Dataset Name: Stroke Prediction Dataset

Group Name: Group 3 On Campus/cloud: cloud

|  |  |  |
| --- | --- | --- |
| Student ID | Student Full Name | Individual contribution \* |
|  | Dale Orders |  |
|  | David Adams |  |
|  | Josh Anthony |  |
|  | Linden Hutchinson |  |
|  | Roger Middenway |  |
| 218330434 | aaron norwood |  |

\* 5 – Contributed significantly, attended all meetings

4 – Partial contribution, attended all meetings

3 – Partial contribution, attended few meetings

1 – No contribution, attended few meetings

0 – No contribution, did not attended any meetings

NOTE: IF ANY OF THE CELLS IN INDIVIDUAL CONTRIBUTION MARK IS EMPTY ALL STUDENTS WOULD GET 3 MARK BY DEFAULT

**(Page count starts here & Minimum font size – 11; Maximum font size - 12)**

|  |
| --- |
| **Section 1: Introduction and getting to know your data** (max 2 pages) |

For this assignment our group voted on using a Stroke Prediction dataset which was downloaded directly from Kaggle. Our aim in using this dataset will ultimately be to train a model capable of accurately predicting whether an individual will have a stroke. In order to achieve this goal, we will first undergo a process of cleaning and preparing the data. Our first looks through the data revealed issues such as missing BMI values, inconsistent data formatting and a large amount of data which was numerical or categorical and was therefore not ready to be compared easily as well as a range of other problems. Our group will be addressing each of these problems in different manners as will be outlined later in this report.

This report will discuss key correlations we have found throughout our analysis, both where they were expected correlations and where those correlations were not as strong as expected. Several visualisations of the data will be presented so as to better understand these correlations. In doing this, we will highlight the most important factors on whether individuals had a stroke throughout this report.

**Hints:**

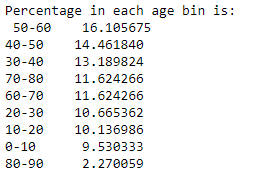
* Why and how the data was collected (the main problems, aims etc.)
* Discussing existing features/variables and their values (numerical, missing values, any classes identified etc.)
* Your initial observations and plans for Exploratory Data Analysis

|  |
| --- |
| **Section 2: Exploratory Data Analysis and Results** (max 7 pages) |

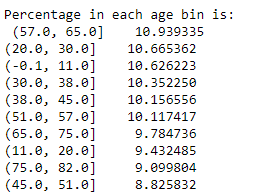
**Data Binning (this will probably move somewhere else in this section, I’ll also shorten it if needed)**

The stroke prediction dataset contained multiple columns with numerical values such as age, BMI and average glucose level which our group wanted to be able to compare to categorical data columns such as whether the patient has had a stroke, heart disease, hypertension and others. To make this possible, we added in binned age, BMI and average glucose levels columns to the data frame so each of these columns could be treated as categorical.

After initially exploring some equal-width binning methods, it became apparent that this method may be problematic; constructing equally wide bins with usually resulted in either bins which were too broad or some bins not containing a high enough percentage of the dataset. For example, when using equal-width binning for age with explicitly defined labels starting at 0-10 and incrementing by 10 with each label, we ended up with the following breakdown of data per bin:



When using an equal-size binning approach and allowing the bin labels to be defined dynamically, we achieved the following percentages per bin for age:



Given the equal-size achieved a better distribution of data to each bin for age and equal-width binning seemed even less important for BMI and average glucose and equal-size binning method was implemented for these variables too. These methods resulted in a similar distribution to those outlined above for age.

**Hints:**

* Data cleansing, like missing data or measurement error, data type of columns, etc.
* Establishing the main structure and key variables; Correlation between variables, looking for abnormality, groups, subgroups
* Investigation interesting patterns, like identifying relationships between variables that are particularly interesting or unexpected.
* Checking assumptions/hypothesis in relation to a specific pattern.
* Using effective visualizations to demonstrate findings/results

|  |
| --- |
| **Section 3: Conclusions** (max 1 page) |

**Hints:**

* Explain what are you main observations after performing Exploratory Data Analysis
* Any suggestion on what problems could be considered for your Machine Learning project?

|  |
| --- |
| **Section 4: References** (max 1 page) |