|  |
| --- |
| **Data Analytics and Visualization (AE2 Report)** |
| Oluwatoyosi Adetola (Q15929469) |
|  |
|  |
|  |

**Table of Contents**

# **Introduction**

Modern organizations now make data driven decisions. These decisions can be guided using the appropriate data, thorough analysis, and other important factors. The combination of math, statistics, big data, analytic tools, specialized programming, and artificial intelligence in other to discover patterns, actionable insights and make data driven decisions is known as Data science (IBM 2022).

Data science combines several fields, including statistics, mathematics, software programming, data engineering, data preparation, data mining, predictive analytics, machine learning, and data visualization. Skilled scientists are generally responsible for it, however entry-level data analysts may also be engaged (Steadman 2022). Simply put, Data science is the study of how to extract non-obvious and practical patterns from huge data sets using a set of concepts, issue descriptions, algorithms, and methods (Kelleher and Tierney 2018).

To characterize a new profession that would enable the comprehension and analysis of the massive volumes of data that were being gathered at the time, the phrase ‘Data science’ was coined in the early 1960s. At that time, it was impossible to foresee the volumes of data that would be generated over the following 50 years. Data science is a discipline that is constantly evolving and constantly developing by utilizing computer science and statistical methods to acquire insights and generate valuable predictions in a variety of industries. In addition to being employed in fields like business and astronomy, data science is also applied in health (Foote 2016).

Large quantities of valuable data on patients’ demographics, treatment plans, outcomes of medical examinations, insurance, etc. are produced by the healthcare sector. Data scientists are interested in the data gathered by internet of things (IoT) devices. The vast amounts of fragmented, structured, and unstructured data generated by healthcare systems may be processed, managed, analyzed, and assimilated with the help of data science. Data science has been used in the healthcare industry for medical imaging, genomics, drug discovery, diseases predictions, health monitoring, tracking diseases, providing virtual assistance, etc. (Subrahmanya *et al.* 2021).

Data science continues to grow in relevance to organizations. Some of the benefits includes:

1. Retailers employ data science to improve consumer satisfaction and loyalty.
2. For fraud detection and individualized financial advice, data science is widely employed in the banking and finance industries.
3. In the healthcare sector, doctors employ data analysis from wearable trackers to monitor their patients' health and make crucial choices. Hospital administrators can improve service and decrease waiting times by using data science.
4. To get insights, one can use data science to examine vast amounts of graphical, temporal, and geographical data. Additionally, it aids in reservoir characterization and seismic interpretation.
5. Businesses may use social media content to track the consumption trends of media material in real-time thanks to data science. As a result, the businesses may produce content tailored to their target audience, evaluate the effectiveness of their content, and suggest on-demand material.
6. Data science is used to analyze how utilities are used in the energy and utility sectors. (Ohri 2022)

**What is Big Data?**

According to Oracle 2021, Big data is simply defined as a data that has a good volume, velocity, variety, variability, veracity, visualization, and value.

**Volume** – The main characteristic of big data. This signifies how big the data is and it’s measured in Zettabytes, Exabytes and Yottabytes.

**Variety** – This refers to the types of data sources. It can either by structured or unstructured.

**Velocity** – This points to how fast the data can be accessed and processed.

**Variability** – This refers to data that keeps changing constantly.

**Veracity** – This deals mainly with accuracy of data.

**Visualization** – This refers to how data is presented.

**Value** – This signifies how beneficial the data is (Oracle 2021).

Big data has its own general challenges, and some are:

1. Lack of professionals: Lack of big data experts is one of the problems that each company deal with. This frequently occurs because, while data handling tools have advanced quickly, most professionals haven't.
2. Confusion in big data tool selection: Organizations often get confused on the best tool to use. They discover that they frequently make poor choices and choose for the wrong technologies. Money, time, effort, and working hours are lost as a result.
3. Data security: One of the intimidating issues of big Data is protecting these enormous repositories of knowledge. Companies frequently postpone data security to later stages because they are so preoccupied with understanding, storing, and analyzing their data sets.
4. Lack of proper data understanding: Companies' efforts to use Big Data fail due to a lack of knowledge. Employees might not be familiar with the definition, sources, processing, and storage of data (Gaur 2020).

Some of the challenges of big data in the healthcare sector are:

1. Costs of implementation
2. Gathering and cleansing of data
3. Security
4. Issues with management and communication
5. Interoperability (Alkhaldi 2021).

**Fetal Health**

According to National Cancer Institute 2011, a fetus is simply an unborn baby. An unborn baby that grows and develop inside a woman’s womb (uterus). The fetal phase in humans starts eight weeks after an egg is fertilized by a sperm and concludes at birth.

According to National Centre for Health Statistics 2019, The birthrate of babies in the United States were 11.0 per 1000 population from women aged 15 to 44. According to Office for National Statistics 2021, the number of live births in England and Wales was 613,936 which was a significant decrease of 4.1% from the previous year. The fertility rate also reached a low record in 2020 resulting in 1.58 children per women.

A screenshot of a computer

Description automatically generated with medium confidence

***Figure 1: Total fertility rate in England and Wales from 1940 to 2020. (Source: Office for National Statistics, 2021)***

According to Macrotrends 2022, a reliable research platform; The current fertility rate increased by 0.06% which translates to 1.76 births per women. Also, about 3 million new born babies die every year (Piri and Mohapatra 2020).

Some of the signs of fetal distress are:

1. Decreased heart rate of the fetus
2. Different or non-existent fetal movement for a long period of time
3. Low amniotic fluid (Cleveland Clinic 2022).

The fertility rate can further be increased by early diagnosis of a fetal distress which would in turn save the fetus if it is not fatal and the life of the mother.

Pregnancy cardiotocography (CTG) is used to track uterine contractions and fetal heart rate. It aids the early diagnosis of fetal distress and the monitoring of fetal health.

CTGs are a quick and affordable way to evaluate fetal health, giving medical practitioners the information, they need to stop infant and mother death. Fetal heart rate (FHR), fetal movements, uterine contractions, and other factors can be determined by transmitting ultrasound pulses and monitoring the response. It is possible to tell if a pregnancy is high- or low-risk by interpreting the CTG. An unusual CTG may point to the need for more research and possibly treatment (Chandraharan 2017).

This report therefore focuses on analyzing the health of fetuses using analytical tools and data science techniques to bring out dependable insights.

## **Aims and Objectives**

The aim of this coursework is to use data analytical tools and visualization techniques reveal insights about some of the factors that affect a fetus’ health.

The specific objectives are to:

1. Get credible data
2. Prepare the data
3. Perform Exploratory Data Analysis (EDA)
4. Visualize the data and extract insights.

# **Methods**

This section consists of information about the dataset, analytical technique used,

## ***Dataset***

The dataset used in this report was gotten from the UC Irvine Machine Learning Repository popularly known as UCI. It can be accessed at: <http://archive.ics.uci.edu/ml/datasets/Cardiotocography>. This dataset was chosen because based on research, it has all the important features that affects the health of a fetus.

The raw dataset contains 2126 records and 22 columns. The total size of the data is 365.5kb and it has no missing values. The data type is also majorly numerical i.e., no categorical values in the dataset. Figure 2 below shows what the data contains.

A computer screen capture

Description automatically generated with medium confidence

***Figure 2: Contents of raw data***

Table 1 below shows further description about the data. What each variable means and their respective data types.

***Table 1: Dataset Description***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Column Names** | **Description** | **Range** | **Data type** |
| 1 | ‘baseline value’ | The fetus heartbeat per minute | 106 – 160 | Ratio |
| 2 | ‘accelerations’ | Heart accelerations per second | 0 – 0.019 | Ratio |
| 3 | ‘fetal\_movement’ | Fetus movements per second | 0 – 0.481 | Ratio |
| 4 | ‘uterine\_contractions’ | Uterine contractions per second | 0 – 0.015 | Ratio |
| 5 | ‘light\_decelerations’ | Light decelerations per second | 0 – 0.015 | Ratio |
| 6 | ‘severe\_decelerations’ | Severe decelerations per second | 0 – 0.001 | Ratio |
| 7 | ‘prolonged\_decelerations’ | Prolonged decelerations per second | 0 – 0.005 | Ratio |
| 8 | ‘abnormal\_short\_term\_variability’ | Percentage of time with an unusual short-term unevenness. | 12 - 87 | Ratio |
| 9 | ‘mean\_value\_of\_short\_term\_variability’ | Mean value of short-term unevenness. | 0.2 - 7 | Ratio |
| 10 | ‘percentage\_of\_time\_with\_abnormal\_  long\_term\_variability’ | Percentage of time with an unusual long-term unevenness. | 0 – 91 | Ratio |
| 11 | ‘mean\_value\_of\_long\_term\_variability’ | Mean value of long-term unevenness. | 0 – 50.7 | Ratio |
| 12 | ‘histogram\_width’ | Width of the fetus heart rate (FHR) | 3 - 180 | Ratio |
| 13 | ‘histogram\_min’ | Lowest frequency of the FHR | 50 -159 | Ratio |
| 14 | ‘histogram\_max’ | Highest frequency of the FHR | 122 - 238 | Ratio |
| 15 | ‘histogram\_number\_of\_peaks’ | Amount of histogram peaks | 0 - 18 | Ratio |
| 16 | ‘histogram\_number\_of\_zeroes’ | Amount of histogram zeros | 0 - 10 | Ratio |
| 17 | ‘histogram\_mode’ | Mode of the histogram | 60 - 187 | Ratio |
| 18 | ‘histogram\_mean’ | Mean of the histogram | 73 - 182 | Ratio |
| 19 | ‘histogram\_median’ | Median of the histogram | 77 - 148 | Ratio |
| 20 | ‘histogram\_variance’ | Variance of the histogram | 0 - 24 | Ratio |
| 21 | ‘histogram\_tendency’ | Tendency of the histogram | -1 - 1 | Ratio |
| 22 | ‘fetal\_health’ (Target Variable) | 1 – Normal  2 – Suspect  3 - Pathological | 1 - 3 | Ordinal |

## **Analysis and Results**

### **Dataset Preparation**

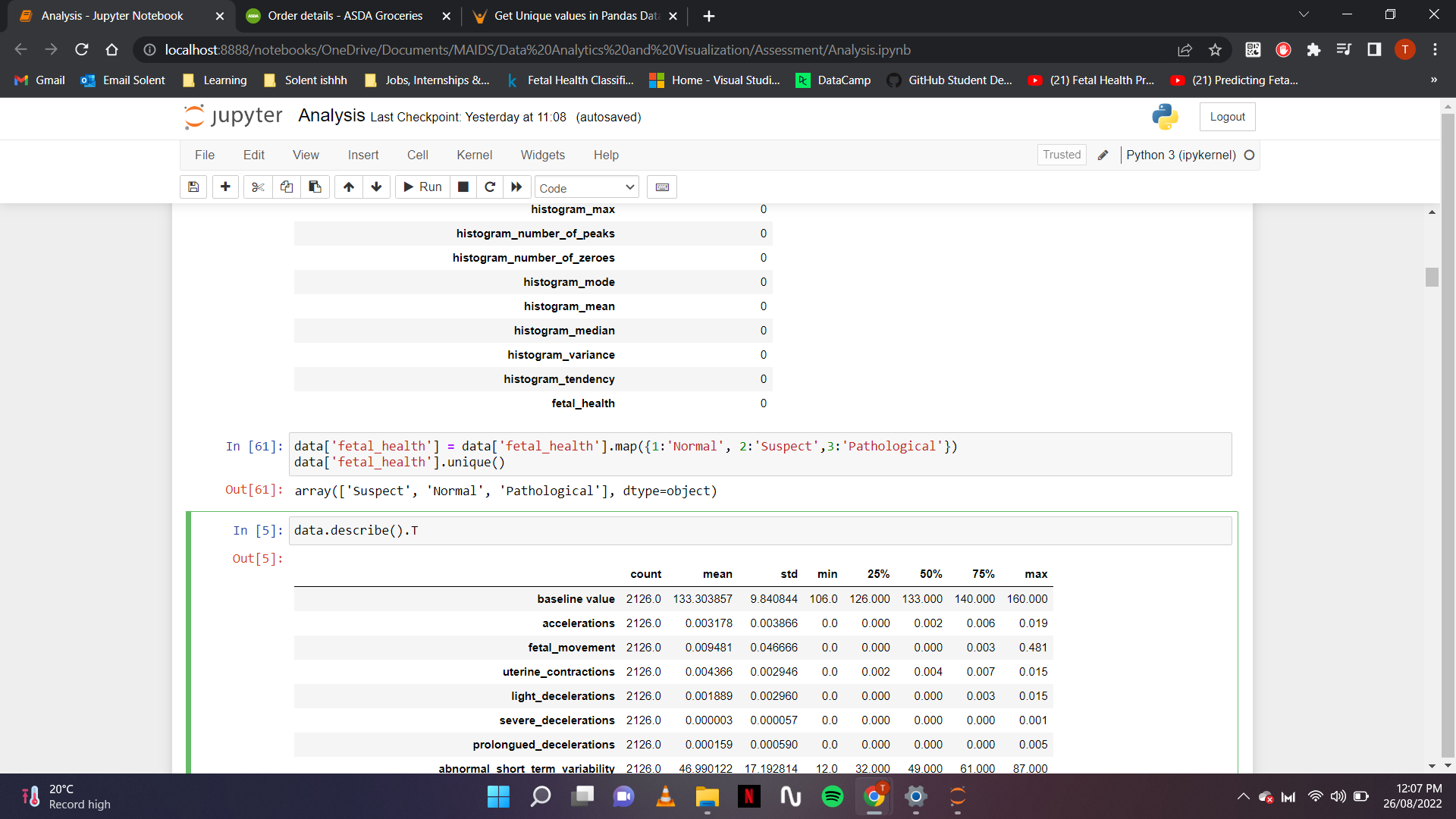
The dataset needed almost no preparation because there were no missing values as shown in figure 3 below.

A screenshot of a computer

Description automatically generated

***Figure 3: Number of missing values***

The target variable was renamed and mapped from its original values (1,2,3) to (Normal, Suspect and Pathological) respectively for better understanding during the analysis. Figure 4 below shows how the target variable was mapped and transformed.



***Figure 4: Target variable mapping***

### **Exploratory Data Analysis (EDA)**

Variety of python libraries were used to achieve the analysis in this report. Primarily EDA is for seeing what the data can tell us.

Since there are no missing values, the balance of the dataset was examined, and it was discovered that the dataset is imbalanced.

Figure 5 below shows a bar chart showing the imbalance.

Graphical user interface

Description automatically generated with medium confidence

***Figure 5: Bar chart showing imbalance***

According to figure 5, the bar chart shows that there are more healthy fetuses than pathological ones which expresses the imbalance. This means that if a machine learning model is used to predict the health of a fetus using this dataset, it is very likely to predict its normal even though it is not normal or it suspects an abnormality.

Graphical user interface, application

Description automatically generated

***Figure 6: Pie chart showing imbalance***

Figure 6 above shows a pie chart displaying the percentage of each of the values present in the target variable.

Figure 7 below shows a univariate analysis of each of the variables present in the dataset. It consists of different histograms illustrating the distribution of all the variables in the dataset. From the figure below, none of the variables are negatively skewed (left skewed). A few of them are normally distributed and positively skewed (right skewed). The variables that are normally distributed are ‘baseline value’, ‘histogram\_mode’, ‘histogram\_mean’, ‘histogram\_median’ and ‘histogram\_max’.

The variables that are positively skewed are ‘acceleration’, ‘light\_deceleration’, ‘mean\_value\_of\_short\_term\_variability’, ‘mean\_value\_of\_long\_term\_variability’ and ‘histogram\_number\_of\_peaks’.

A computer screen capture

Description automatically generated with medium confidence

***Figure 7: Histogram showing distributions***

Figure 8 below is a multivariate analysis using a heatmap to express how the variables in the dataset are correlated with each other. In this figure, ‘histogram\_mean’, ‘histogram\_median’ and ‘histogram\_mode’ are highly correlated with each other. The variables that are highly correlated with the target variable are:

* 1. ‘prolongued\_decelerations’: 0.48,
  2. ‘abnormal\_short\_term\_variability’: 0.47
  3. ‘percentage\_of\_time\_with\_abnormal\_long\_term\_variability: 0.43.

Graphical user interface, chart, treemap chart

Description automatically generated

***Figure 8: Heatmap showing correlations***

Figure 9 below shows boxplots for all the variables in the dataset. The purpose of this is to check for the presence of outliers and from the result, the dataset contains some outliers. One way to deal with outliers would be to drop the variables containing the outliers, but they wouldn’t be dropped in this task because this report focuses primarily on analysis of the dataset.

A screenshot of a computer

Description automatically generated

***Figure 9: Boxplots showing outliers***

Figure 10 below shows how the target variable is clustered. K-means clustering was performed, and the elbow method was used to achieve the optimal number of clusters. According to the figure, the target variables are densely clustered.

A screenshot of a computer

Description automatically generated with medium confidence

***Figure 10: Cluster analysis***

Figure 11 below shows the percentage of outliers contained in each of the variables. The variable with the highest outlier percentage is ‘histogram\_number\_of\_variables’ with almost 24% of outliers in it, followed by the ‘percentage\_of\_time\_abnormal\_long\_term\_variability’ with almost 15%.

Text

Description automatically generated

***Figure 11: Percentage of outliers***

### **Data Modelling and Visualisation**

The independent variables in this dataset are 'baseline value', 'accelerations', 'fetal\_movement', 'uterine\_contractions', 'light\_decelerations', 'severe\_decelerations', 'prolongued\_decelerations', 'abnormal\_short\_term\_variability', 'mean\_value\_of\_short\_term\_variability', 'percentage\_of\_time\_with\_abnormal\_long\_term\_variability', 'mean\_value\_of\_long\_term\_variability', 'histogram\_width', 'histogram\_min', 'histogram\_max', 'histogram\_number\_of\_peaks', 'histogram\_number\_of\_zeroes', 'histogram\_mode', 'histogram\_mean', 'histogram median', 'histogram\_variance', 'histogram\_tendency'.

And the variable to be predicted (target variable) is ‘fetal\_health’.

Figure 12 below shows the importance of the variables in the dataset and from the figure, the two most important variables are ‘histogram\_mean’ and ‘mean\_value\_of\_short\_term\_variability’. The rest are: ‘abnormal\_short\_term\_variability’, ‘percentage\_of\_time\_with\_abnormal\_short\_term\_variability’, ’prolongued\_decelerations’, ‘histogram\_max’, ‘baseline\_value’, ‘uterine\_contractions’, ‘histogram\_mode’, ‘accelerations’.

Chart

Description automatically generated

***Figure 12: Feature importance***

Figure 13 below shows the different health statuses of a fetus based on uterine contractions. Uterine contractions refer to the uterine muscles shortening and tightening. It is measured per second. According to figure 11, any value > 0.0025/s is normal while any value < 0.0025/s is a suspect.

Graphical user interface, application

Description automatically generated

***Figure 13: Fetal health based on uterine contractions***

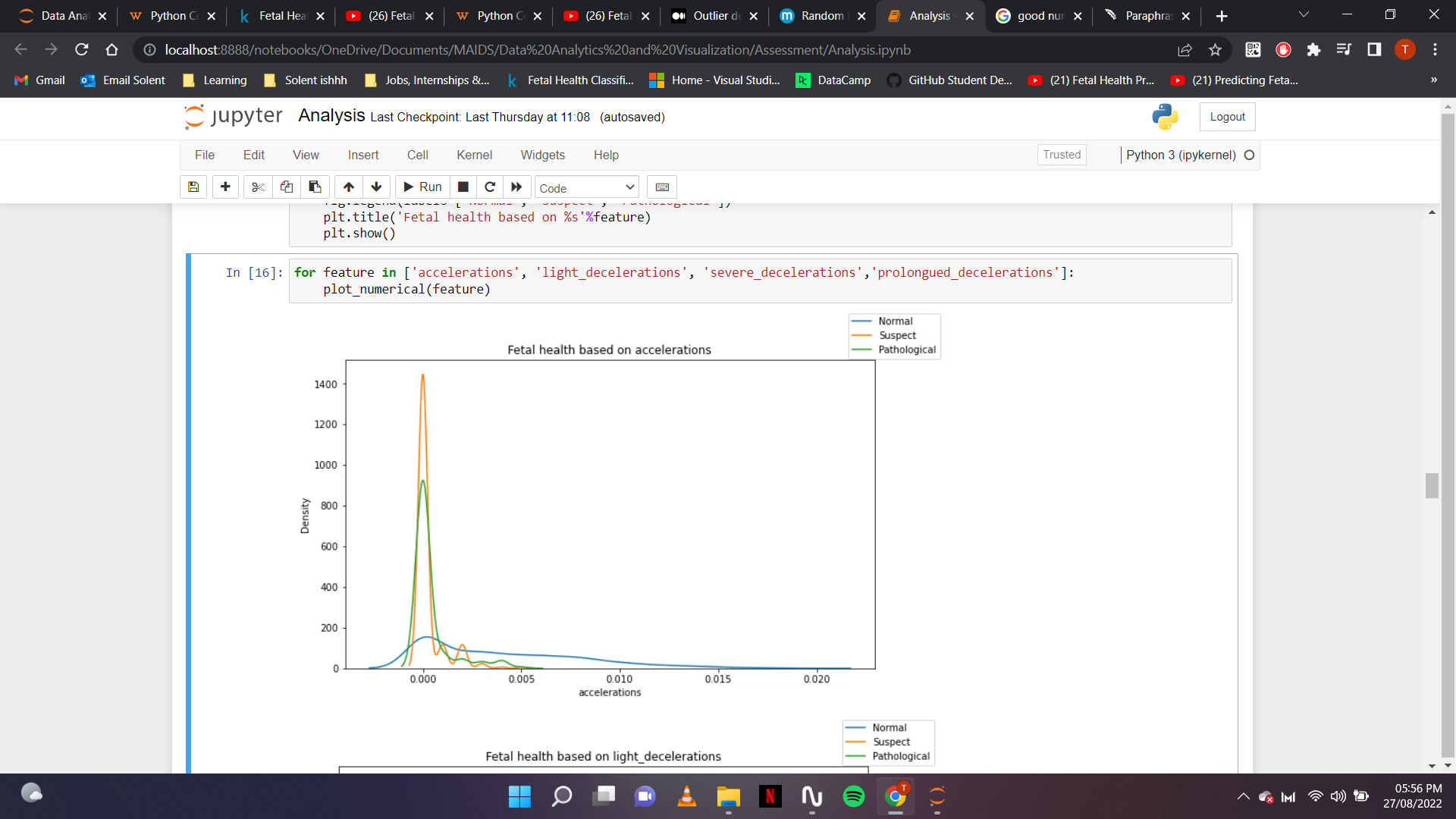
Figure 14 below shows the different health statuses of a fetus based on the baseline value. Baseline value refers to the number of heart beats per minute. According to the dataset, 125beats/m – 135bpm is pathological and any value > 135bpm is considered a suspect.

A screenshot of a computer

Description automatically generated

***Figure 14: Fetal health based on baseline value***

Figure 15 below shows the different health statuses of a fetus based on the heart accelerations per second. According to the dataset, any value > 0.0025/s is normal.



***Figure 15: Fetal health based on accelerations***

Figure 16 below shows the different health statuses of a fetus based on abnormal short-term variability. It refers to the percentage of time with an unusual short-term unevenness. According to the dataset, any value < 50 is normal and any value > 50 is pathological.

Graphical user interface, chart, histogram

Description automatically generated

***Figure 16: Fetal health based on abnormal short-term variability***

Figure 17 below shows the different health statuses of a fetus based on the mean value of short-term variability. It refers to the mean value of short-term unevenness. According to the dataset, any value < 1 is a suspect, 1-2 is normal and any value > 2 is pathological.

Graphical user interface, application

Description automatically generated

***Figure 17: Fetal health based on mean value of short-term variability***

Figure 18 below shows the different health statuses of a fetus based on the percentage of time with abnormal long-term variability. It refers to the percentage of time with an unusual long-term unevenness. According to the dataset, any value between 10 – 70 is a suspect.

Graphical user interface, application

Description automatically generated

***Figure 18: Fetal health based on percentage of time with abnormal long-term variability***

Figure 19 below shows the different health statuses of a fetus based on the mean value of long-term variability. It refers to the mean value of long-term unevenness. According to the dataset, any value < 3 is pathological, any value between 3-10 is a suspect and any value > 10 is normal.

A screenshot of a computer

Description automatically generated

***Figure 19: Fetal health based on mean value of long-term variability***

The link to the complete analysis can be found here:

# **Evaluation**

Alam *et al.* (2022) worked on research covering findings and analysis of multiple machine learning models for fetal health classification. Although the data was examined in the research, it contained very little analysis and a lot more prediction. Out of the Machine learning (ML) models used for prediction, the random forest classifier produced the best result in terms of accuracy.

Mehbodniya *et al.* (2021) article focused on evaluating factors that various CTG-measured parameters have in predicting a fetus' health through ML models. In this article, a few analysis were done. Histplots, regression analysis and correlation analysis were created. The random forest classifier also came out as the best in terms of accuracy, precision, recall and F1-score.

This report focused mainly on analysis and the key findings in this analysis are:

1. Most important features visualized in fig. 12
2. Correlation analysis in fig. 8 and
3. The findings from fig. 13 to 19.

The finding from this report would be very useful to obstetricians now and in posterity. The obstetrician’s visual assessment of the CTG data may not always be correct as they may identify prenatal defects and choose whether to intervene medically before the fetus suffers lasting injury. Data analysis and ML techniques can help to properly identify the future health status on time and in turn save lives, thus increasing the rate of fertility.

## **Limitations and Challenges**

This report is not 100% perfect, it has a few limitations and can be further improved upon. The balance of the data and the cluster analysis can be improved upon. The main challenge faced during this report is time. I would have loved to explore other analytical tools like power BI and be conversant with Excel but there was little to time to learn and still get the report done.

However, further work can be done by making predictions and applying machine learning techniques.

# **Conclusion**

In conclusion, pregnancy-related health issues are a big problem that people throughout the world must deal with. The obstetrician ability to make quick judgments to save the lives of both the mother and the fetus will be aided by this research.

I enjoyed working on this report and while working on this research, I learnt new analytical techniques and got more acquainted with Jupyter notebook.

# **Reference List (Harvard Style)**

# **Appendix**

Add as many screenshots, code snippets, visualisation here even if you have added these in the report. Do not forget to add a short video presentation (up to 7 min long)