Big Data

Coursework

Liam Noonan

S1512127

Lnoona200@caledonian.ac.uk

# The Problem

The provided dataset contains 400 medical records, each with 25 values pertaining to patient information such as their age, blood pressure and blood sugar levels etc and finally whether or not the patient has Chronic Kidney Disease. Presumably the purpose of this investigation is to determine if there is a relationship between any of these data points and the likelihood of a patient having Chronic Kidney Disease. 11 of the values are Boolean with the remaining 14 being numeric.

Removing rows that are missing values leaves 156 data points, a reduction from the original 400. Based on this remaining data, more records are classed as not having CKD than those who do have it. This varies from the original unfiltered dataset which showed a larger number of patients with CKD.

# Construction and Tuning

# Testing

# Discussion

Before getting rid of nulls 237/400 had CKD, after nulls 114/156 did not meaning getting rid of nulls fucked the data set. The rest of the values stayed quite similar.