**Machine Learning I DATS 6202**

**Group Proposal**

**Group 3**

**Rainfall Prediction**

**Instructor: Amir Jafari**

**Authors:**

**Tanmay Vivek Kshirsagar**

**Sudhanshu Deshpande**

**Shreyas Sunku Padmanabha**

**Date: 04/01/2023**

**Problem Statement**

Predict next-day rain by training classification models on the target variable ‘*RainTomorrow*’ using the observations made today.

**Background**

Predicting whether it will rain tomorrow is important for multiple reasons such as safety, agriculture, business, and personal planning. It helps make daily decisions and reduce risk of failure due to unexpected rain.

**Dataset**

The "WeatherAUS" dataset is obtained from Kaggle. The dataset contains 145460 rows and 23 columns, and is available at the following link: <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>. The dataset is large enough to train a machine learning or different algorithms.

**Machine Learning Algorithms**

As the problem type is a classification one, we plan to use multiple classification models such as MLP Classifier, KNN, Random Forest Classifier, SVC.

**Software Used**

We plan to use packages like *scikit-learn* and *keras* as it contains comprehensive list of implemented machine learning and deep learning algorithms. We will also use *pandas*, *numpy* , *matplotlib*, *seaborn* packages to support our analysis.

**Reference Materials**

We plan to use the following websites to gain knowledge on the software:

Scikit-learn: <https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>

"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron: <https://github.com/Akramz/Hands-on-Machine-Learning-with-Scikit-Learn-Keras-and-TensorFlow>

**Performance Metrics**

We plan to use:

* Accuracy - proportion of correct predictions out of total predictions.
* Precision - proportion of true positive predictions out of total positive predictions.
* Recall - proportion of true positive predictions out of actual positives.
* F1-score - harmonic mean of precision and recall.
* ROC curve - graphical representation of true positive rate versus false positive rate.
* AUC - Area Under the ROC Curve, a measure of the classifier's ability to distinguish between classes.
* Confusion matrix - a table that shows the number of true positives, true negatives, false positives, and false negatives.

**Planned Schedule**

|  |  |
| --- | --- |
| Start Date | Description |
| 03/28/2023 | Dataset Selection and Group Proposal |
| 04/08/2023 | EDA |
| 04/15/2023 | Data Pre-processing |
| 04/22/2023 | Feature Engineering |
| 04/26/2023 | Algorithm and Performance Tuning |
| 04/30/2023 | Final Project Reports |