Assessing the impact of key health indicators on Maternal Mortality

Alexandre Silva **- 90004**
Joana Ogura **- 102114**
Pedro Zenário **- 102348**
Ricardo Quintas **- 102154**
Ricardo Simões **- 93674**

Professor:
Maria do Rosário De Oliveira Silva

12 June 2022

Instituto Superior Técnico

Abstract - Scientific applications to healthcare are growing each year within the *IoT* paradigm, and such technologies are capable of measure, store and communicate health indicators of people. Several studies have been conducted in order to assess what indicators can identify problems in pregnancy and post-pregnancy and diminish maternal mortality, especially in developing countries. Therefore, to ascertain which indicators can have meaningful information about this and if they can really predict maternal risk, data was collected from a rural area in Bangladesh. After applying several statistical methods, such as hypothesis tests and machine learning techniques, the results of the study concluded that looking at *Blood Sugar* is of the utmost importance and *SystolicBP* and *Body Temperature* are the other most impactful features, depending on the model at hand. Either way, simple models tend to predict fairly good, with accuracies of 80% or higher.

Keywords: IoT, Robust PCA, PCA, Hypothesis Testing, Entropy, Ordered Logistic Regression, Ordered Probit Regression

1 INTRODUCTION

Maternal mortality is one of the most prominent and urgent problems in today's medicine. The *World Health Organization (WHO)* has identified 17 issues to overcome until the year of 2030 - the *Sustainable Development Goals (SDG)*. One of them, *"Good Health and Well-Being"*, intends to ensure universal health coverage and access to quality health care for people of all ages ([2], [6]). One of the main issues in this domain is maternal mortality. In particular, this complication tends to exist in low income countries, where the proportion of mothers that do not survive childbirth compared to those who do is 14 times higher than in developed countries. For this and many other reasons, it is extremely urgent to propose measures that can potentially minimize these numbers. Fortunately, there have been recent developments in technology applied to healthcare that are positively impacting both the patients and the medical community. In this project, we will analyze data collected within the *Internet of Things (IoT)* based risk monitoring system. The system pulled this data ([5]) from different hospitals, community clinics and maternal health cares from the rural areas of Bangladesh, which is considered a developing country.

2 PRELIMINARY DATA ANALYSIS

2.1 Data Description and Treatment of Outliers

The dataset that we will analyze contains data retrieved from 1014 patients related to health indicators of people from the region of Bangladesh. The descriptive variables are the age of the women in question (*Age*), their systolic blood pressure (*SystolicBP*) and diastolic blood pressure (*DiastolicBP*),

their blood sugar level (*BS*), the body temperature measured in Fahrenheit (*BodyTemp*) and the person's heart rate (*HeartRate*). These variables are all considered continuous-scale variables whilst the target variable (*RiskLevel*), categorical and ordered with 3 ranks, assesses the level of risk associated with the women's pregnancy[1]. There are no missing values and the mean and quantile values do not seem to present irregularities. We also computed other summary statistics such as density plots, box plots and histograms (using the *hist*, *geom_ boxplot* and *geom_ density* functions) in order to better understand the distribution shape of the data. The results are displayed in the appendix section (Figs. 11 to 22), and given that we will mention them in the following sections since it will help us to verify some assumptions on a number of statistical tests, their interpretation is left for later.

The existence and treatment of **outliers** is of the most absolute importance, since they can leverage our analysis in many ways. Of a variety of manners that one could adress this topic, we opted to apply **Robust PCA** (ROBPCA). Although this is a very versatile technique for dimensionality reduction, it can also be used to identify these extreme observations. This method classifies the observations as *regular*, *good leverage points*, *orthogonal outliers* and *bad leverage points* (when an observation is both a good leverage point and an orthogonal outlier), taking as criteria the robust score distance and the orthogonal distance for every point, which is then compared with a cut-off value. More details about this technique can be found in [3].

In order to compute the ROBPCA, we used the *robpca* function from the *rospca* package. The robustness parameter α was defined as 0.75. This can take any real value between 0.5 and 1, and the higher the value, the more efficient the estimates will be for uncontaminated data. On the other hand, setting a lower value will increase the robustness of the algorithm for contaminated samples. Since it's not possible to know if the device that collected the data was working properly all the time, we opted for this conservative value. As for the principal number of components to keep, the criteria introduced in [3] was considered. The results can be seen in Figure 1. Even though there are some orthogonal outliers (18) and quite a few good leverage points (79), there are no bad leverage points. In spite of this, we decided to look further to these 97 records. Two observations had a registered heart rate of 7 beats per minute, which is inconceivable for a normal person, where usual values range between 60 and 100 beats per minute ([8]). Therefore, these observations were discarded.

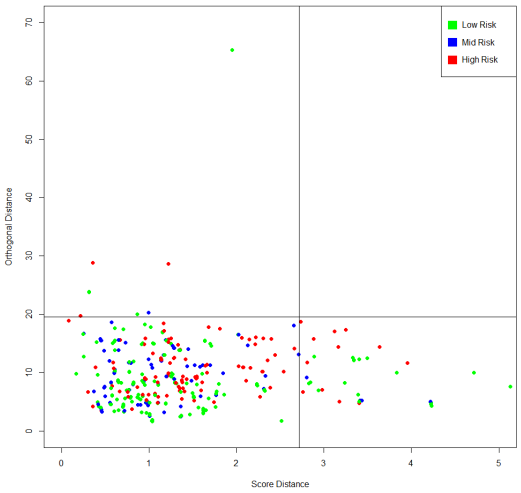


Figure 1: Robust PCA results

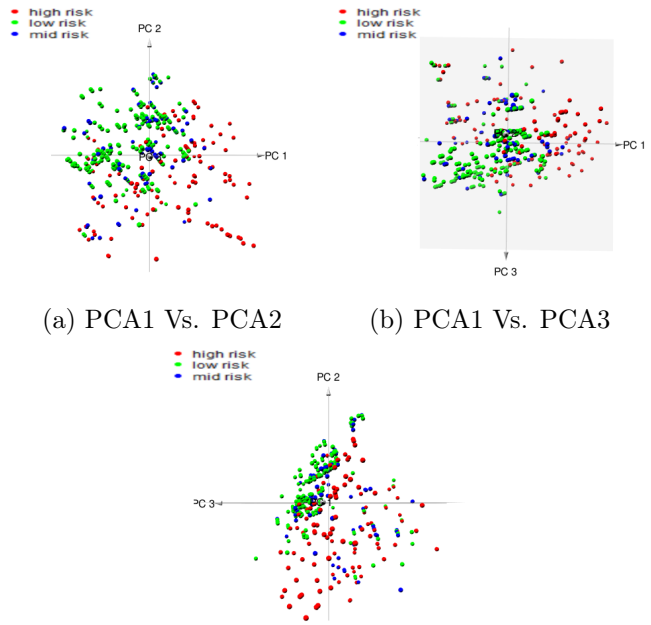
2.2 Data Transformation and Visualization

Even though mathematically rigorous tests are the most precise way to understand data, visualization techniques can also be applied. To this end, we decided to apply **Principal Components Analysis (PCA)** to get a quick "eyeball" over the data, in order to gain a new type of insight of the problem and prepare for the decisions one may have to do up ahead, when dealing with more rigorous statistical tests.

Although PCA is commonly used as a dimensionality reduction technique for datasets composed by a large amount of continuous variables, we will only use it for the purpose of visualization. Since it is a subject studied in many courses, the reader can found a detailed explanation in [4], but the underlying idea is that it converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. It requires standardization of data, since it is a variance maximizing exercise, i.e., it projects the original data onto directions which maximize the variance. This method was computed with the aid of function *prcomp*, implemented in the *stats* package. The 3-dimensional visualization of the PCA results were accomplished with the *pca3d* function, as one can see in Figure 2. The contribution of the variables

[1]From now on, the variables will have the following names in the tables: 1) Age: Age 2) SystolicBP: SBP 3) DiastolicBP: DBP 4) BS: BS 5) BodyTemp: BT 6) HeartRate: HR.

to principal components can be found in Table 1.



(a) PCA1 Vs. PCA2 (b) PCA1 Vs. PCA3

(c) PCA2 Vs. PCA3

Figure 2: PCA visualization through different perspectives

	Age	SBP	DBP	BS	BT	HR
PC1	0.436	0.530	0.522	0.425	-0.274	0.018
PC2	-0.157	0.104	0.123	-0.362	-0.428	-0.797
PC3	0.233	-0.240	-0.304	-0.104	-0.810	-0.358

Table 1: PCA results.

Although PCA does not intend to maximize separability, we can see that large portion of high risk individuals tend to have higher values in the first principal component direction. We can also see that low risk individuals tend to have lower PC1 values and higher PC2 and PC3 values, showing some kind of separability among the different risk groups.

The table also allows us to conclude that PC1 is basically a linear combination of the variables *Age*, *SystolicBP*, *DiastolicBP* and *BS*, while the PC2 is mainly influenced by the *HeartRate*. The PC3 is essentially influenced by *BodyTemp*. This suggests that low risk women are younger with lower blood pressures and sugar levels, and lower body temperatures and heart rates. As for high risk women, these tend be older, with higher blood pressures and sugar levels. Note that for mid risk individuals, it is not so trivial to assess if they tend to have patterns like the other groups seem to have, since they are spread all around the space.

3 Hypothesis Testing

The last section allowed us to get an idea regarding some statistical properties of our sample data. In order to support these theses and draw palpable conclusions, we shall now recur to a more rigorous and robust approach. To achieve this, we will use a variety of methods of statistical analysis, and this section is focused in hypothesis testing. Since the last section showed that there seems to be a discrepancy of values among the different groups, it seems reasonable to assess if some statistics, such as means, are in fact different between them. **Analysis of Variance (ANOVA) test** is a way of doing this. However, this particular method relies on the following assumptions: the data is **independent**; **normal distribution** of the data in each factor level with the **same variance**.

In our particular problem, it seems reasonable to assume that the health indicator measurements for each patient is in no way influenced or related to another patients measurements. The other two presumptions can be verified or refuted more carefully, even though *ANOVA* usually is quite robust against violations of the normality assumptions. When analyzing density plots(Figs. 11 to 16) the distributions don't seem to have a Gaussian shape, for each risk level. This is corroborated by the **Shapiro-Wilk (normality) test** (computed with the function *byf.shapiro*) where the highest *p-value* was less than 10^{-5} (see Table 2), leading us to reject the null hypothesis that the distribution is normal for each level, for the usual significance levels [2]. Throughout the analysis, the significance level we will consider as a cut-off point/critical value is 5%.

	Shapiro-Wilk (Normality) Test					
	Age	**SBP**	**DBP**	**BS**	**BT**	**HR**
L. Risk	< 2.2e-16	< 2.2e-16	1.94e-12	< 2.2e-16	< 2.2e-16	2.11e-11
M. Risk	1.97e-15	< 2.2e-16	1.27e-10	< 2.2e-16	< 2.2e-16	1.07e-09
H. Risk	1.92e-05	1.05e-15	5.36e-15	2.89e-12	< 2.2e-16	4.18e-09

Table 2: Shapiro-Wilk (Normality) Test Results

Violations of the assumption of homogeneity of variances can be more impactful, especially when sample sizes are unequal between levels. Since this is our case (we have 26.8% high risk individuals, 33.1% mid risk and 40.0% low risk), we will verify this variance assumption with a statistical test. Since the results reject normality, we can not apply neither the *F-test* nor the *Bartlett's test*, since they are strongly affected when there is a violation of the normality assumption. Because of this, we will have to resort

[2] 1%, 5% and 10%

to a non-parametric approach, and we decided to apply the *Levene's test* and *Fligner-Killeen's test*. In both tests we draw the same conclusions: due to the very low *p-values*, we reject the null hypothesis of variance homogeneity across groups, except for *Age*, as seen in 3.

	Levene's and Flingner-Kileen's Test					
	Age	**SBP**	**DBP**	**BS**	**BT**	**HR**
p-value Levene's	0.104	2.14e-7	0.005	4.54e-67	1.13e-7	6.536e-5
p-value Flingner-Kileen's	0.029	1.21e-9	0.002	1.36e-96	1.47e-6	1.753e-7

Table 3: Levene's and Flingner-Kileen's Test Restults

One way to overcome the non-homogeneity of variance of the data is, for example, with data transformations, given that they are continuous, reversible and maintain a certain type of order. Nevertheless, even if this technique could solve our variance problem, the conclusions could not be generalized to the original data. The lack of interpretability of these results drove us to discard this kind of procedure. Therefore, we decided to solely apply the **Kruskal-Wallis test**, a non-parametric test, that does not rely on these rejected assumptions. The results for this test can be found below in Table 4.

	Kruskal-Wallis Test					
	Age	**SBP**	**DBP**	**BS**	**BT**	**HR**
p-val	6.56e-22	6.78e-37	9.66e-30	9.68e-67	8.65e-08	1.21e-08

Table 4: Kruskal-Wallis Test Results

Since all *p-values* are low, we reject the hypothesis that for each level, the medians are the same, at the usual significance levels. Even though this test is different from the ANOVA test, since it tests the null hypothesis that the groups have the same median (and not the mean), we still decided to use it. Moreover, the median is a more robust measure to outliers than the mean. Since we still suspected that the means could possibly be different, we decided to apply **pairwise Welch's t-tests** (with *Bonferroni correction*) to assess this question. This is an adaptation of *Student's t-test*, more reliable when the two samples have unequal variances and possibly unequal sample sizes, which is our case. The results can be seen in Tables 5, 6, 7, 8, 9 and 10.

	Pairwise t-tests with non-pooled SD - Age	
	High Risk	**Low Risk**
Low Risk	< 2e-16	
Mid Risk	6.8e-13	0.39

Table 5: Pairwise t-test result for the variable Age

	Pairwise t-tests with non-pooled SD - DBP	
	High Risk	**Low Risk**
Low Risk	< 2e-16	
Mid Risk	<2e-16	0.18

Table 6: Pairwise t-test result for the variable *DiastolicBP*

	Pairwise t-tests with non-pooled SD - SBP	
	High Risk	**Low Risk**
Low Risk	< 2e-16	
Mid Risk	9.7e-13	5.5e-10

Table 7: Pairwise t-test result for the variable *SystolicBP*

	Pairwise t-tests with non-pooled SD - BS	
	High Risk	**Low Risk**
Low Risk	< 2e-16	
Mid Risk	< 2e-16	2.8e-05

Table 8: Pairwise t-test result for the variable BS

	Pairwise t tests with non-pooled SD - BT	
	High Risk	**Low Risk**
Low Risk	< 5.7e-06	
Mid Risk	1	5.2e-06

Table 9: Pairwise t test result for the variable *BodyTemp*

	Pairwise t tests with non-pooled SD - HR	
	High Risk	**Low Risk**
Low Risk	3.7e-08	
Mid Risk	0.00023	0.09781

Table 10: Pairwise t test result for the variable *HeartRate*

For the variable *Age* we infer that there are differences between the means of low and high risk groups, and also between the means of mid and high groups, but not between low and mid risk groups. The same is concluded for the *DiastolicBP* and *HeartRate* variables, which is something that aligns with what we had perceived about the data when we visualized it. However, the *BodyTemp* variable displays a quite interesting behavior. The means between low and high risk groups and low and mid groups are different, however, for the mid risk and high risk groups it seems that there are no differences between those individuals, since the p-value is identical to 1. Lastly, for the*SystolicBP* and *BS* variables we conclude that their means are statistically different between all groups, leading us to the belief that they have more impact in the risk factor.

4 Information Theory

Information theory can also be used to evaluate the impact of our descriptive variables in the risk level. This field was initially developed by Claude E. Shannon in 1984 and is a mathematical approach to the study of coding of information along with its quantification, storage and communication. The *amount of information* presented in a variable can be calculated through means of the **entropy**. Informally speaking, the entropy of a random variable is the average level of *information* inherent in the variable's possible outcome. The more certain a variable is about an event, the less information it will contain. This concept can be expressed mathematically as:

$$H(j) = \sum_{i=1}^{n} -p(x_i)\log_2(p(x_i))$$

where n is the number of possible outcomes and

$$p(x_i) = \frac{n.\ samples\ in\ class\ i\ at\ node\ j}{n.\ total\ samples\ at\ node\ j}.$$

A systematic way to calculate and understand which are the variables that give us less entropy can be defined as:

1. *Calculate the initial entropy of the system, as defined in the equation above.*

2. *Find which variable most reduces the system's entropy. This can be achieved by finding the variable that splits our data into two different groups/nodes (according to a certain threshold) that minimizes the* ***Information Gain (IG)****, defined as*

   $$IG(i,j) = H(i) - p \cdot H(j_{\text{left}}) - (1-p) \cdot H(j_{\text{right}}).$$

   where $IG(i,j)$ is the information gain from transiting from state i to state j, $H(i)$ is the entropy from the previous state i, $H(j_{\text{left}})$ is the entropy of the (new) state j at the left node, $H(j_{\text{right}})$ is the entropy of the (new) state j at the right node, p and $1-p$ are the proportion of elements that transited to the left and right node, respectively. Note that this is a greedy procedure, since we find the variable and its optimal threshold by brute force.

3. *Calculate the information gain (or entropy reduction).*

This procedure is the framework of a **Decision Tree**, which is a well-known supervised learning algorithm. Although the main scope of this work is not to build and fine-tune machine learning models, we will train a decision tree based on the entropy criteria to assess the importance of each variable. We also decided to implement this algorithmic procedure to a smaller partition of our dataset. We split the data in 80% for training and 20% for testing. The top of the tree can be seen in Figure 3. This tree was created using the *DecisionTreeClassifier* function from the *scikit-learn* package, in Python

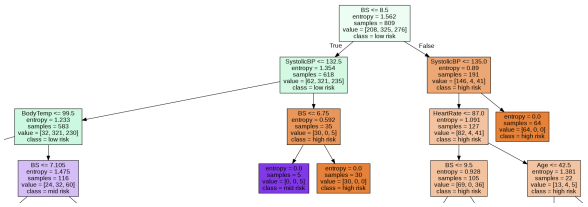


Figure 3: Top of the Decision Tree

The first takeaway is that *BS* is the purest variable in the model, meaning this variable has less uncertainty. After finding the threshold of 8.5 for the variable we can see that the information gain obtained is 0.32, approximately. The variable considered by the algorithm in the following step is *SystolicBP*. These two variables were the ones we concluded that had a greater impact in the risk level by the hypotheses testing methods. Therefore, hypothesis testing and information theory point towards the same conclusion: *BS* and *SystolicBP* seem to have the most impact on maternal mortality risk.

We also allowed this decision tree to grow up to a depth of 10 variable tests per branch, and found that for the test set the model was able to make good predictions, even without a great deal of focus in hyperparameter tuning and the use of other indices. The results are presented in Tables 11, 12 and 13.

		Predicted		
		Low Risk	**Mid Risk**	**High Risk**
Real	**Low Risk**	62	1	1
	Mid Risk	3	65	11
	High Risk	3	15	41

Table 11: Confusion Matrix of the Decision Tree

	Precision	**Recall**	**F1-Score**	**Global Accuracy**
Low Risk (n = 64)	0.91	0.97	0.94	0.83
Mid Risk (n = 79)	0.80	0.82	0.81	
High Risk (n = 60)	0.78	0.70	0.74	

Table 12: Evaluation Metrics of the Decision Tree

Weight. Precision	Weight. Recall	Weight. F1-Score
0.83	0.83	0.83

Table 13: Weighted Evaluation Metrics of the Decision Tree

As one can observe, the predictions are satisfactory, for a model that isn't optimized at its maximum.

5 Cumulative Link Models

Cumulative Link Models are a tool of statistical analysis used to predict and assess the importance of explanatory variables in a model based on prior observations. Both models introduced in this section arise from the cumulative distribution of the response variable. Given that we are dealing with a multi-class problem, and there is a clear **ordering of the categorical variable**, the underlying structure of the response variable is of the utmost importance.

Let g be a transformation function, mapping probabilities to the real line and P_j be the cumulative probability up to class j, i.e., $P_j = P(Y \leq j)$. Then the class of models that we will consider assumes that the transformed cumulative probabilities are a linear function of the predictors, of the form

$$g(P_j) = \beta_j + \boldsymbol{\beta}\boldsymbol{X}.$$

Both methods presented are different ways to model the function g and were computed using the function *polr* from the *MASS* package.

5.1 Ordinal Logistic Regression

The first model we will consider is a variation of the usual **logistic regression model**, that works just for binary outcomes. Instead of applying the *logit* transformation to the response probabilities, we now apply it to the cumulative probabilities, which now have an established order. This method is known as **Ordinal Logistic Regression**. This will enable us to determine which of our variables, if any, have a statistically significant effect on our dependent variable.

Let Y be an ordinal outcome with K categories. The model is now written as:

$$\begin{aligned}\text{logit}(P_j) &= \log\left(\frac{P_j}{1-P_j}\right) \\ &= \beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jp}X_p\end{aligned}$$

where $\beta_{j0}, \beta_{j1}, \cdots, \beta_{jp}$ are the model coefficient parameters with p predictors, for $j = 1, \cdots, K-1$. This model hinges on two fundamental assumptions: No **multi-collinearity** - i.e. that two or more independent variables are not highly intercorrelated with each other; and **proportional odds** - each independent variable has an identical effect at each cumulative split of the ordinal dependent variable. Due to this last premise, the intercepts are different for each category but the slopes are constant across categories, simplifying the equation above to

$$\text{logit}(P_j) = \beta_j + \boldsymbol{\beta}^\intercal \boldsymbol{X}$$

with $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ and $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ column vectors. We also can define the **odds of being less than or equal to a particular category** j as $P_j/(1-P_j)$. If we exponentiate the last equation, we get that these odds are:

$$\frac{P_j}{1-P_j} = \lambda_j e^{\boldsymbol{\beta}^\intercal \boldsymbol{X}}$$

where $\lambda_j = e^{\theta_j}$. It can be hard to interpret this coefficient, as it will depend on the context at hand. However, the general idea is that when $\boldsymbol{x} = 0$, the λ_j are the baseline odds of a response in category j or below. The other factor represents the effect of the covariates in the raising or diminishing of the odds. Since the effect of this change is proportional in the odds, then if a certain combination of covariate values doubles the odds of being in category 1, it also doubles of being in category 2 or below, and so on. For this reason, another name associated with it is the **proportional odds model**.

5.2 Ordered Probit Regression

The **Ordered Probit Model** results from modeling the probit (quantile function associated with the standard normal distribution) of the cumulative probabilities as a linear function of the covariates. Therefore, this model can be written as

$$\Phi^{-1}(P_j) = \beta_j + \boldsymbol{\beta}^\intercal \boldsymbol{X}$$

where Φ is the standard normal cdf. Estimates from the ordered probit model are usually very similar to estimates from the ordered logit model - as one would expect, due to the similarity between the normal and the logistic distributions.

5.3 Modeling and Results

Before doing any kind of calculations, we must check if the assumptions hold for our model. Multicollinearity can occur, for example, when two or

more independent variables are correlated with each other, when an independent variable is a linear combination of other variables or if independent variables provide similar and repetitive results. Hence, we will make use of the **pair correlation** values and the **Variance Inflation Function (VIF)** to check if at least some of these premises hold. Let us first consider the correlation plot in Figure 4

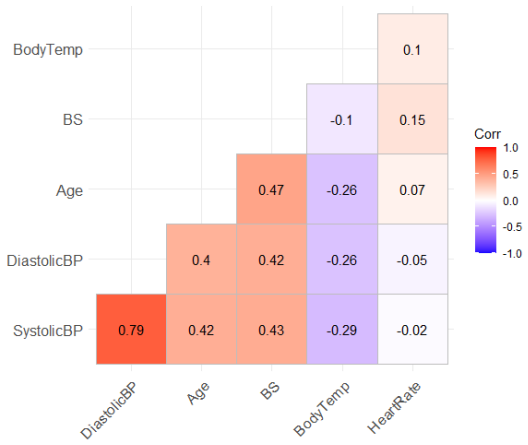


Figure 4: Correlation plot of the data

As we observed, *DiastolicBP* and *SystolicBP* are highly correlated. This is somehow expected, since usually a person with high blood pressure in the diastole also has a high blood pressure in the systole (in comparison to the standard values for both periods). There is also some moderate correlation between other pairs of variables, however it is not enough to conclude that the collinearity is big enough to cause any disruption in our model. Nevertheless, pairwise correlation can be misleading, since a combination of independent variables can explain a very high proportion of the variance of other variable. That's where the *VIF* comes in play. Mathematically, this quantity equals the ratio of the overall model variance to the variance of a model that includes only that single independent variable, when that model is a linear one. High values indicate that it is difficult to assess accurately the contribution of predictors to a model. Table 14 presents the results obtained:

	Age	**SBP**	**DBP**	**BS**	**BT**	**HR**
VIF	1.43	2.83	2.76	1.45	1.15	1.06

Table 14: VIF associated to each variable.

The most general rule of thumb is to consider that there exists strong presence of multicollinearity if the *VIF* values are greater than 5. There is no such case, but we do see the same dependence idea between *DiastolicBP* and *SystolicBP*.

With all of this knowledge, we decided to remove the variable *DiastolicBP*. Weakly or moderate correlated variables can be included, but highly correlated pairs should be avoided. This reason, alongside a moderate value of the *VIF*, leads us to make this decision. This is one of the various procedures one can do to deal with collinear variables. This seems reasonable in this context due to the reasons presented above, however this is not a blanket policy. There are methods that take into account the existence of multicollinearity and they can also be explored in some contexts. For the proportional odds assumption, the **Brant hypothesis test** was used. A more formal explanation can be found in [1], but the general idea is to test the null hypothesis that the coefficient parameters $\boldsymbol{\beta}$ are the same as if they were estimated for each separation j. The results are presented in Table 15.

	Omn.	**Age**	**SBP**	**DBP**	**BS**	**BT**	**HR**
p-val all vars	0	0.42	0	0	0.51	0.79	0.04
p-val remove	0.39	0.43	0.89	NA	0.13	0.85	0.1

Table 15: p-values for the Brant Hypothesis test.

Before removing the *DiastolicBP* variable, the Brant test rejected the idea of proportional odds, due to the presence of multicollinearity. After the removal, we have have moderate to high p-values in all variables so we derive that we do not reject the null hypothesis, and therefore the assumption is not violated.

5.3.1 Ordinal Logistic Regression Results

The results for the coefficients and the two intercepts between low risk and medium risk, and between medium risk and high risk can be seen in Table 16:

	Age	**BP**	**BS**	**BT**	**HR**	**Low/Mid**	**Mid/High**
Coef	-0.012	0.045	0.463	0.458	0.039	55.960	58.276

Table 16: Coefficient and Intercept values - OLR model

Since our main goal is to assess the importance of the covariates, we can calculate confidence intervals for the parameters estimates, either by **profiling the likelihood function** ([7]) or by assuming **nor-**

mal distribution in the standard errors. Informally speaking, if the 95% CI do not cross 0, the parameter estimate is most likely statistically significant. The results for both cases can be seen in Tables 17 and 18:

	Age	BP	BS	BT	HR
2.5% IC	-0.0258	0.0349	0.3883	NA	0.0179
97.5% IC	0.0007	0.0557	0.5445	NA	0.0593

Table 17: Confidence Intervals when profiling the likelihood function - OLR model

	Age	SBP	BS	BT	HR
2.5% IC	-0.0255	0.0355	0.3856	0.4380	0.0178
97.5% IC	0.0007	0.0547	0.5413	0.4784	0.0592

Table 18: Confidence Intervals when assuming normal distribution - OLR model

For the profiled CI, we notice that we were unable to calculate the extremes of the intervals for *BodyTemp*. This is due to specifics of profiling the likelihood function in *R*, where one of the parameters has to span a wide enough range in order to get the interval limits. However, since the results are basically the same for the other variables, we will use them as our baseline. Therefore, we reckon that variable *Age* is not statistically significant to our model, as one would assume due to the small parameter value associated with this variable. Also, even though the confidence intervals do not allow us to make such assertion, *SystolicBP* and *HeartRate* are almost insignificant to the model, whilst *Blood Sugar* and *BodyTemp* are the most important ones. We can also compute the **odds ratio** and their **confidence intervals** (Figure 19.

	Age	SBP	BS	BT	HR
OR	0.9877	1.0462	1.5896	1.5812	1.0393
2.5% IC	0.9749	1.0362	1.4704	1.549	1.0180
97.5% IC	1.0007	1.0562	1.7183	1.613	1.0610

Table 19: Odds ratio and 95% confidence intervals - OLR model

Roughly speaking, odds ratios are used to compare the relative odds of the occurrence of the outcome of interest given exposure to the variable of interest and one of its main uses is to compare the magnitude of various risk factors for that outcome. When this ratio is close to 1, the variable has little to no effect in the odds of the outcome, while a value greater than 1 shows that the variable provokes an increase in the odds and a value smaller than 1 corresponds to a decrease in the odds. Analyzing Table 19, we obtain the same conclusions as for the coefficients. The variable *Age* is statistically insignificant, and the *BloodSugar* and *BodyTemp* variables are the most significant, as supported by our confidence intervals.

5.3.2 Ordered Probit Regression Results

The general results for the Ordered Probit Model are presented in the tables below :

	Age	BP	BS	BT	HR	Low/Mid	Mid/High
Coef	-0.007	0.026	0.265	0.267	0.022	32.536	33.894

Table 20: Coefficient and Intercept values - OPR model

	Age	SBP	BS	BT	HR
2.5% IC	-0.0147	0.0199	0.2251	NA	0.0104
97.5% IC	0.0009	0.0316	0.3071	NA	0.0343

Table 21: Confidence intervals when profiling the likelihood function - OPR model

	Age	SBP	BS	BT	HR
2.5% IC	-0.0145	0.0203	0.2243	0.2559	0.0104
97.5% IC	0.0008	0.0312	0.3063	0.2784	0.0343

Table 22: Confidence intervals assuming Normal distribution - OPR model

As one would expect, the values and conclusions are basically the same, since the models are closely related. We also decided to use this model for prediction, as in the decision tree model. The confusion matrices and metrics of both models are resumed in Tables 23, 24, 25, 26, 27 and 28.

		Real		
		Low Risk	**Mid Risk**	**High Risk**
Pred.	**Low Risk**	65	27	1
	Mid Risk	13	27	26
	High Risk	1	6	37

Table 23: Confusion Matrix for the Ordinal Logistic Regression model

	Recall	Precision	F1-Score	Global Accuracy
Low Risk (n = 64)	0.82	0.70	0.76	0.64
Mid Risk (n = 79)	0.45	0.41	0.43	
High Risk (n = 60)	0.58	0.84	0.69	

Table 24: Evaluation Metrics of the Ordinal Logistic Regression model

Weight. Precision	Weight. Recall	Weight. F1-Score
0.63	0.61	0.61

Table 25: Weighted Evaluation Metrics of the Logistic Regression model

		Real		
		Low Risk	**Mid Risk**	**High Risk**
Pred.	**Low Risk**	66	28	1
	Mid Risk	12	26	26
	High Risk	1	6	37

Table 26: Confusion Matrix for the Ordered Probit Regression model

	Recall	Precision	F1-Score	Global Accuracy
Low Risk (n = 64)	0.84	0.69	0.76	0.64
Mid Risk (n = 79)	0.43	0.41	0.42	
High Risk (n = 60)	0.58	0.84	0.69	

Table 27: Evaluation Metrics of the Ordered Probit Regression model

Weight. Precision	Weight. Recall	Weight. F1-Score
0.63	0.60	0.60

Table 28: Weighted Evaluation Metrics of the Ordered Probit Regression model

These models, as the decision tree, predict accurately the data, with high overall accuracy. This leads us to believe that the variables *BloodSugar* and *BodyTemp* have a high influence in our response variable.

6 Conclusions

Throughout this work our goal was to assess which variables played a major role in the dynamics of the risk level associated with pregnant women. After performing Robust PCA, two outliers were detected, which lead these two observations to be discarded. For the purpose of visualization we then applied PCA and verified some separability between the different risk groups. This analysis corroborated the natural tendency of the human body, that is, low risk women are younger with lower blood pressures and sugar levels, with higher body temperatures and heart rates whilst high risk women have lower blood pressures and sugar levels, lower body temperatures and heart rates.

When performing hypothesis testing, there were a lot of commonly used tests that were rejected, since our data didn't meet their assumptions. Instead, non-parametric tests such as *Levene's test* and *Fligner-Killeen's test* were performed. In both tests the low p-value indicated that there was no variance homogeneity across groups, except for *Age*. Afterwards, the non parametric Kruskal-Wallis test indicated that we should reject the hypothesis, that for each level of Risk, the medians are the same. To ascertain which means were equal and those that were not, we used pairwise Welch's t-tests(with *Bonferroni correction*), which corroborated what we had perceived earlier for the variables *Age*, *DiastolicBP* and *HeartRate*. For the *BodyTemp*, the result was a little different from what we expected but nevertheless, the remaining two variables (*SystolicBP* and *BS*) turned out to be the most significant in the risk factor. Afterwards, resorting to methods originated from Information Theory, in order to assess the importance of each variable, we trained a decision tree based on the entropy criteria. This method turned out to corroborate what we had already seen in the hypothesis testing, that is, *BS* and *SystolicBP* have the highest impact on maternal mortality risk. Finally, we built two predictive models, Ordered Logistic Regression and Ordered Probit Regression. Both models predicted the data with a relatively high accuracy, which showed us that the variables *BS* and *BodyTemp* have a high influence in the Risk Level. This is a little different than what we observed in the Decision Trees, however *BS* keeps having a huge impact on the categorical variable.

In short, the decision tree performed better. This can be due to the fact that our regression models rely in certain assumptions that even though were tested and verified, they always entail some kind of approximation to the truth, and can be misleading. So, if one had to choose a model for prediction, it would be the Decision Tree model.

These results allowed the identification of pri-

mal indicators and gave us a better understanding of what can potentially affect maternal mortality. However, the number of potential indicators collected was fairly poor. One of the ways to improve these models in the future is to work with medical researchers so that more data could be gathered about new potentially good indicators, and most likely leading us to better results.

References

[1] Rollin Brant. "Assessing proportionality in the proportional odds model for ordinal logistic regression". In: *Biometrics* (1990), pp. 1171–1178.

[2] *Goal 3: Ensure healthy lives and promote well-being for all at all ages*. `https://www.un.org/sustainabledevelopment/health/`. Accessed: 2022-06-03.

[3] Mia Hubert, Peter Rousseeuw, and Karlien Branden. "ROBPCA: A new approach to robust principal component analysis". In: *Technometrics* 47 (Feb. 2005), pp. 64–79. DOI: `10.1198/004017004000000563`.

[4] Richard A. Johnson and Dean W. Wichern. "8 - Principal Components". In: *Applied Multivariate Statistical Analysis*. Prentice Hall, 1999.

[5] *Maternal Health Risk Data Set*. `https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set/`. Accessed: 2022-06-03.

[6] *Maternal mortality*. `https://www.who.int/news-room/fact-sheets/detail/maternal-mortality`. Accessed: 2022-06-03.

[7] David Patterson Patterson. *Profiling the likelihood confidence intervals for GLM's*. URL: `http://www.math.umt.edu/patterson/ProfileLikelihoodCI.pdf`.

[8] *What's a Normal Resting Heart Rate?* https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979. Accessed: 2022-06-03.

Appendix Images

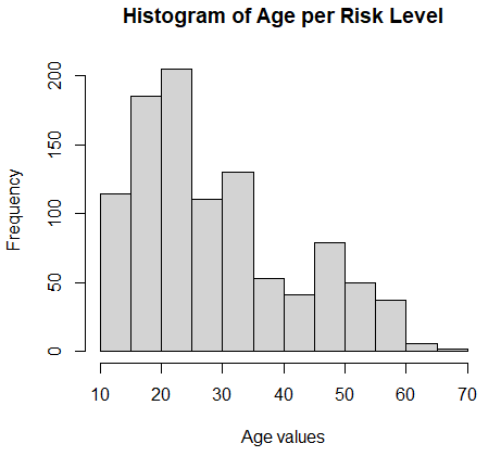


Figure 5: Histogram of Age

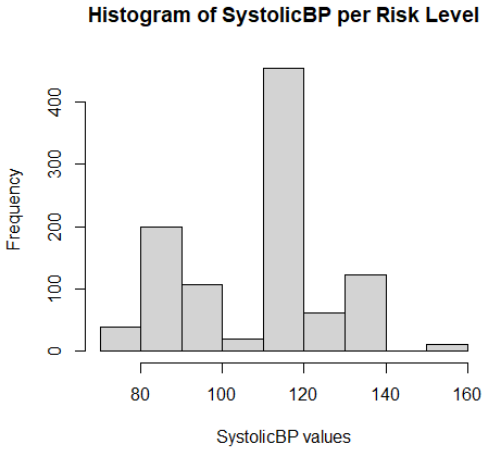


Figure 6: Histogram of SystolicBP

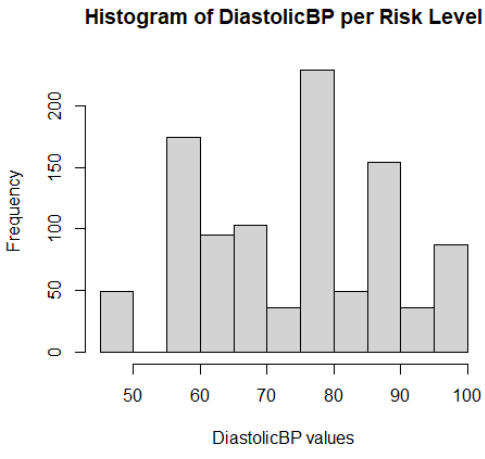


Figure 7: Histogram of DiastolicBP

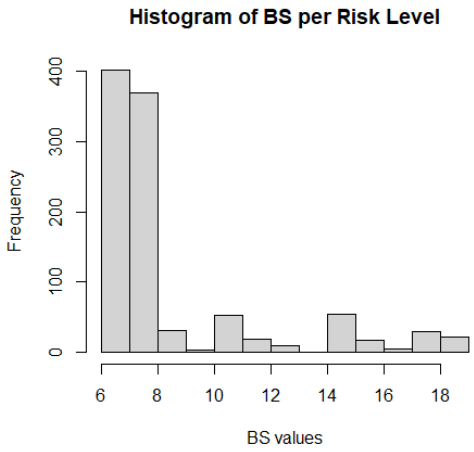


Figure 8: Histogram of BS

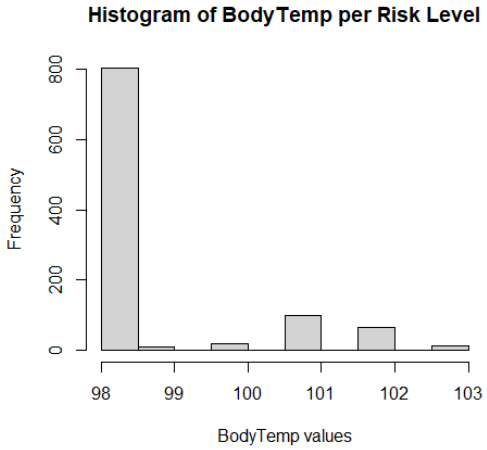


Figure 9: Histogram of BodyTemp

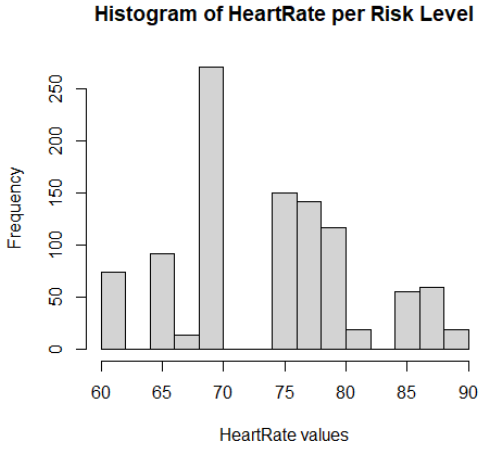


Figure 10: Histogram of HeartRate

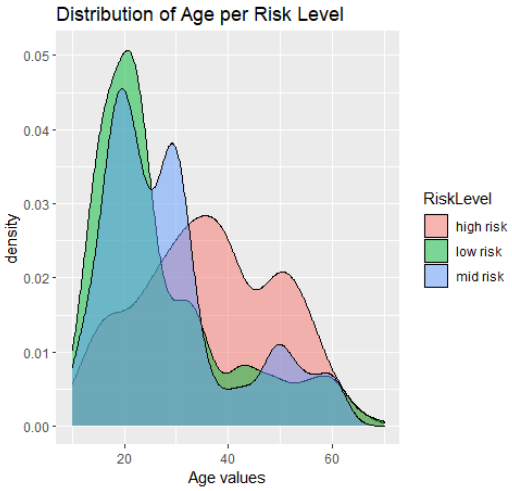


Figure 11: Density plot of Age

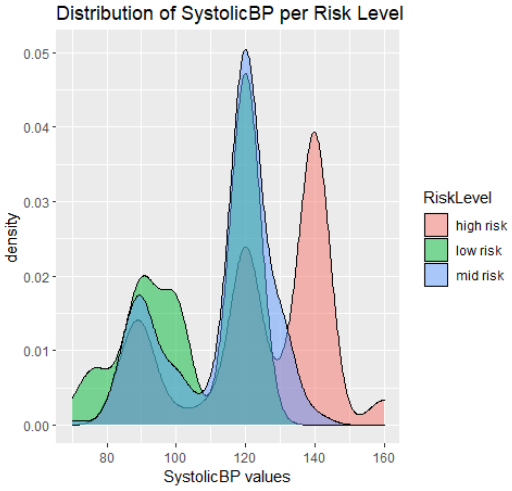


Figure 12: Density plot of SystolicBP

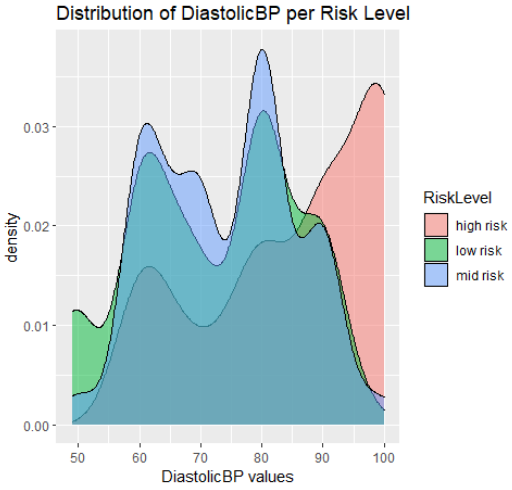


Figure 13: Density plot of DiastolicBP

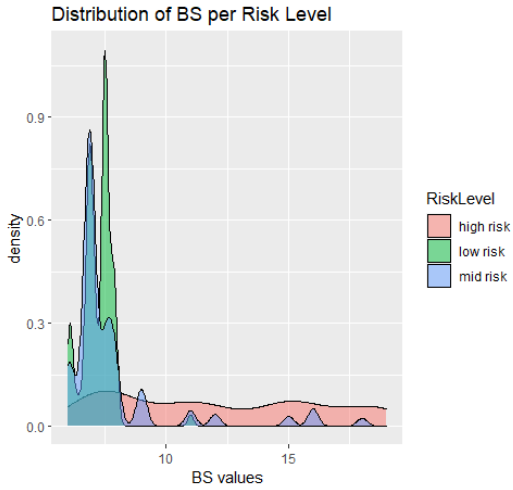


Figure 14: Density plot of BS

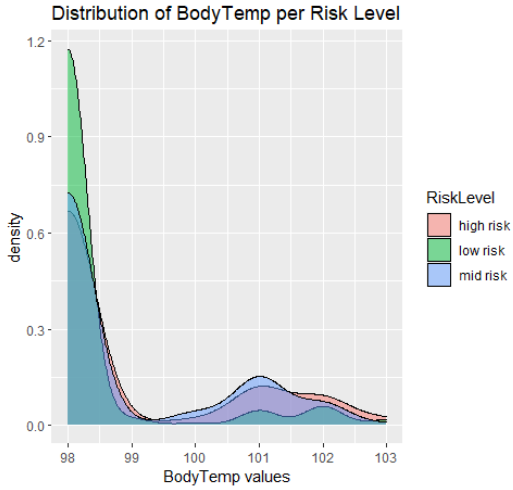


Figure 15: Density plot of BodyTemp

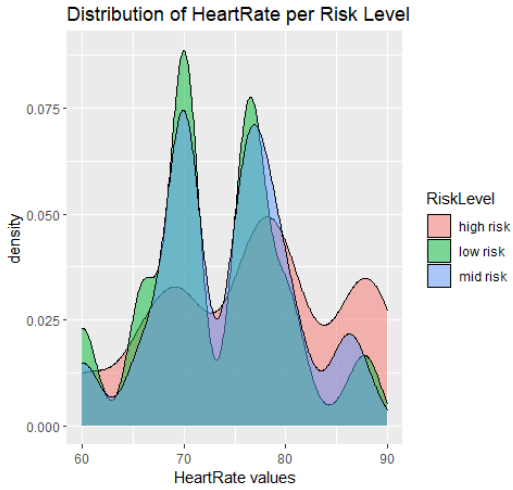


Figure 16: Density plot of HeartRate

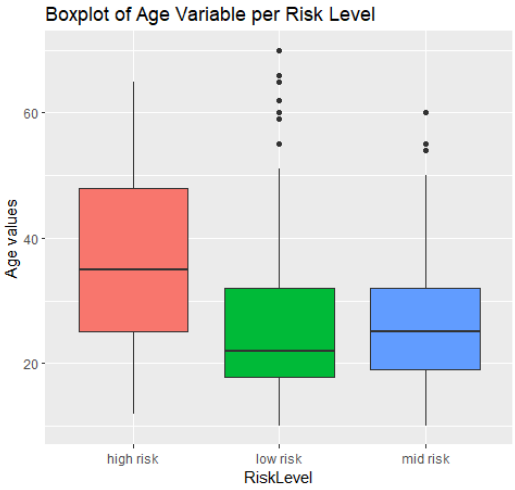


Figure 17: Boxplot of Age

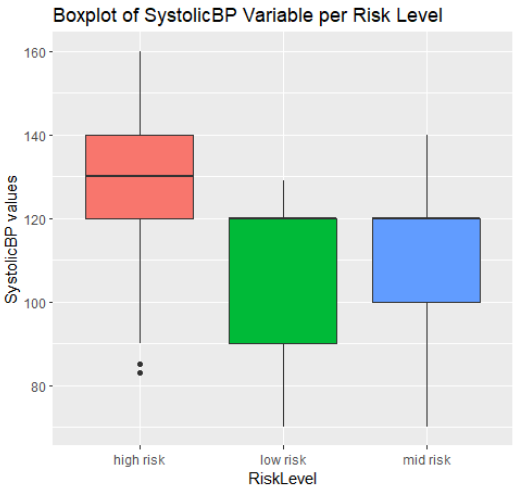


Figure 18: Boxplot of SystolicBP

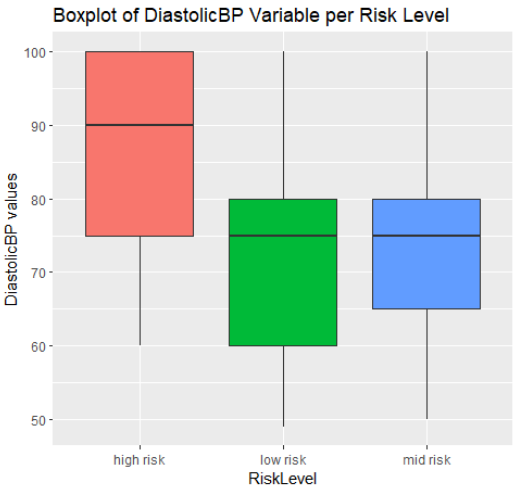


Figure 19: Boxplot of DiastolicBP

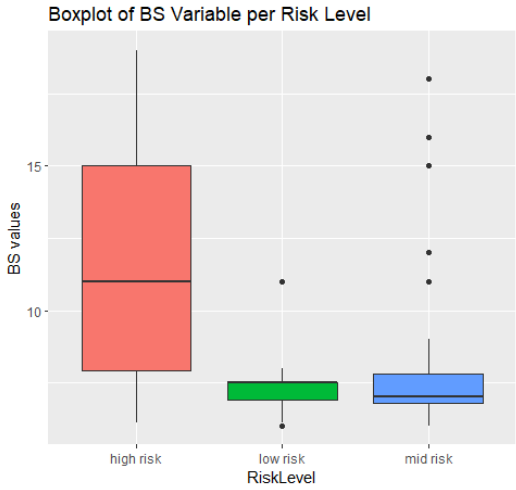


Figure 20: Boxplot of BS

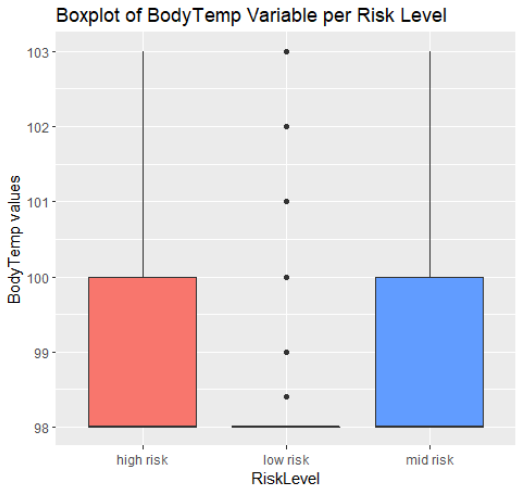


Figure 21: Boxplot of BodyTemp

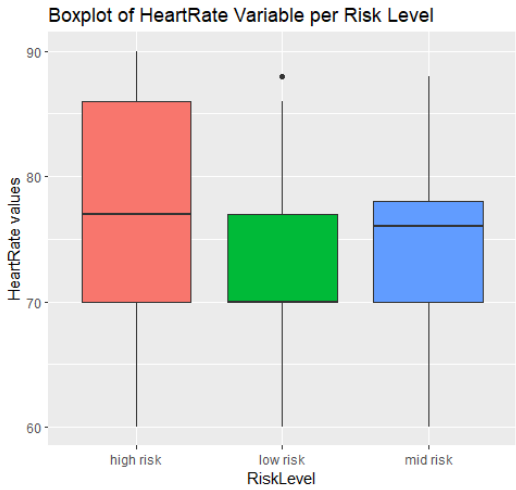


Figure 22: Boxplot of HeartRate