
# Fundamentals Of Big Data Analytics

# WeatherSense Project


| | |
|---|---|
| Muhammad Bilal | L22-7551 |
| Umair Imran | L22-8370 |
| Akhyar Chaudhary | L22-7461 |


## Title And Description

- **Title:** WeatherSense: Comprehensive Weather Analysis and Forecasting
- **Description:** Using the power of Pyspark to perform in-depth analysis and make robust predictions on comprehensive and detailed weather data.

# Abstract

The main motive behind this research paper is to provide a comprehensive understanding about the nature of the project, the reason to opt it, what is its significance and what are the technical aspects of it which make the topic opted worth giving time to. Choosing Weather Data to analyze it and preform cleaning, exploratory data analysis, visualizing the data and making predictions using advanced machine learning models provides a solid foundation to handle big data in practical terms.


# Introduction

- **Background:** We were given a big data project to choose a topic, gain data and perform big data techniques to gain insights from the large dataset. Our group decided to take a practical topic related to meteorology. Weather and climate have always been a fascinating thing for me and to use the weather data to apply big data techniques made it even more interesting. However, in the field of data science, we must always have a problem to deal with in the context of data and we should know how to build our hypothesis around it. So, we took the weather data of US and performed all the tasks related to project.
- **Problem Statement:** How to handle big weather data and apply machine learning on it to gain useful predictions?
- **Objectives:** To determine weather conditions in the US based on the data lake (1991-2021) including precipitation,

Maximum temperature, minimum temperature and possibility of climate change.

H0: Change in Precipitation

H1: No change in precipitation

H0: Climate Change

H1: No climate Change

H0: Possible increase in Maximum Temperature

H1: No increase in Maximum Temperature

H0: Possible decrease in Minimum Temperature

H1: No decrease in Minimum Temperature

- **Significance:** Weather and Climate directly affect us and our surroundings. The eco-pool comprised of all the living beings is affected by it. Hence, it's significance cannot be underestimated.

## Literature Review

- **Prior Work:** We already have worked on similar datasets and have hands-on experience to handle datasets like this one. The ability to work with dependent features to determine response variables is an approach used for

numerical, categorical datasets, however, the difference is between the choice of the machine learning model. In this case, we will use Multi-Linear Regression

- **Theoretical Frameworks:** Heat transfer concepts, Stefan-Boltzmann law of radiation, changes in temperature, precipitation and evaporation

## Methodology

- **Data Description:** The name of the dataset is US weather data (1991-2021) providing details about the weather conditions in the US in a generalized way. We found this dataset on Kaggle. It was roughly around 8gb and had around 150 million rows. It has 9 features (ID, Date, Maximum Temperature, Minimum Temperature, Precipitation, Evaporation, Elevation, Longitude, Latitude).
- **Data Preprocessing:** Firstly, we removed the duplicate and redundant rows. Secondly, we removed the nulls from the rows. Thirdly, we founded the outliers but did not drop them as they were essential in finding climate change and weather anomalies. Furthermore, we did exploratory data analysis (EDA) to visualize the features, drew a heatmap to check for correlation and multi-collinearity.
- **Tools and Techniques:** We have used Spark's API for python, called Pyspark. We have used Pysaprk's built-in features for data cleaning. We have used matplotlib and seaborn for exploratory data analysis and visualizations.

- **Analytical Methods:** Pyspark provides a machine learning library to carry out regression and classification tasks. Here, we specifically used the multi-linear regression approach as the dataset was numerical and linear. We first assembled the features to create a synchronous dimension using Vector Assembler. Then, we split the assembled data into training and testing portions. Afterwards, we applied multi-linear regression using pyspark's built-in machine learning library. We predicted Precipitation, Maximum Temperature and Minimum Temperature from it.

## Results

- **Model Evaluation:** For precipitation, the RMSE value was 64.643820 and the R^2 value was 0.560100. For Maximum Temperature, the RMSE value was 44.0829 and R^2 value was 0.6717380. For Minimum Temperature, the RMSE value was 41.310504 and R^2 value was 0.7039022.
- **Interpretation:** The predictions tell us that the precipitation for the upcoming years in US will relatively be normal. The minimum temperature coefficients tell us that it will continue to be stable according to the past data. The maximum temperature coefficients tell us that the temperature is being increased as each year progresses, but the change is negligible. A robust criterion to find out if there is a possible climate change is to determine if the temperature is increased by 0.0015 degrees centigrade each year. However, based on our dataset, the change is even less than the stated threshold value.

# Citations


- Adams, R.M., K.J. Bryant, and R. Weiher, 1995. Value of improved long-range weather information. Contemporary Economic Policy 13: 10-19.
- Allen 1990 Global climate change and U. S. agriculture. Nature 345:219-224.