**PES University, Bangalore**

(Established under Karnataka Act No. 16 of 2013)  
 **UE19CS203 – STATISTICS FOR DATA SCIENCE**

**Project details**

Project Title :

Student details :

|  |  |  |  |
| --- | --- | --- | --- |
| Sl. No. | Name | SRN | Section |
| 1. | Shamanth K M | PES1UG19CS444 | G |
| 2. | Sanjiv Sridhar | PES1UG19CS432 | G |
| 3. | Sahana B Manjunath | PES1UG19CS411 | G |
| 4. | Sayonika Das | PES1UG19CS439 | G |

**Abstract**

The dataset under consideration is Weather Dataset – Temperature Prediction (Weather in szeged : 2012 - 2015). This project is about combining programming skills, knowledge of mathematics and statistics to extract meaningful insights out of the raw data. Here, the data regarding weather is cleaned i.e., missing data in all kind of data present in the dataset are handled. Unimportant features are dropped, some are imputed and duplicated or irrelevant records are removed. The dataset is visualized to exhibit the insights from the data using different graphs, which is useful in removing outliers if any. The data is then normalized and standardized as per requirements, checked for gaussian distribution using plots and statistical tests (i.e., hypothesis testing : Null hypothesis is supported or rejected). The variables in the data are now checked for mutual relationship (correlation) and draw inferences from it.

**Problem Statement (**Introduction: Introduction about the problem statement that you have considered to study (write two paragraphs).**)**

The Weather dataset – Temperature Prediction (Weather in szeged : 2012 - 2015) has records of weather on an hourly basis of 4 years. This dataset is and raw and unclean. It needs to be cleaned i.e., stripped of missing values or say NaN values, duplicate or irrelevant records, typos and inconsistent capitalization to be either corrected or removed as per requirement. The dataset now, is ready for further operations. The data is now visualized to observe the insights, which is essential to detect the presence of outliers and therefore process the data inorder to remove the outliers if any present in the dataset.

The data is now to be normalized and standardized. The mean of the columns under importance and consideration for the analysis are asserted to values in the range of 0 – 1. The data now is visualized to check for gaussian distribution. It is then followed by statistical tests (hypothesis testing). The Null hypothesis is either supported or rejected. Lastly, the data variables are checked mutual relationship (correlation). This depicts how the variables are related to different other important variables.

**Dataset**

The dataset Weather – Temperature Prediction (Weather in szeged : 2012 - 2015) is sourced from a popularly known website : Kaggle.com. This dataset is about weather on an hourly basis for 4 years ranging from 2012 – 2015. It has 12 features : Formatted Date, Summary, Precipitation Type, Temperature (C), Apparent Temperature ©, Humidity, Wind Speed (km/h), Wind Bearing (degrees), Visibility (km), Loud Cover, Pressure (millibars), Daily Summary. It consists of 35,078 records that include true and real values that are relevant and also errored vales and missing values. The dataset is unrefined and may contain features that might not be useful at the end.

**Preprocessing or Data Cleaning** (Write the techniques that are used by you to clean your dataset. Also, mention about why cleaning is required with respect to your dataset (one paragraph). **It may include outlier removal if any.** )

This is part where the data is essentially cleaned. All the rows where the Precip Type and Summary is NaN have been removed by selecting the other rows with notna function because we cannot find the mean of categorical data. Then the "Loud Cover" column has been dropped using the drop function as it contained only one value that is 0 and it is absolutely not useful at all for prediction. Preprocessing from sklearn has been imported and the LabelEncoder function is used to encode the categorical values to numerical values as it makes it easy to use them for prediction. After that the values have been imputed in the columns containing NaN values using mean as the strategy so that we can use considerable values for training using the SimpleImputer function that has imported from sklearn. All the rows that are duplicate of other rows have been removed so that the model does not train on the same values again. This concludes the process that has been performed and the data is now ready to be visualized and processed for further insights.

**Exploratory Data Analysis (**What are the insights obtained from EDA. Represent them in terms of various graphs along with explanation. (two to three paragraphs). **)**

**Hypothesis Testing**  (Come out with appropriate hypotheses w.r.to a population parameter of your dataset. Perform Hypothesis Testing and make conclusion(one paragraph). )

**Results** (Results and Discussion of your problem statement(one paragraph).)