UNIVERSITY OF THESSALY
FOUNDED 1984

Problem Solving Environments and Applications in Data Science ECE525

Report

Friday 16 December 2022

Fall Semester

Project 5
Linear and Logistic Regression

Anastasia Psarou 2860
Christos Arseniou 2730

Supervisor: Elias Houstis

1 Introduction

In this project, we focused on implementing linear regression and logistic regression, with the use of the sklearn library.

2 Data

The data that was used for this project came from the National Health and Nutrition Examination Survey (NHANES). This survey was conceived in the early 1960s to provide nationally representative and reliable data on the health and nutritional status of adults and children in the United States

3 Exploratory Data Analysis

Firstly, we implemented Exploratory Data Analysis. This analysis is the process of performing initial investigations on data, in order to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

4 Regression

Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. So, algorithms are trained to understand this relationships and afterwards, the model can be leveraged to predict the outcome of new and unseen input data, or to fill a gap in missing data.

The regression model defines a linear function between the X and Y variables that best showcases the relationship between the two. It is, usually, represented by a slant line, where the objective is to determine an optimal 'regression line' that best fits all the individual data points.

Furthermore, the regression model uses a cost function to optimize the weights. The cost function of linear regression is the root mean squared error or mean squared error (MSE). Fundamentally, MSE measures the average squared difference between the observation's actual and predicted values. The output is the cost or score associated with the current set of weights and is generally a single number. The objective here is to minimize MSE to boost the accuracy of the regression model.

4.1 Linear Regression in height and weight columns

In general, linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. In our project, we used height as the independent variable and weight as the dependent.

In the figure, the height and weight variables are being represented with a linear function, that best showcases the relationship between these two.

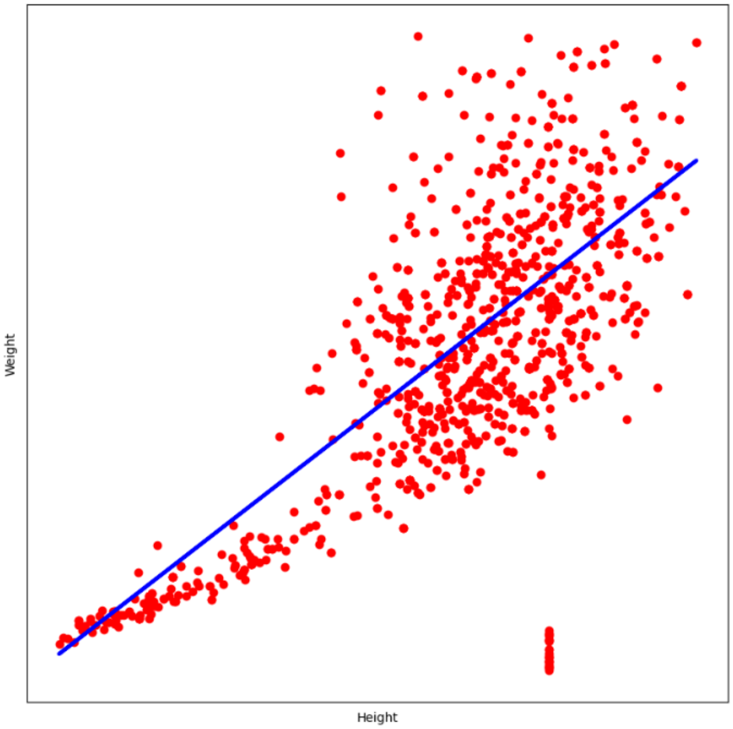

Figure 1: Linear Regression for height and weight

4.2 Multiple Regression at the effect of physical activity, age, and gender on testosterone levels

In simple words, multiple regression is linear regression, with the use of more than one independent features. The model evaluates different weight combinations, in order to determine the line best fits the data. In this example, our plot is represented in the 3d axis, as we want to take into consideration more than two columns for our results.

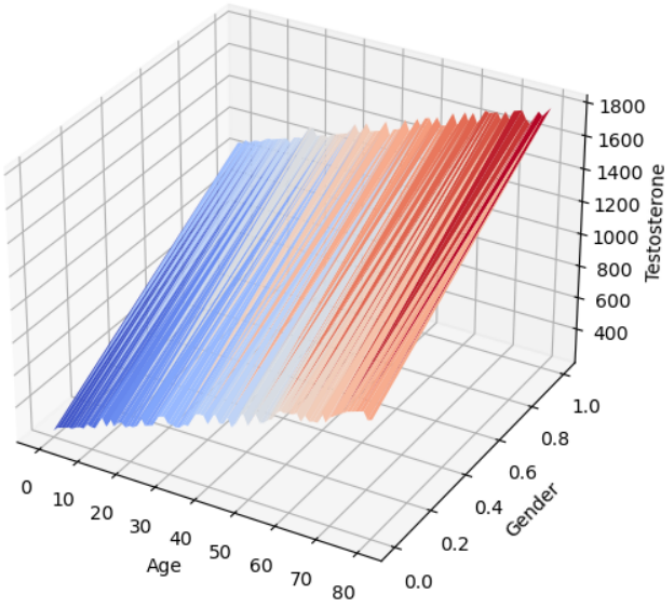

Figure 2: Multiple Regression

5 Logistic Regression

Logistic regression is applicable in cases where there is one dependent variable and more independent variables. The main difference between multiple and logistic regression is that the target variable in the logistic approach is discrete (binary or an ordinal value). Some other differences are that logistic regression isn't represented using linear structures and that it normally contains only two outcomes, ranging between one and zero.

In our project, logistic regression is implemented to fit 'Insured' column using 'Race', 'Age', 'Income' and 'SleepHrsNight' columns. The problem is that the data in 'Insured' column are undersampled. Hence, value 0 occurs very few times in comparison to value 1.

5.1 Without Over Sampling

Firstly, we implemented logistic regression without making any changes in the 'Insured' column. The result was having value 1 always being predicted and value 0 was never being predicted. Also, the accuracy resulting from the perceptron algorithm was quite high. Hence, we decided to balance the dataset and then implement again the logistic regression algorithm.

5.2 With Over Sampling

In order to balance the dataset, SMOTE() algorithm was used and after that the logistic regression algorithm was applied. After implementing these, we noticed that value 0 is also predicted and the accuracy got decreased around 40%.

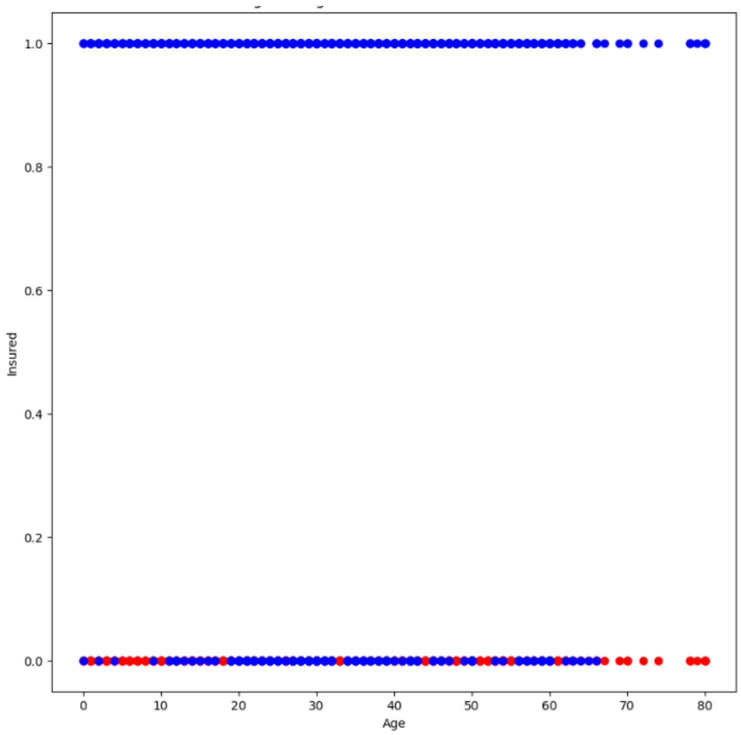

Figure 3: Logistic Regression

6 Conclusion

In a nutshell, linear regression and logistic regression are the two most famous and commonly used algorithms when it comes to machine learning. Linear regression is used to handle regression problems whereas logistic regression is used to handle the classification problems. They are supervised machine learning algorithms and they both can be used to make informed decisions. Logistic regression requires a more balanced dataset, in order to achieve satisfying results, whereas linear regression can be applied in a more imbalanced continues dataset.