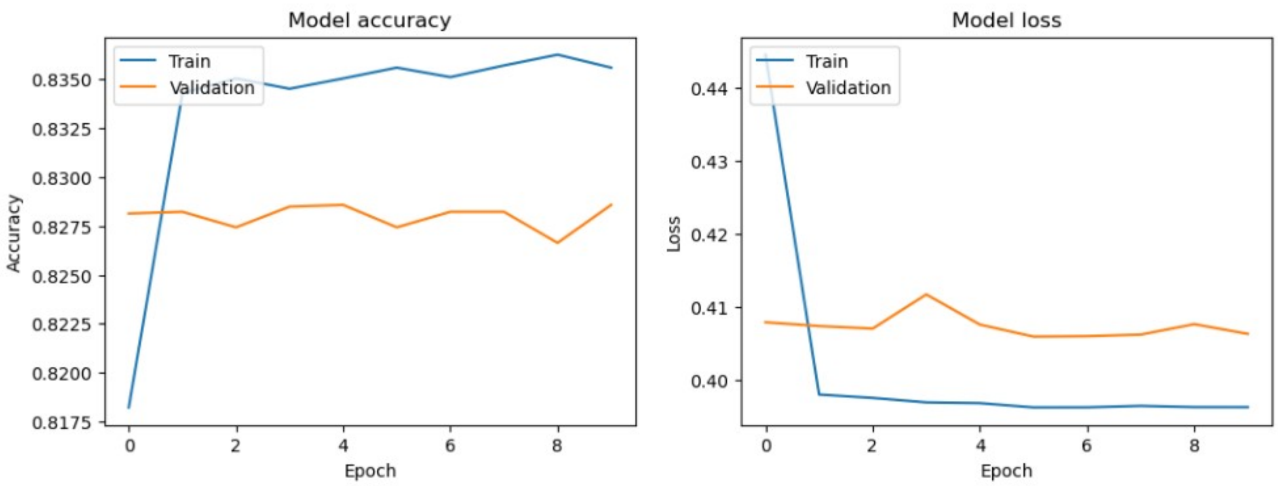
FOR DEEP LEARNING:

BY LSTM:



FOR ADVANCED CLASSIFICATION:

ARTIFICIAL NEURAL NETWORK (ANN):



CONCLUSION:

In Conclusion, the integration of Machine Learning and Deep Learning into Weather Forecasting models represents a significant leap forward in our ability to predict and understand complex atmospheric phenomenon.

Through the utilisation of advanced algorithms, big data analytics and innovative technologies we are poised the accuracy and reliability of weather forecast enabling better preparedness for extreme weather events and improving overall societal resilience.

As we have explored through this report, machine and deep learning models bring about a paradigm shift in extracting patterns and insights from vast datasets that traditional methods may struggle to process effectively. The continuous learning capability of these models enable them to adapt to evolving weather conditions providing forecasts that are not only precise but also more adaptable to dynamic nature of weather system.

The intersection of machine learning and weather forecasting holds immense promise for revolutionising our understanding of atmosphere and improving our ability to anticipate and respond to weather-related events. Embracing these Technology advancements is not just a scientific imperative but a practical necessity for building a more resilient and adaptive society in face of an ever-changing environment.

REFERENCES:

1. http://www.bom.gov.au/climate/data

2. http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

3. http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data

4. Australia Weather Data (kaggle.com)

5. Introduction to Machine Learning Algorithms: Linear Regression | by Rohith Gandhi | Towards Data Science

6. Textbooks:

7. 1. Data Mining: The Textbook 2015 Edition, Kindle Editionby Charu C. Aggarwal .

2. Data Mining: Concepts and TechniquesBy Jiawei Han, Jian Pei, Micheline Kamber

GITHUB REPOSITORY LINK:

GitHub - Pareesh-Sharma/ML_Project