Introduction to Machine Learning and Data Mining

Panagiotis Papadamos - s205637
Dimitrios Loukas - s220514
Ziyu Zhang - s212657

Introduction and Data Description

Data Description

The data set consists of 1014 observations (pregnant women), each of which has 7 attributes, namely: Age, Systolic Blood Pressure, Diastolic Blood Pressure, Blood Sugar, Body Temperature, Heart Rate and Risk Level, which are the responsible and significant risk factors for maternal mortality. In this project we get a basic understanding of our data by feature extraction and visualization in order to do further analysis, which is predicting the risk level of maternal mortality from the first six attributes.

References

Marzia Ahmed
Email: ahmed.marzia32'@'gmail.com
Institution: Daffodil International University, Dhaka, Bangladesh

Summary of previous analysis

The present dataset, "Maternal Health Risk Data Set Data Set", was created by Marzia Ahmed and Mohammod Abul Kashem in a research regarding a risk prediction model for maternal health. The results were published by IEEE in the paper with title *IoT Based Risk Level Prediction Model For Maternal Health Care In The Context Of Bangladesh.*

In this paper were used analytical tools, as well as machine learning algorithms in order to study the risk level on the basis of risk factors in pregnancy. The data collected in this research came from various sources (IoT device and Hospitals and maternity clinics in Dhaka and Bangladesh). This research consists mainly of three parts. First one is the data collection through the IoT sensors of the device (temperature sensors, heart rate sensors etc), second one is the data processing and last is the transfer of the data to the cloud for further analysis.

This paper proposes a maternal healthcare model for observing a pregnant woman, as well as fetal health. More in depth, this model consists of four continuous processes:

1. The health data of the patients are collected from the various sources, which in this case are Hospitals and maternity clinics in Dhaka and Bangladesh
2. The collected data are stored in the local server and the cloud server.It is worth noticing that the data were analyzed by ranking system, meaning that high risk factors are ranking first, while mid risk factors are ranking lower. Three categorial risks have been considered, low risk, mid risk and high risk, more specifically the total data size is 1014 of which 406 low risk, 336 mid risk and 272 high risk. The most probability risk factor according to their results is **BS(Blood Sugar)**, especially for a mother affected by diabetes which is three time higher than **BP(Blood Pressure)** and other factors. BP is also a significant risk factor.
3. The stored data were classified and predicted using a machine learning algorithm according to the trained data. After implementing various groups of machine learning algorithms (Decision Tree, Random Forest, Support Vector Machine, Sequential Minimal Optimization, Logistic Regression, Naïve Bayes, Lbk, Logistic Model and others), in both Weka and Python, the **Decision Tree** algorithm was chosen giving the highest accuracy of 97%. Furthermore in pursuance of

the best prediction, the accuracy in respect to the maximum tree depth was studied and was observed that depth 8 gave the maximum accuracy. A hyperparameter tuning method was used to tune the classifier in order to obtain the best parameters. Also data mining and statistical approaches have been done in order to find significant factors such as Chi-squared or Gain ratio. After preprocessing the data, all the risk factors were encoded in an object type array by Implementing sklearn and then the risk level as the class for the prediction. Then the dataset was split to 70% training set and 30% test data. The model was trained by the Decision Tree algorithm.

4. Finally the predicted result were sent to both the sources (Hospital and emergency services related with) and to the patient.

To summarize the **Decision Tree** algorithm provided the highest accuracy for classification and prediction of the risk among other 15 machine learning algorithms. The most significant risk factor for maternal health was found to be **BS(Blood Sugar)**, followed by **BP(Blood Pressure)**.

Goals

We are given the records of six attributes for a pregnant woman and have to determine what is the risk level of maternal mortality. This is a multi-class classification problem since there are three categories to choose from: low level, mid level and high level. To get a understanding of data, we firstly transform data to a more available value. For example, we change the risk level from string to ordinal value. After that we do feature Extraction, which aims to reduce the number of features in a data set by creating new features from the existing ones. Then visualize data with Principal component analysis(PCA).

Attribute Explanation

Attributes Type

The chosen dataset contains 7 attributes (features), namely **age**, **systolic blood pressure**, **diastolic blood pressure**, **blood glucose**, **body temperature**, **heart rate** and **risk level**. The 6 of them will be later used as independent variables in order to predict the 7th one (risk level). In this subsection we will describe the type of each attribute.

We will start by categorizing the attributes as discrete and continuous. The 5 out of 7 attributes are discrete, namely **age**, **systolic blood pressure**, **diastolic blood pressure**, **body temperature**, and **risk level**). We should note that **systolic blood pressure** and **diastolic blood pressure** are measured in mmHg, but their values are discrete integers numbers. The **blood glucose** is a continuous attribute as it can take any real value between two numbers. Finally **heart rate** is categorized as continuous. Although it seems to be discrete based on its value, there is an observation that has **heart rate** value equal to 98,4. This value can be a mistake, but as long as we don't have evidence to conclude this we categorize the attribute **heart rate** as continuous (the distance between to values of the attribute is not standard).

We will now categorize the attributes as ordinal, nominal, interval or ratio. The attributes **age**, **systolic blood pressure**,**diastolic blood pressure**, **blood glucose**, **body temperature** and **heart rate** are ratio as their absence (value 0) has a physical meaning. So, for example, regarding blood glucose levels, an observation (patient) with a value 0 would mean the physical absence of glucose in the blood of the specific patient. The last attribute (**risk level**) is ordinal. Even though in the dataset is given as a string, one can order the values of the attribute, based

Table 1: Statistics Summary

Statistics	Age	Systolic BP	Diastolic BP	Blood Gl.	Body Temp.	Heart Rate	Risk Level
Mean	29.87	113.2	76.46	8.73	98.67	74.3	1.87
Variance	181.38	338.37	192.63	10.84	1.88	65.36	0.65
STD	13.47	18.39	13.88	3.29	1.37	8.08	0.81
Range	60	90	51	13	5	83	2
Min	10	70	49	6	98	7	1
Max	70	160	100	19	103	90	3

on the magnitude of the risk level (low, mid or high). For our computations we will transform this values to "low risk"=1, "mid risk"=2 and "high risk"=3 in order to make them numerical and use on-out-of-K encoding, if need be, to perform the principal component analysis.

Data Issues

We checked the dataset for missing or corrupted values using Python (can be found in our code). In order to do so, we check every value of the dataset in order to see if it's an integer or a float number (after we transformed the 7th attribute) and used a counter to keep track of the missing values. This check didn't concern cases that missing values might have already been altered to some arbitrary one. There are no missing or corrupted values in the dataset and so we didn't have to take measures and deal with them (like replacing them or discard the values or the related observations). In this stage we hadn't checked the dataset for outliers, something that would be easier when the Principal Component Analysis will be implemented.

Statistics

The Table 1 shows the basic summary of statistics of the chosen dataset. For the sake of completeness we have included the values also for the last variable, namely **risk level** as we transformed its values to numerical. Although it doesn't make much sense to compute the mean of the last attribute we do so, but we also compute the mode (equal to 1) that can give more insights regarding our dataset. The summary of statistics of course cannot substitute prediction models, but can probably give us a first idea of what to expect. For example, we can see that the mean of the **age** attribute is low, as well as the mode of the predicted attribute. This could imply a relation between **age** and the **risk level**, something that we cannot conclude only using statistics. The statistics also help to have a deeper idea of our dateaset and ususally detect outlier. For example, we can see that the minimum value of **heart rate** is 7 and we can be confident that this value is an outlier (even the world record is 27bits per min). We can handle the outlier in different ways. For the two observations in our dataset with **heart rate** value equal to 7 we chose to change the value with the mean of the heart rate value (74 instead of 74.3 in order to keep the attribute discrete). Note that in Table 1 we show the summary of the statistics before the changes.

Principal Component Analysis

Initial Visualization

We firstly chose to visualize our data attributes against each other in pairs in order to see if we can have an initial idea about relations between attributes or even clear separation of the classes

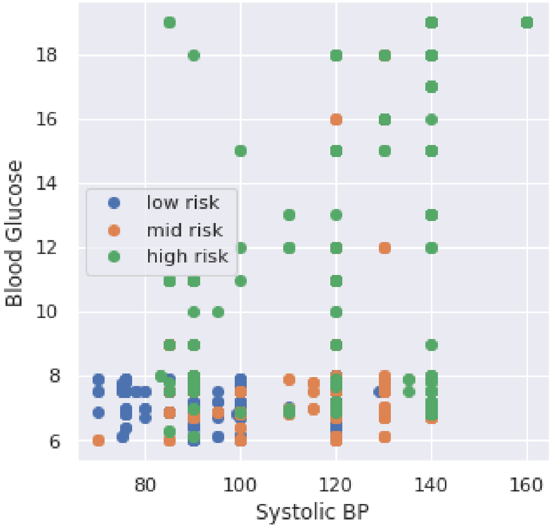


Figure 1: *Systolic Blood Pressure against Blood Glucose*

in the data visualization. Our dataset consists of many observations with same values for specific attributes so it was difficult to get some informative results by just visualizing the data. In 6 we show the visualization of **Systolic Blood Pressure** against **Blood Glucose**. In the graph we can see a rudimentary separation of the classes (left bottom 'low risk', middle bottom 'mid risk' and middle and top 'high risk'). The rest visualizations can be found in the appendix. In the next subsections we use PCA in order to visualise our data in a 2-dimensional space and have more informative results.

Outliers

In data analysis, the identification of outliers is very important. An outlier requires special attention. It may be the result of a measurement or recording error, an observation from a different population, or an unusual extreme observation. Boxplots are a way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). This type of plot is used to easily detect outliers. Figure 2 shows the age distribution of different risk groups. Observations that fall outside bounds are marked as hollow circles and they are outliers. Since we observe an outlier, we try to determine its cause. There are no corrupted values in the dataset, it is not legitimate to simply drop the outlier, so we remained them. We can also run the analysis both with and without it, and state how the results changed. In the report we included the boxplots for the attribute **age**, but in the appendix one can find more boxplots (Running the code can give all of them).

Distribution of Variables

We need to know how the data are distributed to determine the most appropriate statistical analyses to use. The normal distribution is the most important probability distribution in statistics. Characteristics that are the sum of many independent processes frequently follow normal distributions, like heights, blood pressure, measurement error, etc. Here we describe the distribution of age using a histogram and thus the y-axis doesn't describe number of observations,

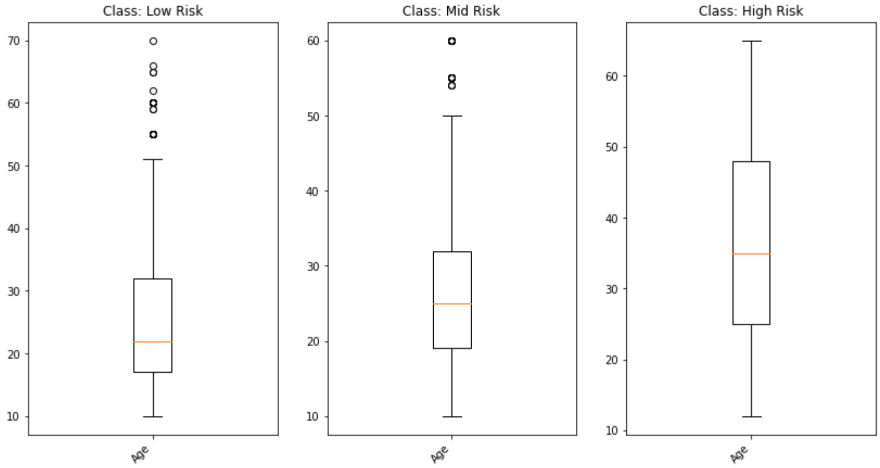


Figure 2: *Age Boxplot*

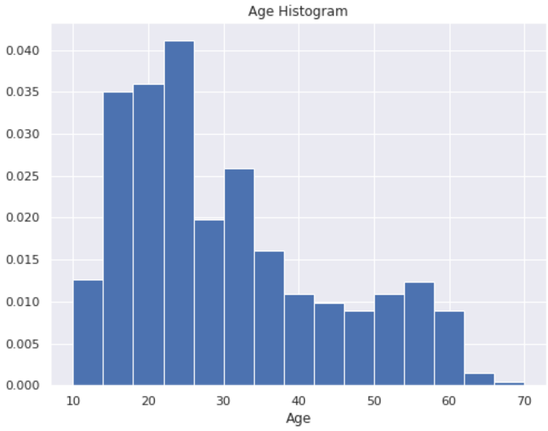


Figure 3: *Age Distribution*

but probabilities instead. Figure 3 shows age distribution is a right-skewed distribution, which is difficult to assign a typical value as there is no clear center point on a right-skewed graph. By contrast, with normal distribution, because the mean, median, and mode are all equal and come at the center of the data set, we can easily use this value to more accurately generalize the data. The Figure 4 shows that the attribute **Heart Rate** follows normal distribution. The mean value and the variance can be found in 1. In the appendix one can find more histograms.

Correlation

Correlation tells us how related attributes are. Figure 4 is a correlation matrix which shows correlation coefficients between variables. Each cell in the table shows the correlation between two variables, and light colors indicate higher correlation. In the table, we can see that the first six attributes are all positively correlated with the risk level, and the correlation coefficient between blood pressure and risk level is the highest.

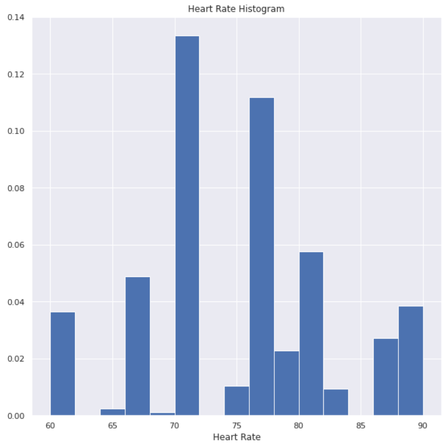
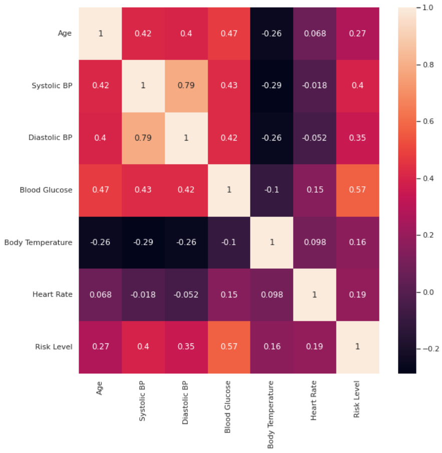


Figure 4: *Heart Rate Distribution*



	Age	Systolic BP	Diastolic BP	Blood Glucose	Body Temperature	Heart Rate	Risk Level
Age	1	0.42	0.4	0.47	-0.26	0.068	0.27
Systolic BP	0.42	1	0.79	0.43	-0.29	-0.018	0.4
Diastolic BP	0.4	0.79	1	0.42	-0.26	-0.052	0.35
Blood Glucose	0.47	0.43	0.42	1	-0.1	0.15	0.57
Body Temperature	-0.26	-0.29	-0.26	-0.1	1	0.098	0.16
Heart Rate	0.068	-0.018	-0.052	0.15	0.098	1	0.19
Risk Level	0.27	0.4	0.35	0.57	0.16	0.19	1

Figure 5: *Correlation Matrix*

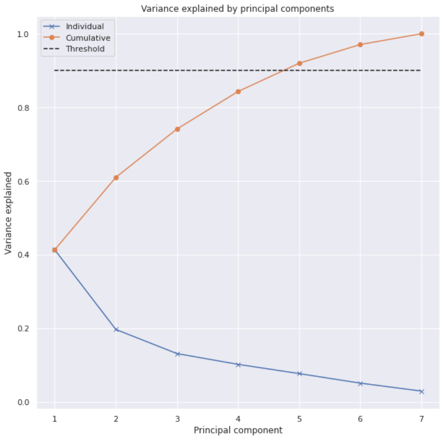


Figure 6: *Variance explained by principal components*

Variance explained by principal components

To reduce the dimensionality of the data we apply PCA. We first standardize by subtracting the mean, and include the step where we divide by the standard deviation since the ratings have substantially different scale. After carrying out an SVD, we obtain the matrices U, Σ, and V. From PCA we can find out how much of the variation in the data each PCA component accounts for. We plot the percent of variance explained by the principal components as well as the cumulative variance explained. Figure 5 shows more than 90% of the variation in the data is explained by the first 5 principal components.

Projection on Principal Components

The principal component analysis gave us better visualization of the dataset. By construction, the first principal component explains most of the data's variance, the second principal component explains the second most and so on. That means, essentially that the direction of the principal components is the one that is shown in the figure. The first principal components is the x-axis of the plot and the second principal component is the y-axis of the plot, because as we see the data are more spread in the x-axis than the y-axis. The projection onto the plain create by the two first principal components gives also a good visualization of the differentiation of the classes, a fact that gives us evidence that classification is feasible. We could even try to classify a new observation by calculating the coordinates and project it onto the plain created by the components. Then, based on where the observation lies we could classify it using the visual results, but classification is something we will achieve using models.

In terms of the attributes, we can describe a principal component based on which attributes is correlated with. We can use different thresholds for the level of correlation, but in our case a threshold of 0.4 could give us an idea of what the first two principal components describe. More specifically we can see that PC1 is correlated with **Blood Glucose** and **Body Temperature**. That means that observation with high value of **Blood Glucose** and **Body Temperature** will have high values on the PC1 (based on the values on the PC1). For the second principal component we can see it describes the attributes **Age**, **Blood Glucose** and **Risk Level**. With the same way we could reason for all the principal components.

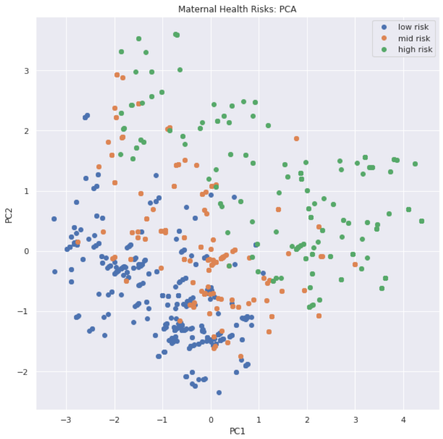


Figure 7: *Data projected onto the two first principal components*

Discussion

In this project we assimilated various information regarding the pre-processing of the data, the visualization, as well as the interpretation of it. We valued the importancy of **standardization** of the data by comparing principal component results though plots and graphs. We observed that putting variables on the same scale helped us understand and visualize better our results. We discovered the significance of the **visualization** of the data and how important its role is, in terms of better understanding our data, as well as in terms of communicating with each other and other researchers. Furthermore we deduced major detail points in the process of formatting our data that require attention and can potentially lead to faulty results, such as **outliers** or **missing values** from our dataset. Moreover we established though graphs that the more principal components we take into account the better we can explain the variable.

Finally we observed that our primary machine learning aim appears to be feasible based on our visualizations and results. More specifically, in the graph *PC1 vs PC2* we can clearly see that there are distinguished clusters for each risk factor, which means that classification is possible. Lastly if we consider the correlation matrix we can see that there are attributes that are highly correlated, such as Blood Glucose and Risk Level, which is a positive sign that a linear regression is feasible.

Exam Questions

Question 1

Time of day and *y(Congestion level)* are both ordinal variables, because there is a clear ordering in the values. The $x_2 - x_7$ attributes are all ratio because they have all the properties of an interval variable, as well as a definition of zero. The correct answer is **C**.

Question 2

In order to calculate the p-norm distance we use the *Equation (4.17)* (for $1 \leq p < \infty$) and the *Equation (4.18)* (for $p = \infty$), which are stated in *page 58* in the book *Introduction to Machine Learning and Data Mining.* The correct answer is **A**.

Question 3

The variance explained by each principal component is given by *Equation (3.18)*. As such we find *VarExp*PC1—4 = 86.7%, so the answer is **A**.

Question 4

Time of day has negative coefficients of PCA2 whereas **Broken Truck**, **Accident victim**, **Immobilized bus** and **Defects** have positive coefficients resulting in a positive projection onto the first principal component. The answer is **D**

Question 5

The correct answer is A. We used the type for Jaccard similarity $J(s_1, s_2) = \frac{f_{11}}{f_{11}+f_{10}+f_{01}}$. From the given texts we calculated the terms of the type. We had in mind that no stopwords and stemming is used. The calculation gave us 0.1538 and so we chose A.

Question 6

Table 2: Work distribution

Section	**Papadamos (s205637)**	**Loukas (s220514)**	**Zhang (s212657)**
Data Description			✓
Summary of previous Analysis		✓	
Goals			✓
Attribute Types	✓		
Data Issues	✓		
Statistics Summary	✓		
Initial Visualization	✓	✓	
Outliers			✓
Distribution of Variables			✓
Correlation			✓
Variance Explained			✓
PCA	✓	✓	
Discusion		✓	
Coding	✓	✓	
Exam Question 1		✓	
Exam Question 2		✓	
Exam Question 3			✓
Exam Question 4			✓
Exam Question 5	✓		
Exam Question 6			

Appendix

In the appendix we added plots that we couldn't add in the main report due to the report's size. Figure 10 shows all the attributes plotted against each other.

We also added the boxplots of **Diastolic BP** and **Blood Glucose**. The rest can be found by running the code. Finally, in this appendix one can find the histograms that weren't included in the main report.

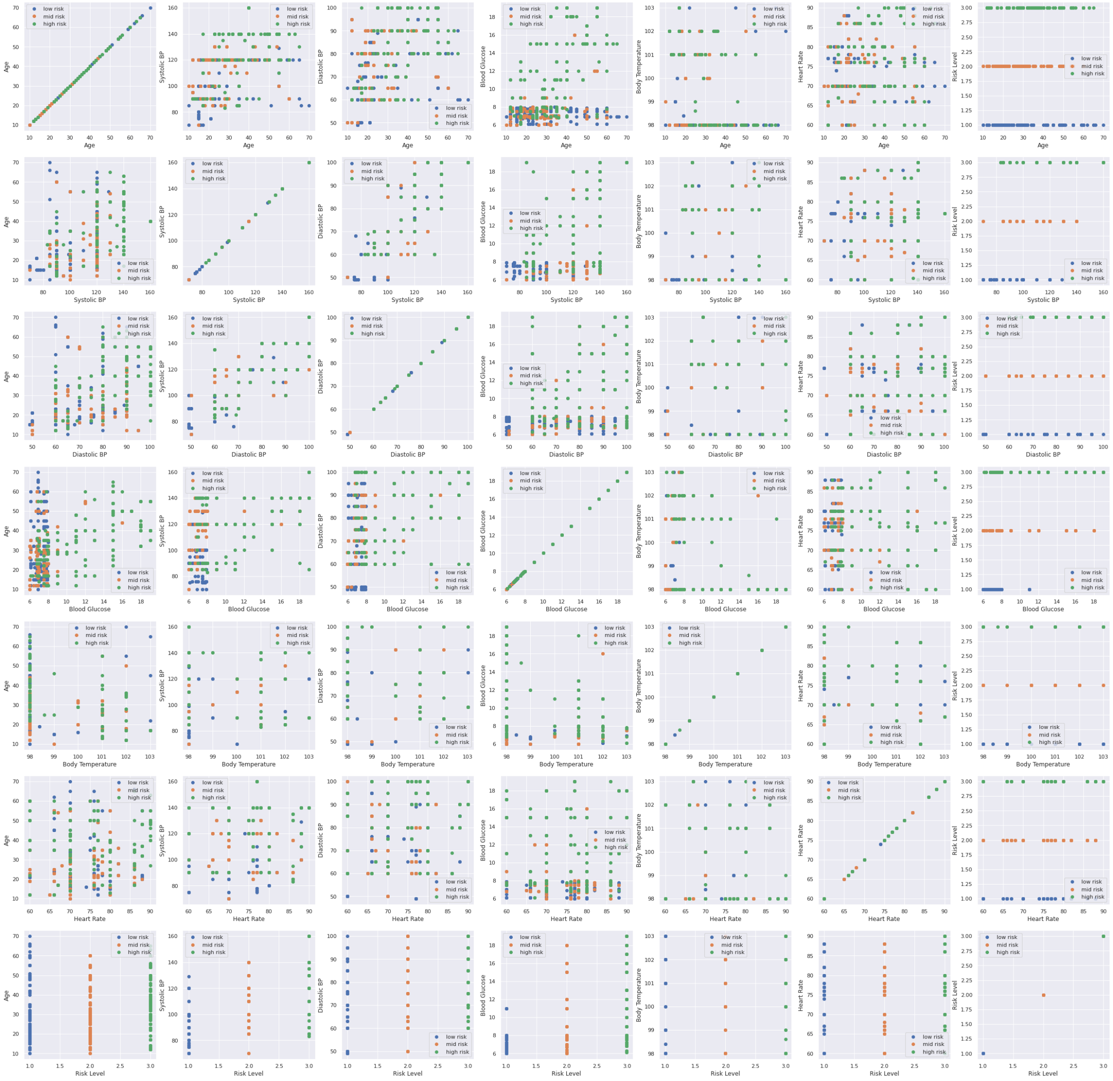


Figure 8: *Attributes projected against each other*

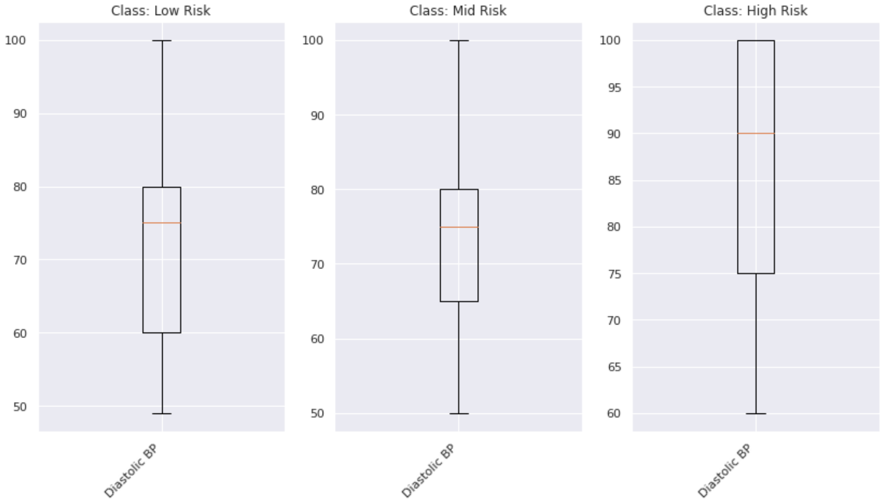


Figure 9: *Attributes projected against each other*

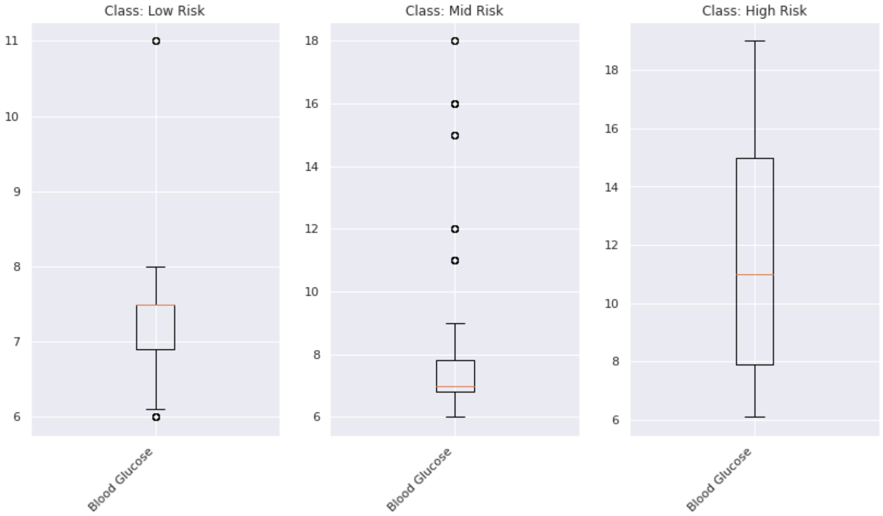


Figure 10: *Attributes projected against each other*

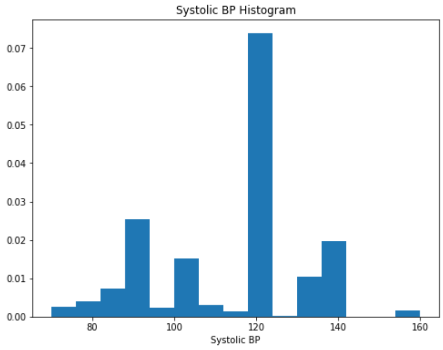


Figure 11: *Systolic BP Histogram*

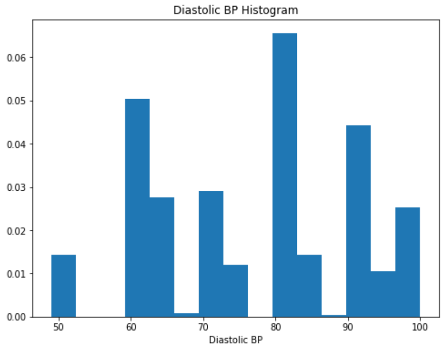


Figure 12: *Diastolic BP Histogram*

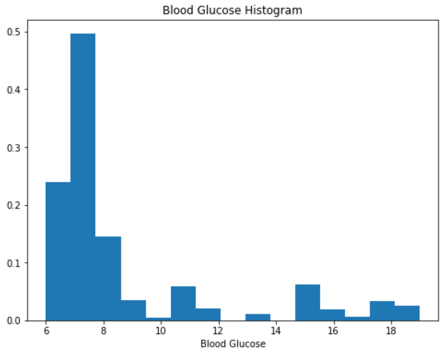


Figure 13: *Blood Glucose Histogram*

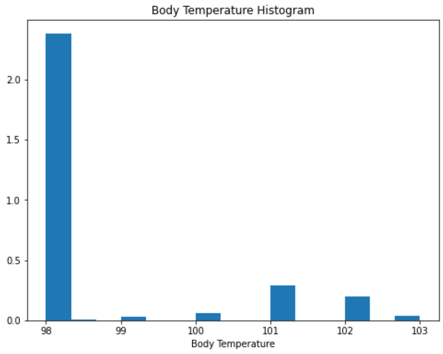


Figure 14: *Body Temperature Histogram*

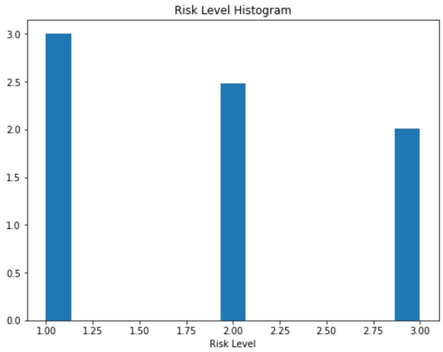


Figure 15: *Risk Level Histogram*