**Introduction**

For this assignment we were split into pairs to explore a dataset through machine learning practices. The dataset we chose to work on is about football players. In this dataset we will be trying to classify the release clause of a player based on many other attributes. This dataset is about the stats of all the football players in FIFA 2019 and contains 89 attributes. One of these attributes is the release clause but not every player has a release clause. We will be trying to build a model using the other 88 attributes to predict the release clause of the players which don't have one.

**Data exploration**

The FIFA 2019 dataset has 89 attributes. Some of these attributes are irrelevant to predicting the release clause of an players. Such attributes would be such as player name, link to club logo and link to country of birth flag. We also need to change a few data columns to change the content into a format which we can use. After cleaning up the dataset we will create histograms out of the data to check if they are normally distributed. Another exploration we are going to perform is making an heatmap to check for correlations. One thing we also discovered was that we are dealing with nominal and numerical attributes.

**Pre-processing**

During the pre-processing of the data, we need to convert all the data we have to formats we can use. For this we put everything into a scale from 0-1. Underneath you will find a table showing all the data conversions we have done.

|  |  |
| --- | --- |
| Wage, Value, Release Clause | Changed to a numeric value and then divided it by the highest value of each column to put it in a 0-1 range |
| Preferred foot | Encoded it with 1 = right foot and 0 = left foot |
| International Reputation, Weak foot, Skill moves | Converted the scale from 1-5 to 0-1 |
| Work rate | We applied one-hot encoding and divided it up into 9 different columns |
| Overall, Potential, Skill stats | Converted the scale from 1-100 scale to a 0-1 scale |
| Age | I changed it from a scale between 0-45 where 45 years is the oldest player in the dataset to a scale where this is all in a range of 0-1 |
| Body type | Converted to 3 columns using one-hot encoding |
| Height | We converted the height from feet and inches to only inches than we went from a 0-85 scale and converted it to a 0-1 scale |
| Weight | To make this column usable we removed the suffix of lbs and then changed the scale from 1-250 to a 0-1 scale. |

**Models**

During this assignment we were implementing a k-nearest neighbour model and a naïve bayes model. These 2 models are trying to predict the release clause of a player. K-nearest neighbour does that by comparing the most similar entries to the one you want to predict. By doing this you can guess the most likely release clause. Naïve bayes model does it a bit different. Naïve bayes uses probability of every attribute happening and multiplying those together for nominal cases and in this numerical case you will use the same principle but now in relation to the gaussian distribution. Both models were applied in this assignment and tried. The Gaussian distribution seemed to be giving the best results.