Data Science 874 Post block Assignment 1

Jeandre de Bruyn
19768206

ABT before cleaning:

Feature	Count	Miss%	Card	Min	1st Qrt	Mean	Median	3rd Quart	Max	Std Dev
age	5110	0.00%	104	0.08	25	43.23	45	61	82	22.61
avg_glucose_level	5110	0.00%	3979	55.12	77.25	106.15	91.89	114.09	271.74	45.28
BMI	4909	3.93%	418	10.30	23.50	28.89	28.10	33.10	97.60	7.85

Table 1: Continuous feature ABT

Feature	Count	Miss%	Card	Mode	Mode Freq	Mode %	2nd Mode	2nd Mode Freq	2nd Mode %
hypertension	5110	0.0%	2	0	4612	90.30%	1	498	9.70%
heart_ disease	5110	0.0%	2	0	4834	94.60%	1	276	5.40%
ever_ married	5110	0.0%	2	TRUE	3353	65.62%	FALSE	1757	34.38%
work_ type	5110	0.0%	5	Private	2925	57.24%	self-employed	819	16.03%
residence_ type	5110	0.0%	2	Urban	2596	50.80%	Rural	2514	49.20%
smoking_ status	5110	0.0%	4	never smoked	1892	37.03%	Unknown	1544	30.22%
stroke	5110	0.0%	2	0	4861	95.13%	1	249	4.87%
gender	5110	0.0%	3	female	2994	58.60%	male	2115	41.40%

Table 2: Categorical feature ABT

Feature study:

There are the following features in the dataset:

ID

Each entry in the dataset has a unique ID

Gender

Categorical type with 3 possible values: Male, Female or other. There is only one value for "other", and the rest are either male or female, as seen in the following pie chart. We can see from the chart that there are considerably more females present than males in this dataset. There are almost 1.5 times as many females as there are males. For the convenience of modelling, the single entry of "other" gender will be removed.

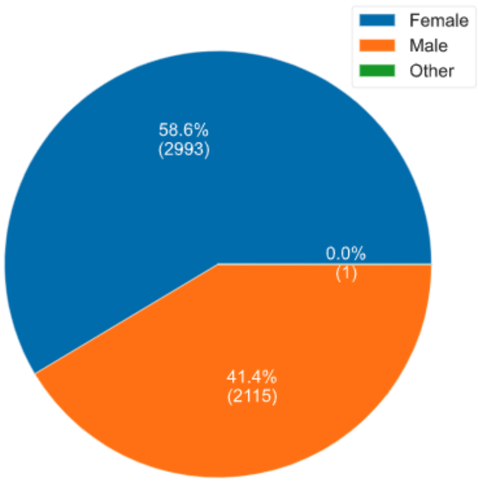


Figure 1: Pie chart showing the gender distribution

Age

Absolute numerical data type with possible values being real numbers between the minimum of 0.08 and maximum of 82. Below is a kernel density estimate (KDE) of age in the dataset. The best way to describe this distribution is unimodal (normal), despite the slight hump towards the end.

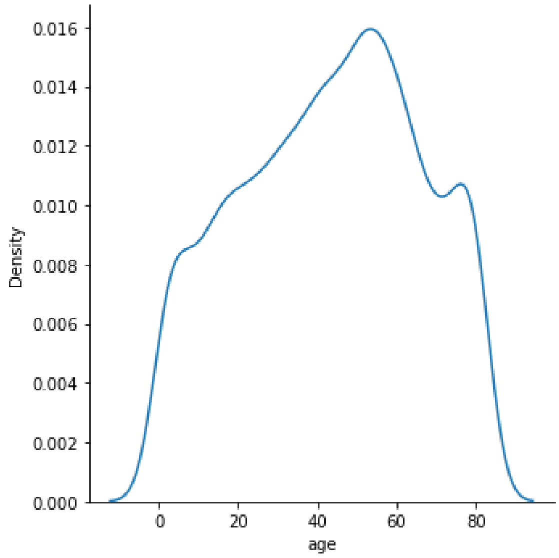


Figure 2: KDE showing the distribution of age through the dataset

Regarding data quality, we find that there are some extreme minimum values such as 0.08, 0.16 and so on. There are 43 values less than 1, 77 values between 1 and 2, 55 values between the values of 3 and 2. However, these are all marked as work type children, but so they won't be marked for data cleaning. Using a blanket rule on ages under 1 like 0.08 and removing the 0. Changing the value to an 8 will not work for them all since there are values like 0.64, which would change to an age of 64 and designated as a child.

However, at age 13, things take a twist. There are several entries whose work type is not "children" but instead something else like Self-employed. For example,

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
251	16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	Unknown	0
410	54975	Male	7.0	0	0	No	Self-employed	Rural	64.06	18.9	Unknown	0
455	7351	Male	13.0	0	0	No	Private	Urban	92.14	23.2	never smoked	0
939	16556	Male	13.0	0	0	No	Never_worked	Rural	111.48	20.8	Unknown	0
1063	42821	Female	13.0	0	0	No	Private	Rural	60.69	24.0	smokes	0
1789	13862	Female	13.0	0	0	No	Never_worked	Urban	70.93	22.9	never smoked	0
1809	18179	Male	13.0	0	0	No	Private	Rural	99.44	21.0	never smoked	0
1976	46577	Female	13.0	0	0	No	Private	Urban	77.63	31.7	never smoked	0
2112	9199	Male	13.0	0	0	No	Self-employed	Urban	74.19	31.1	formerly smoked	0

Figure 3: Screenshot showing entires with age 13 and under and that arent classified as work type "children"

In this case, there are several outliers: work type, smoking status, BMI. One can try and guess what values were entered incorrectly, or the entry might just be valid and just an anomaly. Not all entries will fit the norm as there is always noise in the data

Hypertension

A boolean value of either 0 meaning no; or 1 meaning yes. This feature, like some of the others, has very few positive cases and is underrepresented at 9.7% positive for hypertension.

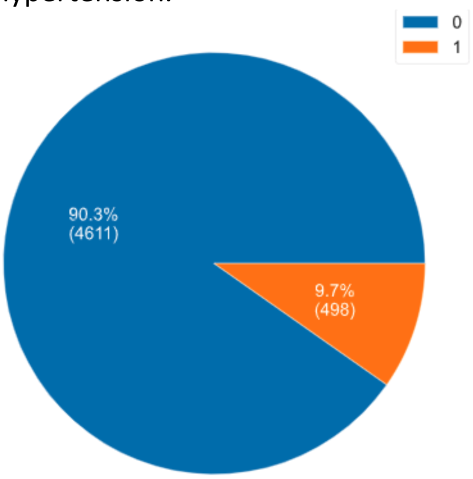


Figure 4: Pie chart showing the distribution of positive entries for hypertension versus negative entries

Heart disease

A boolean value of either 0 meaning no; or 1 meaning yes. Very few positive entries at around one in every twenty being positive for heart disease.

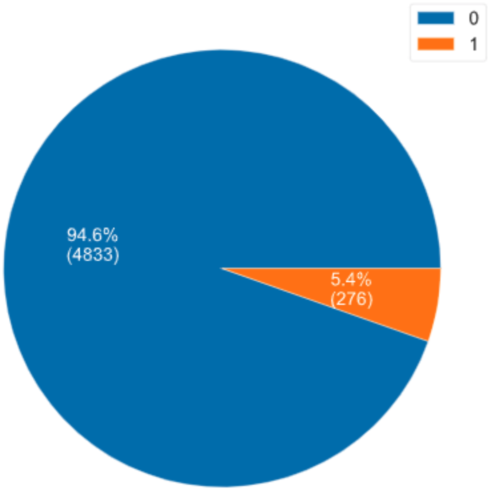


Figure 5: Pie chart showing the distribution of positive entries for heart disease versus negative entries

Ever Married

Categorical value with possible values being True or False. Around double the number of entries is married versus unmarried with a 65.6:34.4 ratio.

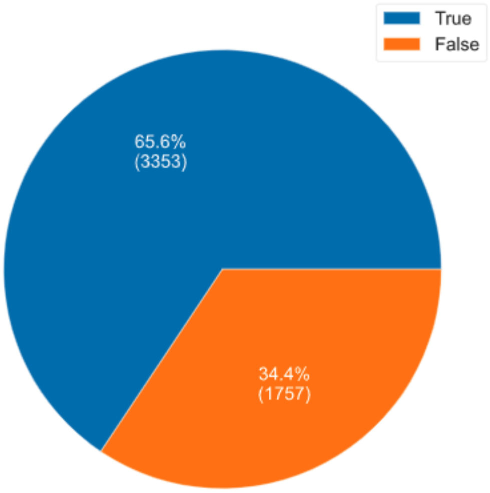


Figure 6: Pie chart showing the distribution of married versus unmarried entries

Work Type

Categorical value with possible values being: (Private, self-employed, children, Govt_job, Never_worked). Over half the entries fall under the "private" entry with the other possible values more or less equally sharing the rest of the distribution except for never worked which makes up a small 0.4%.

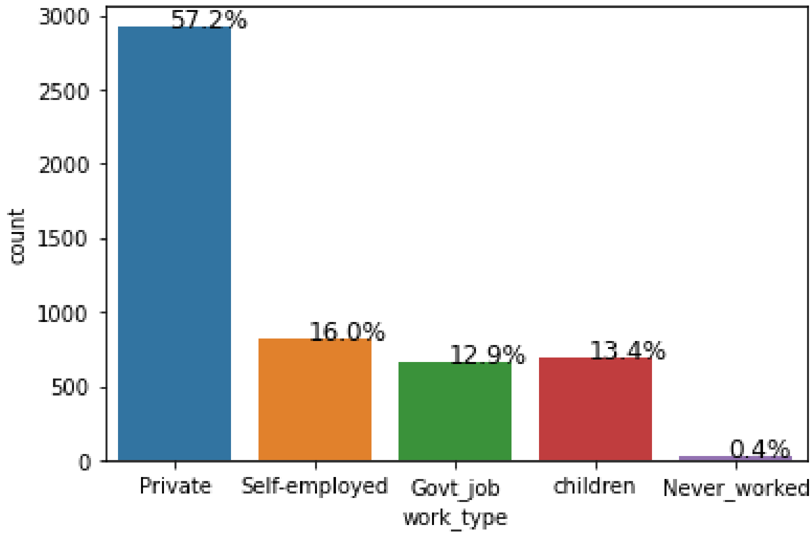


Figure 7: Histogram showing the different work_type values distribution

Residence Type

Categorical with two possible values being Urban or Rural. A near 1:1 split between urban and rural.

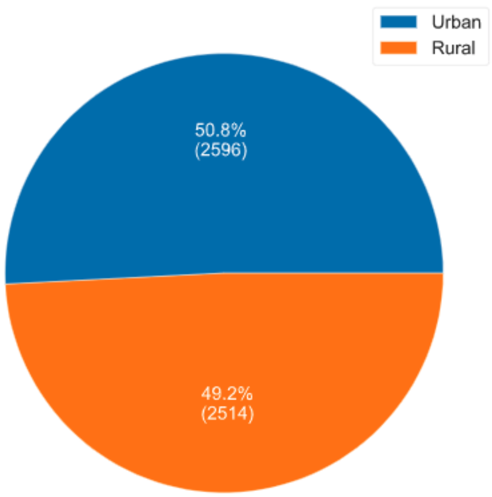


Figure 8: Pie chart showing the distribution of Urban and Rural entries in the dataset.

Avg_glucose_level

Real number with a minimum of 55.12 and a maximum of 271.74. This is a clear multimodal distribution with a clear second grouping around a value of 200.

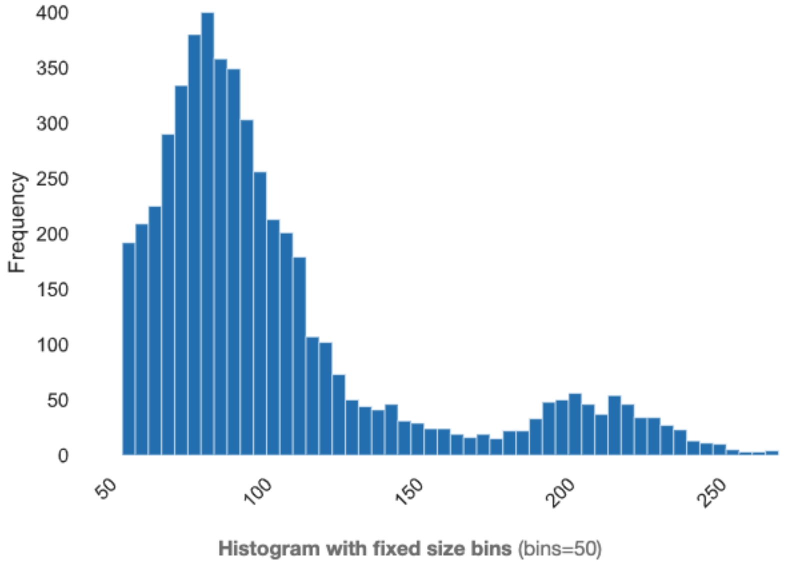


Figure 9: Histogram showing the distribution of average glucose level

Bmi

Real number with a minimum of 10.3 and a maximum of 97.6. There are 201 missing values which accounts for 3.9% missing.

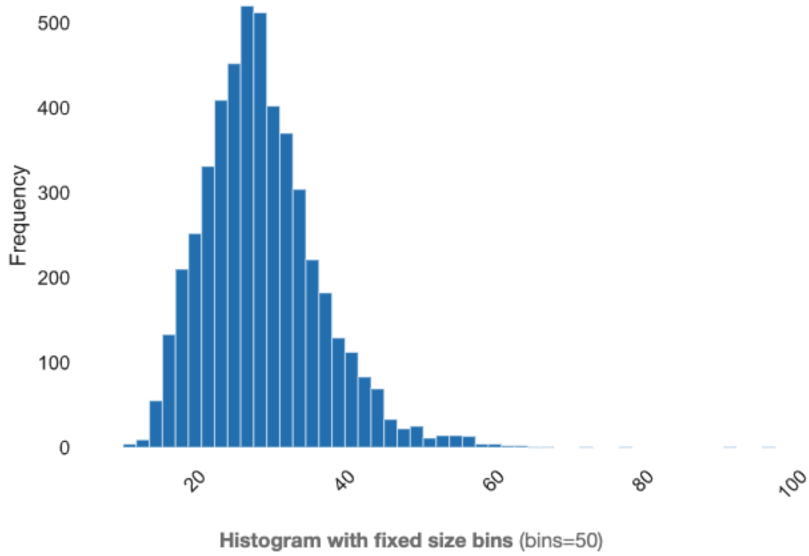


Figure 10: Histogram showing the distribution of BMI

Smoking_status

Categorical variable with possible values being: (never smoked, Unknown, formerly smoked, smokes).

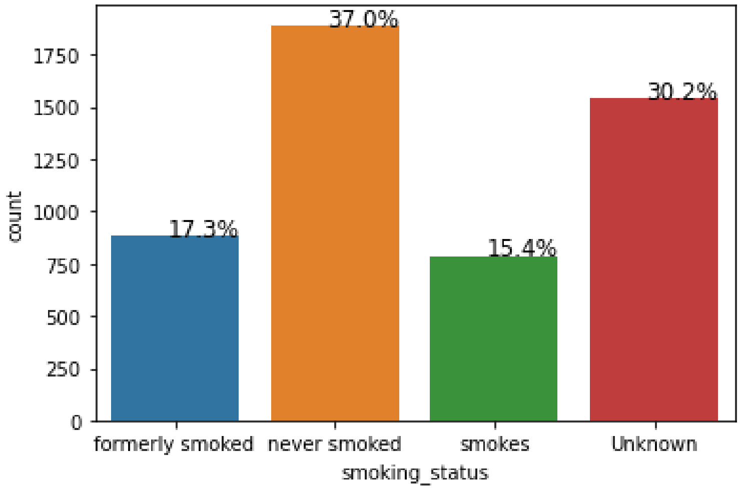


Figure 11:Histogram showing the distribution of possible smoking_status values

Stroke

A boolean value of either 0 meaning no or 1 meaning yes. Very few positive entries at around one in every twenty being positive for Stroke.

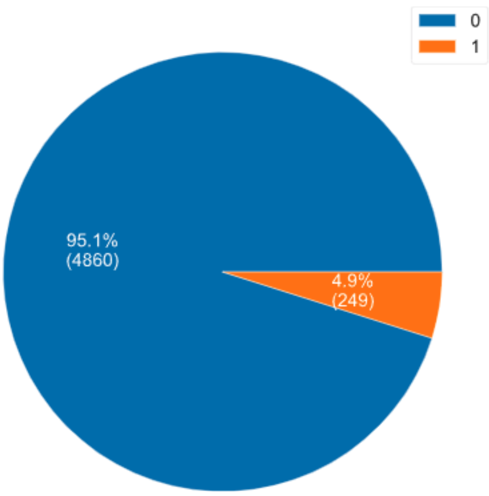


Figure 12:Pie chart showing the ratio of patients who have suffered a stroke versus not

Data Quality Plan

Feature	Data Quality Issue	Potential handing strategies
BMI	missing values(3.9%)	drop all rows with missing BMI values
BMI	outliers(high)	clamp transformation(0,50)
gender	outlier	remove "other" entry

Table 3: Table detailing the data quality plan

Dropping the rows that contain missing BMI entries leaves us at with 4909 entries in the dataset rather than 5110, a very small loss.

To handle the outliers in the BMI feature we will clamp the possible values to a maximum of 50. This will affect 79 rows which equate to 1.61% of the dataset (4909 entries) after the missing values have been dropped. This upper BMI value of 50 has been chosen from the chart shown below taken from the WHOs recommended BMI values. There are some outliers on the lower end of the BMI entries but they are close to the lower end of the WHO scale and so will be left alone.

Category	**BMI range - kg/m^2**
Severe Thinness	< 16
Moderate Thinness	16 - 17
Mild Thinness	17 - 18.5
Normal	18.5 - 25
Overweight	25 - 30
Obese Class I	30 - 35
Obese Class II	35 - 40
Obese Class III	> 40

Figure 13: Table showing WHO BMI brackets

Comparing the 3rd quartile to median range and the 3rd quartile to maximum range before and after the clamp transformation we see that the upper and lower ranges are much more similar rather than the huge difference before.

Min-1st Quartile range	**1st Quartile-median range**	**Median-3rd Quartile range**	**3rd Quartile-Max range**	
13.2	4.6	5	64.5	**Before clamp**
13.2	4.6	5	16.9	**After clamp**

After the clamping of the upper BMI values, the single entry for "other" gender is removed taking the total number of observations 4908 and the gender feature is much more suitable for use in a model.

Post Data Clean ABT

Feature	**Count**	**Miss%**	**card**	**Min**	**1st Qrt**	**Mean**	**Median**	**3rd Quart**	**Max**	**Std Dev**
age	4908	0.00%	104	0.08	25	42.868	44	60	82	22.56
avg_glucose_level	4908	0.00%	3851	55.12	77.07	105.3	91.68	113.5	271.74	44.43
BMI	4908	0.00%	418	10.3	23.5	28.89	28.1	33.1	50	7.47

Table 4: Continous feature ABT post data clean

Feature	Count	Miss%	Card	Mode	Mode Freq	Mode %	2nd Mode	2nd Mode Freq	2nd Mode %
hypertension	4908	0.00%	2	0	4457	90.81%	1	451	9.19%
heart_ disease	4908	0.00%	2	0	4665	95.05%	1	243	4.95%
ever_ married	4908	0.00%	2	TRUE	3204	65.28%	FALSE	1704	34.72%
work_ type	4908	0.00%	5	Private	2810	57.25%	self-employed	775	15.79%
residence_ type	4908	0.00%	2	Urban	2490	50.73%	Rural	2418	49.27%
smoking_ status	4908	0.00%	4	never smoked	1852	37.73%	Unknown	1483	30.22%
stroke	4908	0.00%	2	0	4699	95.74%	1	209	4.26%
gender	4908	0.00%	2	female	2897	59.03%	male	2011	40.97%

Table 5: Categorical feature ABT post data clean

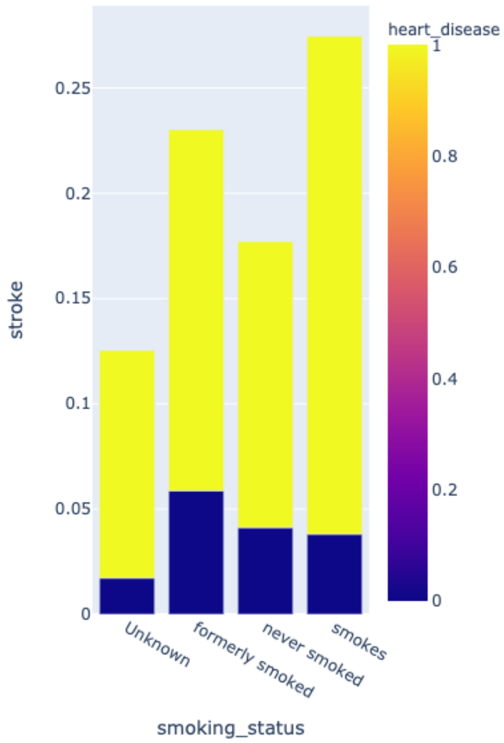
Model Choice

The choice of model is between an SVM or logistic regression. These two are selected because they are both highly proven in the classification category and the medical field with applications like cancer detection. Logistic regression would probably be the better fit since SVM only tries to find the best line that separates the positive class and the negative class whereas logistic regression can be more fine-tuned and have more complex decision boundaries[1]. This is advantageous in a dataset such as this where there are a large number of input variables.

[1] https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16#:~:text=Difference%20between%20SVM%20and%20Logistic%20Regression&text=SVM%20works%20well%20with%20unstructured,is%20based%20on%20statistical%20approaches.

Relationships

Using a scatter plot with so many data points can obscure some of the information that can be gained. The following chart is a stacked bar chart showing the most informative relationship in the dataset. The mean value of heart disease categorised by smoking status and whether or not the individual has had a stroke.



What we can see from this chart is that if for the entries that do not have heart disease (blue stack) then the mean value of stroke (height of blue stack) are much lower than the entries that have heart disease (yellow stack). Comparing the heights of the blue stacks we can see that smoking_status has an impact on the mean value of stroke. Former smokers seem to have more positive entries for stroke than current smokers. Having heart disease however we can see drastically affects the average stroke value. As such we can conclude from the chart that both heart disease and smoking_status have a predictive power, but heart disease is the more informative of the two.