Regression and Classification Using Different Machine Learning Algorithms on Forest Fire Dataset

Har Shobhit Dayal,
School of Computing Sciences and Engineering,
VIT Chennai, Tamil Nadu, India 600127
Email:dayalharshobhit@gmail.com
9537264094

In this project classification and regression algorithms were applied on Forest Fire Dataset to build various machine learning models. Various classification and regression algorithms were applied in order to see how each algorithm works on the dataset.Algorithms were fine tuned in order to provide the best results.

Through this project we try to develop a machine learning model which helps in predicting whether a particular area in forest will be burned or not burned, based on various climatic conditions and indexes.

Methodology

This paper introduces different machine learning algorithms for classification and regression on Forest Fire dataset in order to classify the points as burned or not burned.
In this project we have made use of the best models for both classification and regression to get accurate results. The methods used for classification are Perceptron,Logistic regression and Decision Trees.The methods used for regression are linear regression and Neural Network.K-Means clustering was also used for feature extraction.
The data has been converted to a log scale to get a proper range of values as initially the points are scattered randomly and a smooth Gaussian curve is not obtained.

Database – Forest Fire dataset

The aim of Forest Fire dataset is to predict the burned area of forest fires, in the northeast region of Portugal, by using meteorological and other data. It consists of measurements with respect to forest fire in particular area and what were the surrounding conditions at that time.
The parameters is as follows:
1)x axis spatial coordinate
2)y axis spatial coordinate
3)Month
4)Day
5)Fine Fuel Moisture Code (FFMC)
6)Duff Moisture Code (DMC)
7) Drought Code(DC)
8)Initial Spread Index (ISI)
9)temperature
10)Relative humidity
11)wind
12)rain
13)Burned area(ha).
There are 517 examples(measurements) in the dataset.

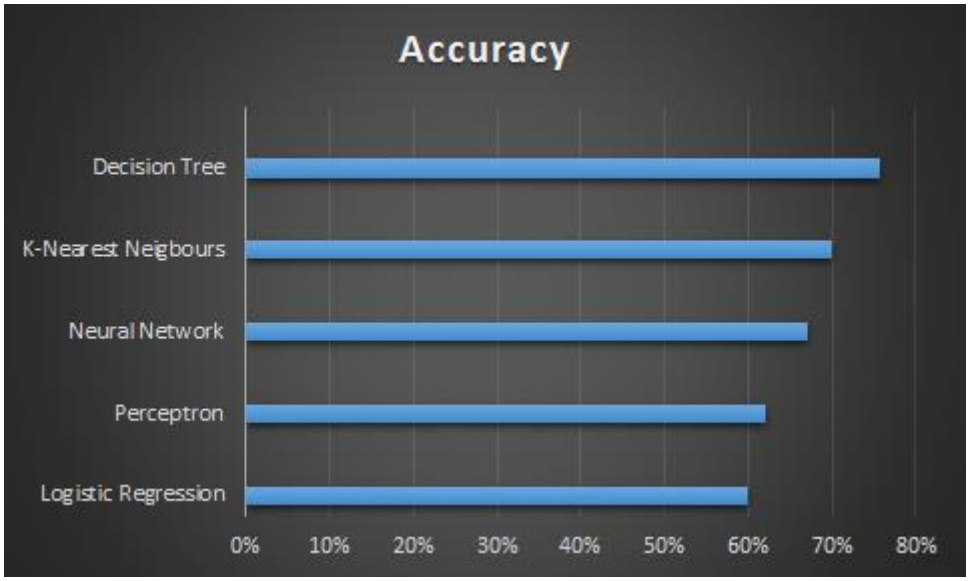
The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger and it includes six components : Fine Fuel Moisture Code (FFMC),Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Build up Index (BUI) and FWI.
The first three are related to fuel codes: the FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity.
The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components. Although different scales are used for each of the FWI elements, high values suggest more severe burning conditions.

Algorithm

All the models were performed by keeping bias and variance in check so that the models are not biased towards a particular input.
The steps followed in each of the machine learning models is as follows:
1) Import the forest fire dataset
2) Import the python machine learning libraries such as pandas, numpy, scikitlearn, matplotlib etc.
3) Apply different machine learning algorithms as mentioned before. These machine learning algorithms used in this research work include-:
1. logistic regression
2. Perceptron
3. Neural networks
4. k nearest neighbours
5. Decision trees
6. linear regression
7. k nearest neighbours
4) Observe the plots obtained after applying these algorithms.
5) Compute the accuracy result and compare them
6) Draw possible conclusions.

Results

Algorithm	Accuracy
Logistic Regression	60%
Perceptron	62%
Neural Network	67%
K-Nearest Neigbours	70%
Decision Tree	75.60%



Accuracy score of different algorithms

The Mean SquaredError(MSE) obtained is 0.14739 according to the linear regression algorithm.

Decision tree for classification and linear regression for regression tasks serves as the best algorithms in terms of accuracy for the given dataset.