THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Machine Learning I DATS 6202

(MS In Data Science)

Group Report

Group 3

Rainfall Prediction

Instructor: Amir Jafari

Authors:

Tanmay Vivek Kshirsagar

Sudhanshu Deshpande

Shreyas Sunku Padmanabha

Date: 05/01/2023

ABSTRACT

TABLE OF CONTENTS

1: Introduction

Rainfall prediction is a critical application of machine learning that has a wide range of practical applications, from agriculture to transportation to disaster management. In this project, our objective is to use the dataset to predict whether it will rain on the next day based on various weather observations made on the current day.

We will first investigate the dataset using exploratory data analysis (EDA). We will pre-process the data after examining it to manage missing values, encode category variables, and scale numerical characteristics. Then, using different algorithms such as logistic regression, MLP Classifier and random forest, we will train and evaluate the machine learning models. Finally, we will select the best performing model and fine-tune its hyperparameters using cross-validation. The result will be a machine learning model that can accurately predict whether it will rain on a given day based on the weather observations. This project has practical applications in weather forecasting and risk assessment and can help inform decision-making in various industries.

2: Dataset Description

The weather dataset contains daily weather observations from various weather stations across Australia, spanning from 2007 to 2017. The dataset includes 142,193 instances and 24 features, including temperature, humidity, rainfall, wind speed, and direction, among others. The target variable is *'RainTomorrow'*, which indicates whether it rained on the following day. The data is in a structured format, with mostly numerical and categorical features. The dataset has missing values, which will require data pre-processing before modelling. The weather dataset is a suitable candidate for binary classification tasks related to rain prediction and risk assessment.

3: Machine Learning Algorithms

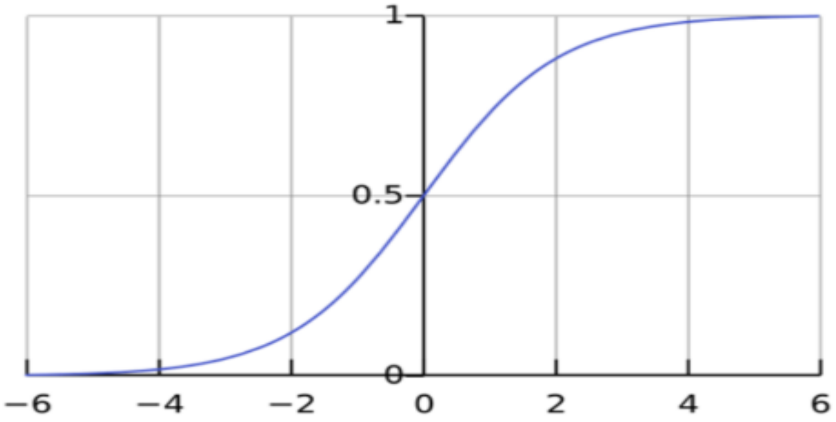
Machine learning algorithms use statistical models and algorithms to identify patterns in data and make predictions or decisions based on that data. As we have a binary classification problem, we have used the following methods.

3.1. Logistic regression

Logistic regression is a machine learning algorithm used for binary classification tasks. It models the relationship between input features and the output variable using a logistic function, which returns a value between 0 and 1. Once trained, the model can be used to predict the probability of the output variable being one of two values based on new input features. Logistic regression is simple and interpretable but has limitations in handling non-linear relationships.

The model parameters are learned by minimizing the cost function using an optimization algorithm like gradient descent. The logistic function has an S-shaped curve that approaches 1 as input values increase and approaches 0 as input values decrease.

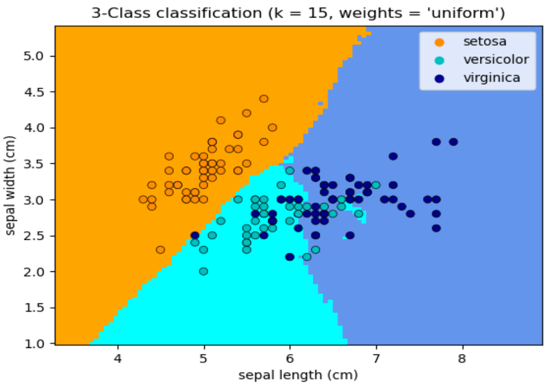
$$f(x) = \frac{1}{1 + e^{-x}}$$



3.2. KNN (K-Nearest Neighbors) Classifier

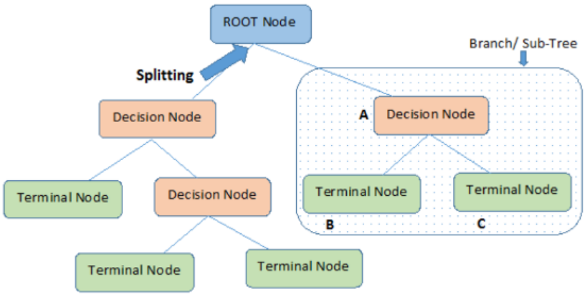
KNN Classifier is a non-parametric machine learning algorithm used for classification tasks. It is based on the idea that similar data points tend to belong to the same class. The algorithm determines the class of a new data point by finding the k nearest neighbors in the training data and assigning the most common class among them as the predicted class for the new data point. The algorithm is simple and easy to implement but can be computationally expensive for large datasets. It also requires careful selection of the hyperparameter k, which determines the number of neighbors to consider.

In summary, K Neighbors Classifier is a simple and effective non-parametric algorithm for classification tasks. It works by finding the k nearest neighbors in the training data and assigning the most common class among them as the predicted class for a new data point.



3.3. Decision Tree Classifier

Decision Tree Classifier is a machine learning algorithm used for classification tasks. It partitions the feature space recursively into subsets based on input features and assigns a class label to each leaf node of the tree. The algorithm determines the best split for each node by maximizing a measure of purity. The decision tree can be graphically represented as a tree structure with each internal node representing a split on an input feature and each leaf node representing a class label. It can handle both categorical and numerical data but is prone to overfitting.

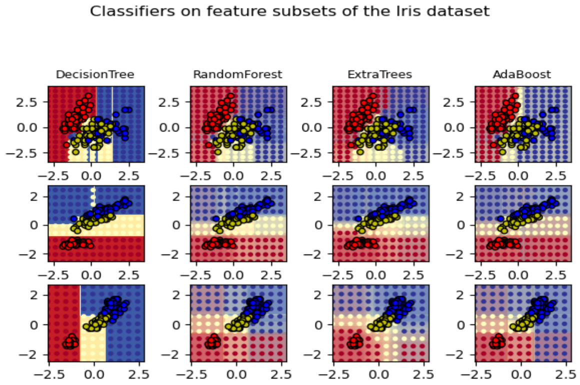


3.4. Random Forest Classifier

Random Forest Classifier is an ensemble machine learning algorithm used for classification tasks. It works by combining multiple decision trees, each trained on a different subset of the training data and a different subset of the input features. The algorithm determines the class of a new data point by aggregating the predictions of all the trees in the forest.

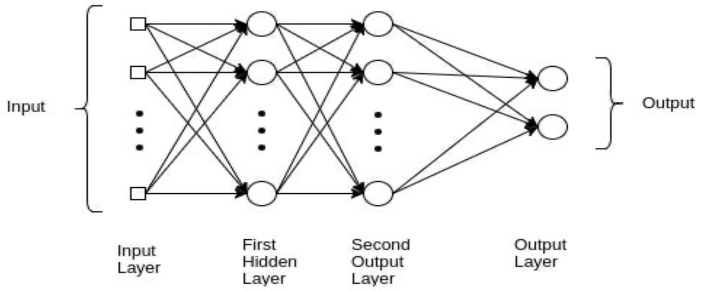
Random Forest Classifier is a powerful algorithm that can handle high-dimensional data and is less prone to overfitting than Decision Tree Classifier. However, it can be computationally expensive and difficult to interpret.

The random forest can be represented graphically as a collection of decision trees, with each tree trained on a different subset of the training data and input features.



3.5. MLP Classifier

MLP Classifier is a type of neural network used for classification tasks that consists of multiple layers of nodes connected to each other. It was developed in the 1980s to handle non-linearly separable problems and can learn complex non-linear relationships between inputs and outputs by adjusting weights to minimize a loss function. MLP Classifier uses activation functions and softmax function to produce class probabilities. It is a powerful algorithm but can overfit if the model is too complex or if there is not enough training data. The basic architecture of MLP Classifier includes an input layer, one or more hidden layers, and an output layer with nodes corresponding to class labels.



4. Experimental Setup

Hello world

5. Results

Hello world

6. Summary and Conclusions

Hello world

References

https://aws.amazon.com/what-is/logistic-regression/
https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/
https://scikit-learn.org/stable/modules/neighbors.html/
https://medium.com/@AI_with_Kain/understanding-of-multilayer-perceptron-mlp-8f179c4a135f

Appendix

Hello world