**Assignment 1**

**Data Exploration and Classification**

**Semester 1 2024**

**Student Name: William Bank Sukjaem**

**Student ID: 18029208**

## **PAPER NAME:** Foundations of Data Science

**PAPER CODE:** COMP615

**Due Date:** Sunday 14 April 2024 (midnight)

**TOTAL MARKS:** 100

**Instructions:**

1. **The following actions** may be deemed to constitute a breach of the General Academic **Regulations Part 7: Academic Discipline,**

* Communicating with or collaborating with another person regarding the Assignment
* Copying from any other student work for your Assignment
* Copying from any third-party websites unless it is an open book Assignment
* Uses any other unfair means

1. **Please email** [**DCT.EXAM@AUT.AC.NZ**](mailto:DCT.EXAM@AUT.AC.NZ) **if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
2. **Attach your code for all the datasets in the appendix section**.

Table of Contents

[**PAPER NAME:** Foundations of Data Science 1](#_Toc163139672)

[Introduction to the dataset (100-200) 2](#_Toc163139673)

[1. Data Exploration 2](#_Toc163139674)

[Data Classification Models - Create Decision Tree 6](#_Toc163139675)

[Results and Discussion (400-500) 6](#_Toc163139676)

[References 6](#_Toc163139677)

# Introduction to the dataset (100-200)

The chosen data covers the topic of Maternal Health Risks in pregnant women with the data being collected from different maternal health care places in rural areas of developing countries, more specifically Bangladesh. The research relevant to the dataset analyses biometric data from wearable IoT devices from women during maternity with the goal being to identify, compare, and analyse relationships between the collected biometric data to mitigate and/or reduce maternal health risks.

This analysis is under the following assumption:

* The information from the dataset is accurate.
* The women chosen for the accumulation of the dataset are moderately healthy.

# Data Exploration

**1.1 Data Types and Statistics Summary**

The data set consists of 1014 instances and 7 attributes where 4 of the attributes are of type integer, 2 of type float, and 1 of type object. Six of the seven attributes are numerical being Age, SystolicBP, DiastolicBP, BloodSugar, BodyTemp, and HeartRate while the remaining is categorical being RiskLevel.

To briefly explain these attributes. The systolic blood pressure and diastolic blood pressure is the upper and lower values of blood pressure, respectively, which is measured in millimeter of mercury (mmHg). Blood sugar, which measures blood glucose levels (mmol/L), and body temperature, measured in Fahrenheit (°F).

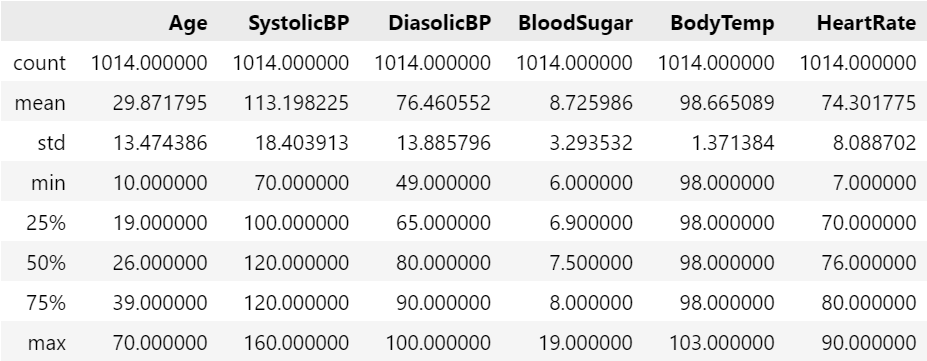


Figure 1: Summary statistics of continuous numerical features

The summary statistics in figure 1 provides key insight into biometric data of women in maternity and interesting statistical data. It is important to note that some of the data is not fully representative of the general population of women in maternity since the collection of data for the dataset was conducted within a single country.

In this dataset, the mean age of women in maternity is 29.8 years old with the lowest being 10 and maximum being 70. In this case, the minimum and maximum age are on the extreme ends of the spectrum which in reality will be quite rare when comparing amongst the general population indicating potential outliers. The mean age is normal when comparing to New Zealand’s average age of mothers at the time of child birth (Statistica, 2023).

The mean blood sugar level is 8.7mmo/L and the median being 7.5mmo/L. The minimum and maximum values being 6.0mmo/L and 19.0mmo/L respectively. The values for blood sugar has an extremely large range of 13mmol/L. According to Diabetes Qubec, baseline maximum for blood sugar level in pregnant women is around 7.8mmol/L putting the median within range for normal blood sugar levels. The mean has a high value due to it being skewed by the high maximum value.

The mean heart rate is 74.3 bpm where the minimum is 7.0 bpm and maximum being 90 bpm. The minimum heart rate is abnormally low and is most likely an outlier. Heart rate is slowest when a person is sleeping which puts the heart rate at around 40 to 60 bpm. The range for heart rate is extremely large at 87 bpm.

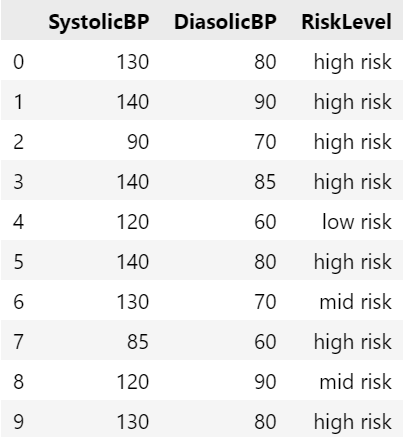


Figure 2: Table comparing Blood Pressure to Risk Level

For the blood pressure values, the mean SystolicBP is 113.20mmHg and DiasolicBP is 76.46mmHg. For the minimum and maximum, SystolicBP has values of 70mmHg and 160mmHg, and DiasolicBP has values of 49mmHg and 100mmHg. The mean values for blood pressure is normal and within the expected range healthy range of 120mmHg and 70mmHg or lower (Heart Foundation, n.d). However, when blood pressure exceeds 120mmHg, there is an overall higher risk of health complications shown when comparing RiskLevel to SystolicBP and DiasolicBP seen in figure 2 that shows first 10 rows of the dataset. High blood pressure is seen in women 20 weeks before or after who have Chronic Hypertension or Preeclampsia which is a serious disorder resulting from high blood pressure (ACOG, 2022). These abnormally high values, where SystolicBP is greater than 140mmHg, may be potential outliers in the dataset.

**1.2 Dataset Cleanliness**

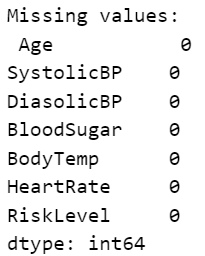


Figure 3: Missing Values

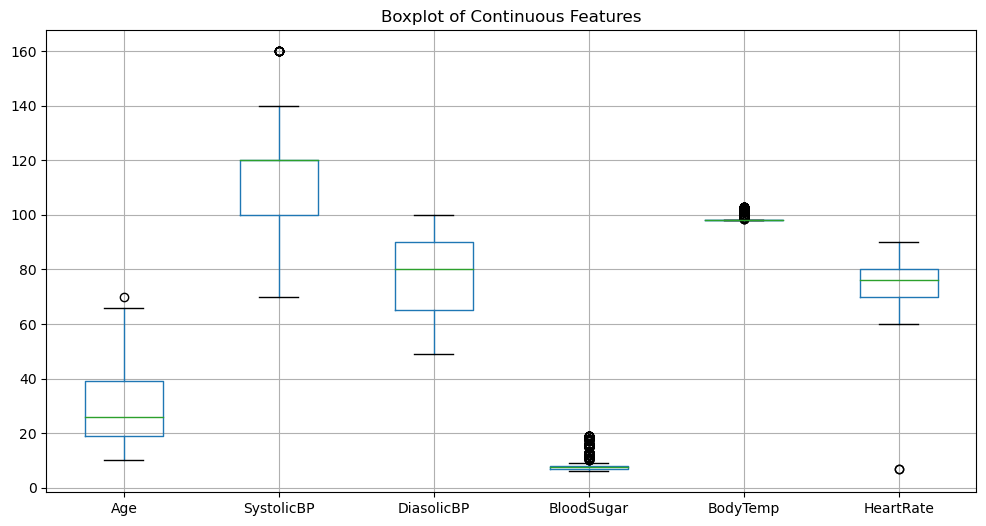


Figure 4: Boxplot of Continuous Features

In figures 3 and 4 helps in determining the cleanliness of the dataset. Figure 3 shows that there are no missing data in each attribute and in figure 4 shows the distribution of data points across each continuous attribute.

Upon looking at figure 4 and based on previous analysis on the statistics summary, there are certainly outliers that are not within reason of the general sample population of the dataset due to their extreme variance relative to the mean in their respective categories. By using the z-score strategy to identify and remove these potential outliers will give a better representation of the sample population.

A diagram of a graph

Description automatically generated with medium confidence

Figure 5: Boxplot of Continuous Attributes without Outliers

Figure 5 shows what the distributions of the dataset look like after removing the extreme outliers detected using the z-score method.

**1.3 Illustration and Explanation of Features**

*This section of your report must discuss the dataset and any features you consider relevant to the analysis and modelling task.*

*~~• How many features (attributes) and instances exist, and what data types are these?~~*

*• ~~Provide summary statistics of the continuous numerical features.~~*

*~~• Perform an initial exploration of the provided dataset to assess its cleanliness. Describe the steps taken to address both data cleanliness evaluation and data cleaning strategies.~~*

*• Illustrate the features of your dataset using meaningful boxplots, histograms and grouped scatter plots (remember, these plots allow you to analyse the individual distribution of features and the relationship between them).*

*• Explain what you can learn from your data exploration and visualisations provided*

# Data Classification Models - Create Decision Tree

# Results and Discussion (400-500)

# References

[**https://www.heart.org/en/health-topics/high-blood-pressure**](https://www.heart.org/en/health-topics/high-blood-pressure)

[**https://www.niddk.nih.gov/health-information/diabetes**](https://www.niddk.nih.gov/health-information/diabetes)

[**https://medlineplus.gov/ency/article/001982.htm**](https://medlineplus.gov/ency/article/001982.htm)

[**https://www.statista.com/statistics/1081327/new-zealand-median-age-of-mothers-at-childbirth/**](https://www.statista.com/statistics/1081327/new-zealand-median-age-of-mothers-at-childbirth/)

[**https://www.heartfoundation.org.nz/wellbeing/managing-risk/managing-high-blood-pressure**](https://www.heartfoundation.org.nz/wellbeing/managing-risk/managing-high-blood-pressure)

[**https://www.acog.org/womens-health/faqs/preeclampsia-and-high-blood-pressure-during-pregnancy#:~:text=You%20have%20gestational%20hypertension%20when,normal%20blood%20pressure%20before%20pregnancy**](https://www.acog.org/womens-health/faqs/preeclampsia-and-high-blood-pressure-during-pregnancy#:~:text=You%20have%20gestational%20hypertension%20when,normal%20blood%20pressure%20before%20pregnancy)**.**

[**https://www.diabete.qc.ca/en/diabetes/information-on-diabetes/diabetes-in-pregnancy/**](https://www.diabete.qc.ca/en/diabetes/information-on-diabetes/diabetes-in-pregnancy/)