9 Machine Learning and Real-world Data (av308)

You are a football academy manager and you want to design a classifier for deciding which players to recruit to your academy. For this purpose you gather recent data (2020 and onwards) on players about to finish high school. You construct the following table, where each row represents a potential recruit. The column `Success` indicates whether the player was considered a successful recruit or not; it is the label we are trying to predict.

Success	Goals	Position	Gender
Y	Many	Attack	M
N	None	Goalkeeper	F
Y	Few	Defender	M
Y	Few	Attack	M
N	None	Defender	M
N	Few	Defender	F

You want to develop a model to help you in deciding whether you should recruit a player based on their features. You decide to use a Naive Bayes Classifier.

(*a*) Define and estimate the parameters of the Naive Bayes Classifier. [4 marks]

(*b*) There is a suspicion that the model you estimated is biased. If a model is biased, it treats certain groups of the population unfairly, i.e. it never predicts their members to be successful. Identify two such groups and demonstrate the problem by constructing appropriate test instances. [4 marks]

(*c*) Explain which property of the classifier you developed in part (b) enables you to make this judgement based on the parameters you have estimated. Give one reason why it is a useful property for the model to have and one reason why this is a problematic property. [2 marks]

(*d*) You are given more data from an earlier time period, pre-2020. How would you incorporate it in your experimental setup for the classifier you developed in part (a)? [4 marks]

(*e*) The model you developed predicts success or not based on a snapshot of the potential recruits at a single point in time. However you are recruiting for an academy, so you want to take into account the trajectory of the players over a number of years. Propose a modelling solution for this. [6 marks]