Estimation of obesity levels based on eating habits and physical condition Data Set

Project Specification

The goal of this project is to estimate the obesity levels of individuals based on various factors such as their age, height, weight and their habits in general, physical habits, eating/drinking habits and smoking habits.

Only 23% of the values in the dataset provided for this project are genuine, it means that 77% of the data was generated and probably needs cleaning.

The purpose of this project is to gain knowledge and experience in handling errors in the data, commonly known as "outliers“, and applying supervised learning algorithms that were covered in our theoretical class.

Related Work (References)

During the cleaning process of the dataset, we found that reading the attribute descriptions and reviewing the article provided was helpful for a better understanding of the dataset. It helped us understand whichs features could be more important and which types of “outliers” we could have.

Additionally, we found the scikit-learn webpage to be useful for implementing the supervised learning algorithms.

Algorithms to Implement

Like we said in the project description, the dataset will need some cleaning and for that we will need to see the data that is wrongly classified. One possible bad classification that we can think is related to the body mass index, bmi = weight / (height*height).

The algorithms to implement are described in the introduction of the project. The goal is to compare different types of classifiers and identify which one performs best for the given task. The errors generated by each classifier will be analyzed to gain insight into their strengths and weaknesses.

Work Implemented

In this phase, we have already implemented every algorithm described in the introduction of the project and we already cleaned the dataset as we talked in the previous slide.

The dataset was decreased from 2111 rows to 1941 and the best algorithm with an accuracy of 93% is the DecisionTreeClassifier and 88% on the KNeighborsClassifier. An analysis of the errors generated by each classifier remains to be conducted.