Machine Learning Project Mona Shah – N01479948

* A project work under Mihai Albu Dec 16th, 2022

## Introduction

The Project demonstrates observations obtained from both the Supervised and Unsupervised Learning approaches and predicts the target variable ‘Climate’ for the dataset: The Shifts Weather Prediction. The dataset consists of 50K instances and around 129 features that help predict whether the climate is tropical, mild temperature or dry. Link of the dataset is mentioned under the References.

## Methods

As mentioned in the introduction we will be discussing the Supervised and Unsupervised Learning approaches. The algorithms used for both approaches are as follows:

* Naïve Bayes

It is a supervised learning algorithm based on Bayes' theorem and used to solve classification problems.

It is called naive because it assumes that the occurrence of certain traits is independent of the occurrence of other traits. For example, if you identify fruits by color, shape, and taste, red, globular, sweet fruits are perceived as apples. Therefore, each trait contributes individually to making apples without interdependence.

Bayes' Theorem, also known as Bayes' Rule or Bayes' Law, is used to determine the probability of a hypothesis based on prior knowledge. It depends on conditional probabilities.

Naïve Bayes Classifier Algorithm

* K-Means Clustering Algorithm

K-Means clustering is an unsupervised learning algorithm used in machine learning or data science to solve clustering problems.

Diagram

Description automatically generated

* Logistic Regression Algorithm

Logistic regression categorically predicts income as the dependent variable. Therefore, the result must be a categorical or discrete value. It can be either yes or no, 0 or 1, true or false, etc., but instead of an exact value of 0 and 1, it returns probable values ​​that fall between 0 and 1.

Diagram

Description automatically generated

## Results

Following are the results of the above mentioned methods for the weather dataset. We are using Jupyter Notebook for calculating the above algorithms.

* Naïve Bayes

Loading the Dataset

Graphical user interface, text, application, email

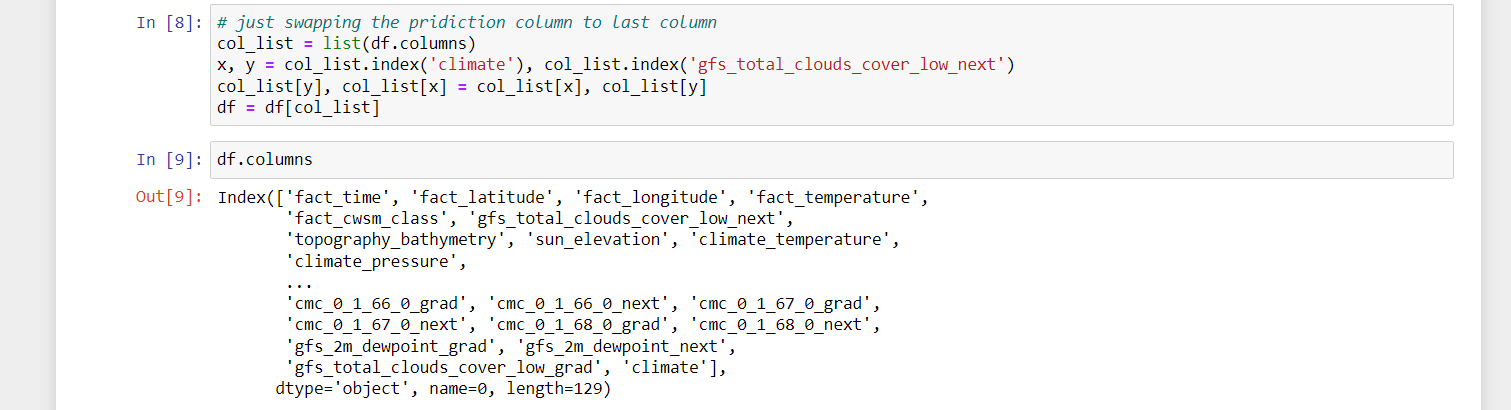
Description automatically generated

Removing the header row

Graphical user interface, text, application

Description automatically generated

Putting the target column at the last place

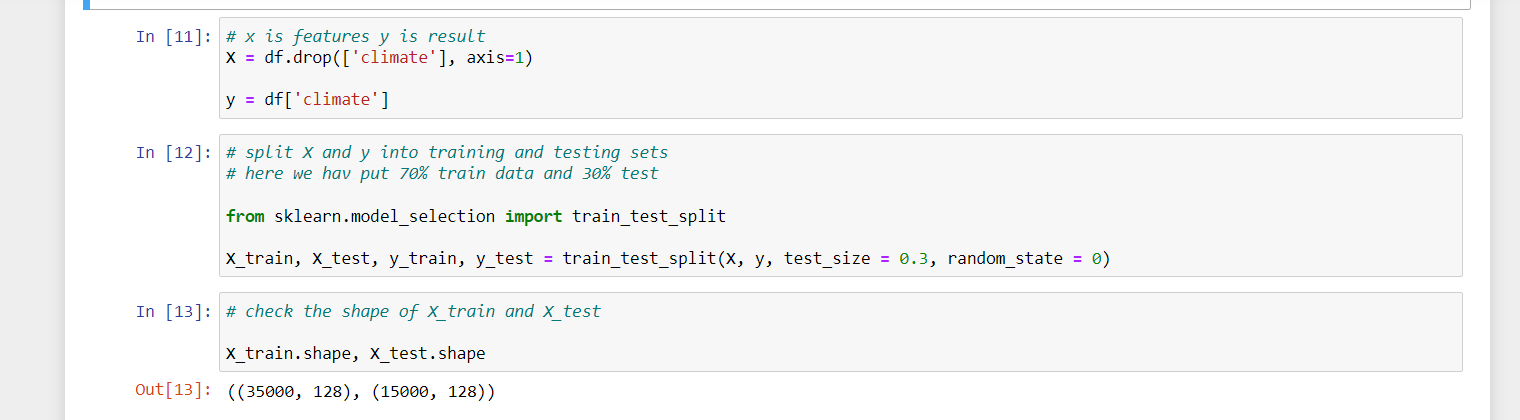


Dataset Summary

Background pattern

Description automatically generated with low confidence

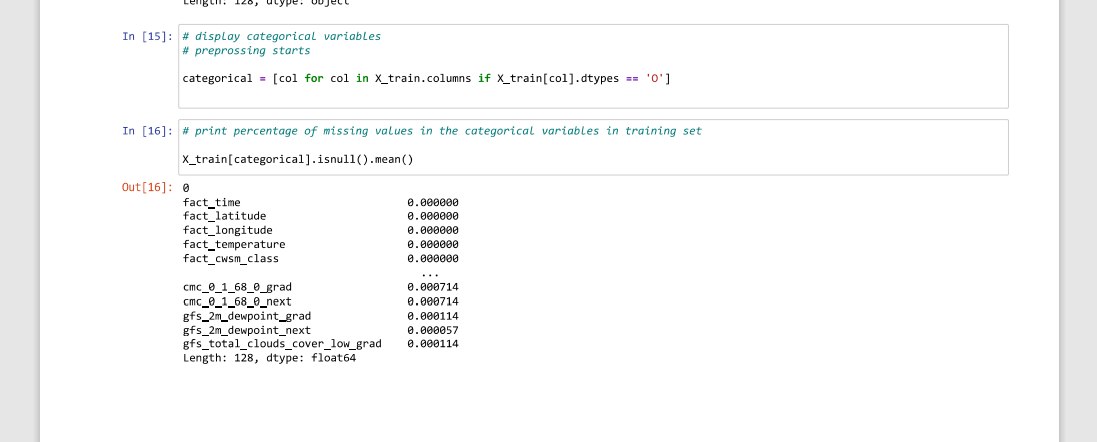
Dividing the dataset into training and testing sets



Graphical user interface, application

Description automatically generated

Preprocessing the dataset to replace null values with the most frequent ones



Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Table

Description automatically generated

Performing Naïve Bayes Algorithm

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

* K-Means Clustering Algorithm

Loading the Dataset

Graphical user interface, text, application, email

Description automatically generated

Removing the header row, putting the target column at the last place and dataset Summary

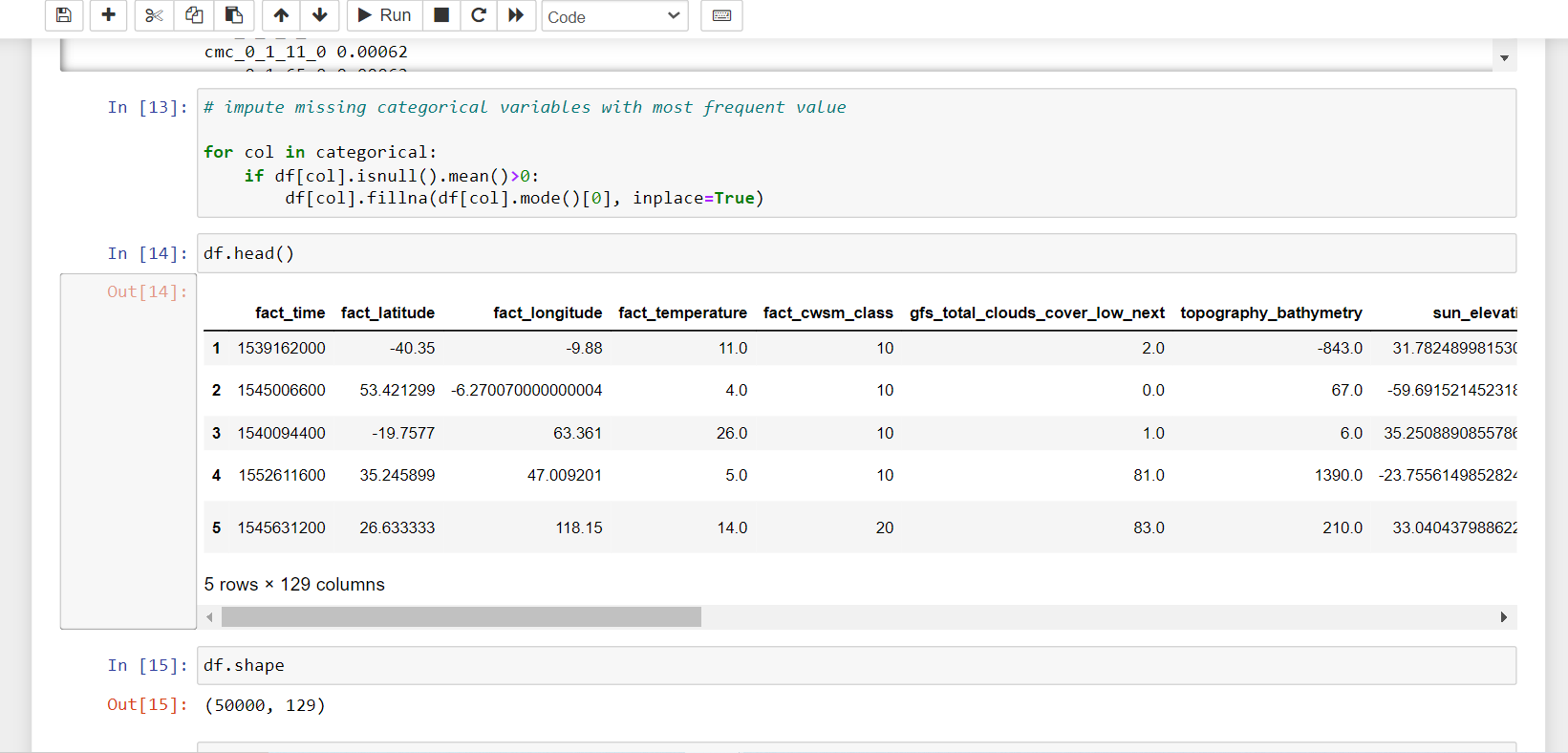
Graphical user interface, text, application, email

Description automatically generated

Preprocessing the dataset to replace null values with the most frequent ones

Graphical user interface, text, email

Description automatically generated



Performing K-Means Clustering Algorithm (Clusters = 3)

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Performing K-Means Clustering Algorithm (Clusters = 21)

Graphical user interface, text, application, email

Description automatically generated

Chart, bar chart

Description automatically generated

* Logistic Regression Algorithm

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

## Discussion And Observations

From the above results we have the following Observations that can be discussed below:

* The dataset needs to be preprocessed for all the algorithms. In other words, replacing the null values with the most frequent ones.
* After performing Naïve Bayes, the accuracy was as follows:

Training set accuracy: 69.4286

Test set accuracy: 69.8133

* After performing K Means with 3 components, the accuracy was as follows:

K=3 KMeans -> 38.0560%

* After performing K Means with 21 components, the accuracy was as follows:  
  K=21 KMeans -> 68.9000%
* After performing Logistic Regression, the accuracy was as follows:

Logistic Regression Accuracy - 80.8560%

Thus, we can conclude that Logistic Regression suites best for the Weather Dataset.

## Feature Reduction

* PCA

To reduce dimensionality in machine learning, principal component analysis is an unsupervised learning approach. It is a statistical procedure that uses orthogonal transformation to change the observations of correlated features into a set of linearly uncorrelated features. The Principal Components are these newly altered features. One of the widely used tools for exploratory data analysis and predictive modelling is this one. It is a method for identifying significant patterns in the provided dataset by lowering the variances.

Some of the terminology of PCA is as follows:

**Dimensionality** is the quantity of characteristics or variables in the dataset in question. The dataset's amount of columns makes this easier to determine.

The **correlation** between two variables indicates how closely they are related to one another. For instance, if one variable changes, the other variable also changes. The correlation score lies between -1 and +1. Here, -1 denotes an inverse relationship between the variables, and +1 denotes a direct relationship between the variables.

**Orthogonal**: It establishes that there is no correlation between the variables, and as a result, there is none.

If a non-zero vector v is given along with a square matrix M, those are called **eigenvectors**. In the event where Av is v's scalar multiple, v will then be an eigenvector.

The term **Covariance Matrix** refers to a matrix that contains the covariance between two variables.

## Results

* Performing PCA

Graphical user interface

Description automatically generated

Graphical user interface

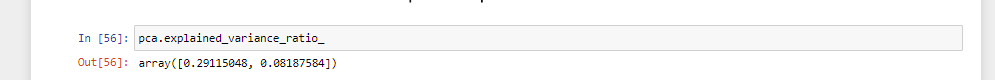
Description automatically generated with medium confidence

Text

Description automatically generated

Chart, scatter chart

Description automatically generated



* Performing Naïve Bayes

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

* Performing K Means for Cluster =3

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

* Performing Logistic Regression

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

## Discussion And Observations

From the above results we have the following Observations that can be discussed below:

* The dataset needs Standard Scaling for the PCA algorithms.
* After performing PCA, the accuracy of Naïve Bayes is as follows:

Training set accuracy: 71.8600

Test set accuracy: 72.2067

* After performing PCA, the accuracy of K Means with 3 components is as follows:

K=3 KMeans -> 43.8600%

* After performing PCA, the accuracy of Logistic Regression is as follows:

Logistic Regression Accuracy - 72.6960%

One of the observations here is that for Logistic Regression after PCA, the accuracy got reduced from 80.85% to 72.69%

Still, we can conclude that after performing dimensionality reduction, both, Logistic Regression and Naïve Bayes suites the Weather Dataset.

## Weka Feature Reduction

Graphical user interface, text

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, table

Description automatically generated

## References

* <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>
* <https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning>
* <https://www.javatpoint.com/logistic-regression-in-machine-learning>
* <https://www.javatpoint.com/principal-component-analysis>
* Dataset - <https://www.kaggle.com/datasets/keshan/the-shifts-weather-prediction-dataset>
* <https://github.com/udacity/machine-learning/blob/master/projects/capstone/report-example-1.pdf>

**Thank You!!!!!!**