CO3093 COURSEWORK 2 - Report

Big Data & Predictive Analytics - Classification & Clustering

Ihtasham Chaudhry

Department of Informatics
University of Leicester
28th March 2018

Exploring the data

In this section we will explore the data-set `Diabetes 130-US hospitals for years 1999-2008`.

Exploring the data as a whole

To visualise and explore the data-set it's important to extract key-information as some of the columns in the data-set are not key in analysing the data.

Firstly, we must remove any rows/columns that have missing values and that we may not be able to use in order to conduct any further experiments or analysis on. There seem to be many rows with `?` values, i.e. unknown values and mostly on the columns `weight` and `payer_code`. So we can remove these as they will not be relevant in either clustering or analysing the data-set.

From the numerical data we can make the following analyses:

- The mean time spent in hospital by a patient who has been admitted is approximately 4 days.

At a basic level we can visualise the distribution of diabetic patients by race and gender, the results of this can be seen below.

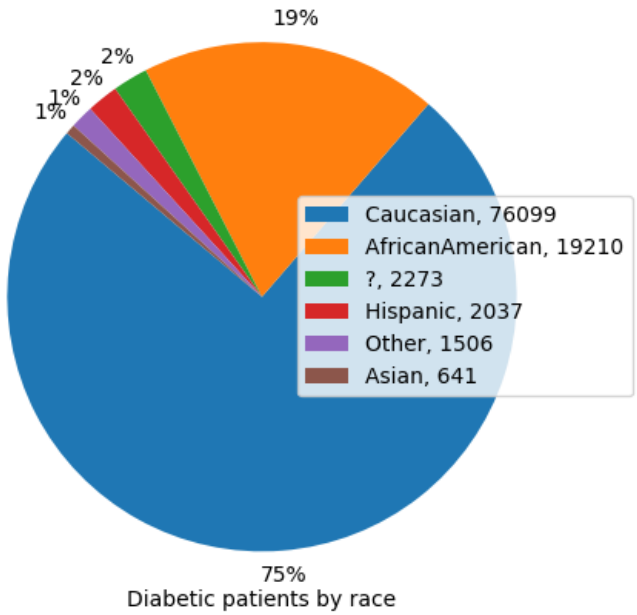

Figure 1: Proportion of diabetics by race

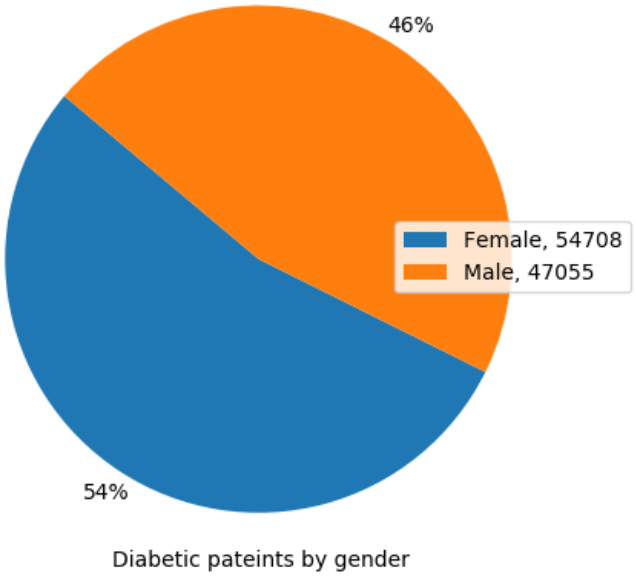

Figure 2: Proportion of diabetics by gender

From this we can see that majority of diabetic patients admitted to US hospitals from 1999 to 2008 are mostly Caucasian (75%) by analysing the gender split, it's almost equal split however a higher percentage of females (54%) than males (46%).