

**WiDS ‘22 - ‘23 Final Documentation**

**Daily weather prediction using machine learning and deep learning (UID-44)**

**Mentors:** Annie D’Souza, Amruta Parulekar

|  |  |  |
| --- | --- | --- |
| **Team Member Name** | **Roll Number** | **Email-Id** |
| Nitish Beck Bhagat | 22M1853 | 22m1853@iitb.ac.in |
|  |  |  |

**Introduction to Problem Statement**

|  |
| --- |
| Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. Ancient weather forecasting methods usually relied on observed patterns of events, also termed pattern recognition. However, not all of these predictions prove reliable. Now, weather can be predicted by various parameters such as temperature, humidity, wind speed, precipitation, pressure which are more reliable when combined with machine learning models. |

**Existing Resources**

|  |
| --- |
| Remote sensing technology, weather radar which could detect direction and movement of rain, meteorological satellites which capture the distribution of clouds across earth from space are indispensable methods for weather observation. With weather observational data, weather maps can be produced to track the movement of weather systems and to predict weather. From consecutive weather maps, one can track the direction and speed of movement of weather systems such as high-pressure areas, low-pressure areas, cold fronts, and warm fronts, so as to assess the local weather changes brought by these weather systems. Nowadays we rely on computer models to make weather forecasts. The technology concerned is called numerical weather prediction. In recent years, many researches have been made in the application of artificial intelligence in weather forecasting, especially the use of deep learning methods. |

**Proposed Solution**

|  |
| --- |
| The dependency of weather on temperature, pressure, precipitation, wind speed, humidity can be used to make machine learning model which will predict the weather by studying the past data and find correlation between different parameters. The model can be further trained and its accuracy can be increased by initial data preprocessing and feature engineering which will provide us a robust prediction. |

**Methodology & Progress (Mention the work done week-wise)**

|  |
| --- |
| **Methodology:**   * The dataset provided by the analytics club is imported along with other python libraries. Brief description of dataset is shown. * Following **data preprocessing** is done for the unstructured data which include finding imputing the missing values with mean, mode depending upon the number of missing values and checking for any duplicate values. If there is any unnecessary data, it is dropped. * Once data is cleaned, we plot a correlation plot as heatmap to find any correlation between dependent variables. Later, if correlation above a threshold value, we drop the feature or reduce the multicollinearity by using **Boxcox, log transformation**. Also, to check **multicollinearity** we are using heatmap and **VIF (Variance Inflation Factor)** score. * **Exploratory Data Analysis** is done such as count plot, scatterplot, joint plot, kde plot (to check the skewness of the feature values) for better understanding and extract information from the plots. * **Feature engineering** is carried out to create dummy variables (**one hot encoding**) for categorical features. Also, new features like month & year are created from the formatted date feature (string type). * Now, we can apply various ML models to our dataset.   **Linear Regression:**   * **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Here, the target variable is humidity. * Data scaling is not done for this model. * Model is created and x, y values are trained and values are predicted. * Here, we are using **R2** as an evaluation score.   **SVM ( Support Vector Machine):**   * Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks. But, it is widely used in classification objectives. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. * Dataset is scaled using standard scaler. * Model is trained & values are predicted. * Here, we are using **R2** as an evaluation score.   **Random Forest:**   * Random forest is an ensemble technique capable of both classification & regression task using multiple decision tree and a technique called **bootstrap** & **aggregation**, commonly known as **bagging.** The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. * We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called **Bootstrap**. * In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called **Aggregation**. * Model is created and x, y values are trained and values are predicted. * Here, we are using **R2** as an evaluation score.   **Artificial Neural Network:**   * Keras is a simple tool for constructing a neural network. It is a high-level framework based on tensorflow, * We standardize our dataset using standard scaler. * **Sequential** specifies to keras that we are creating model sequentially and the output of each layer we add is input to the next layer we specify. * **model. add** is used to add a layer to our neural network. We need to specify as an argument what type of layer we want. The **Dense** is used to specify the fully connected layer. * In our neural network, we are using different combinations of hidden layers and using the one with best score. * Now we need to specify the loss function and the optimizer. It is done using compile function in keras. Since it is a regression problem, loss function used is **mse** and optimizer is **adam** * Training step is simple in keras. model.fit is used to train it. * Now we can check the model’s performance on test data: * Here, we are using R2 as an evaluation score. |
| Week 1: Introduction to Data Science, machine learning  Week 2: Exploratory Data Analysis  Week 3: Linear regression, SVM  Week 4 :Random Forest, Introduction to Deep Learning  Week 5: Artificial Neural Network |

**Results**

|  |
| --- |
| Different model were applied to the dataset and following R2 score is obtained:    link to drive folder/ GitHub page: <https://github.com/NitishBB/WIDS-weather-prediction> |

**Learning Value**

|  |
| --- |
| Basics concepts of machine learning and deep learning and their implementations, parameters ,hyper parameter tuning, standardizing ,dealing with data, exploratory data analysis were learned from this project. |

**Tech-stack Used**

|  |
| --- |
| Python,numpy,pandas, seaborn, matplotlib, tensorflow |

**Suggestions for others**

|  |
| --- |
| This project helped me a lot on how to apply machine learning and deep learning on real life datasets and constant mentorship helped me to finish this project. |

**Contribution by each Team Member**

|  |
| --- |
| Individual Project |

**References and Citations**

|  |
| --- |
| <https://towardsdatascience.com/exploratory-data-analysis-in-python-a-step-by-step-process-d0dfa6bf94ee>  <https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2>  <https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/>  <https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/>  <https://www.analyticsvidhya.com/blog/2022/03/an-overview-on-long-short-term-memory-lstm/> |