# Machine learning project

The dataset I want to use is the soccer dataset on Kaggle. This contains details about matches, players, betting odds and match events. It has up to 25000 matches and up to 10000 players. The link to the dataset is: <https://www.kaggle.com/hugomathien/soccer>.

This project looks interesting to me because one of my hobbies is soccer and I like to watch and play soccer. Trying to predict soccer matches looks very interesting to me.

I am going to try to predict the outcomes of soccer matches. With the different variables like previous matches, player stats, betting odds and current wins and losses. I think it is an option to use random forest or decision trees. Naïve bayes can be an option but I think the features are correlated too each other. With naïve bayes the assumption is that features aren’t correlated.

With all the data I can use multiple features for predicting the match, I need to determine what features are import and use the most important features. I also need to create variables on my own. Things like previous matches against an opponent need to be queried and are not in the tables yet.