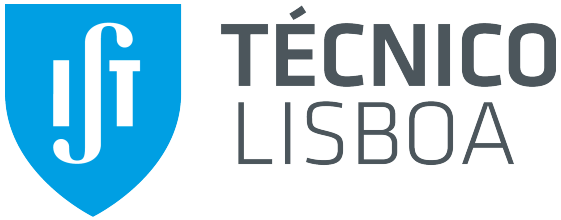


Biostatistics

Assessing the impact of key health indicators on maternal mortality

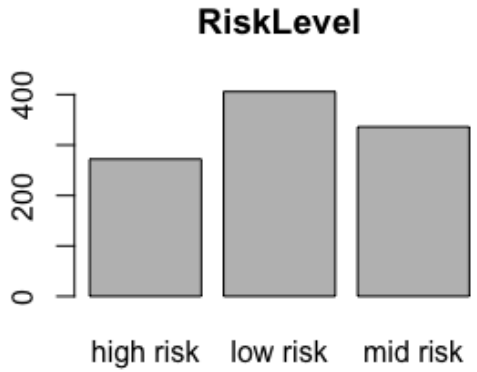
Exploratory Data Analysis

* Age (num)
* Systolic blood pressure (SystolicBP)(num)
* Diastolic blood pressure (DiastolicBP)(num)
* Blood sugar level (BS)(num)
* Body temperature (BodyTemp)(num)
* Heart rate (HeartRate)(chr)

num is for numerical type of variable and chr is for cathegorical type of variable

Exploratory Data Analysis

* 1014 observations, 817(80%) for train and 202(20%) for test.
* Target variable (RiskLevel) is categorical and ordered in 3 ranks. The proportion for each Level is 26.8% high risk individuals, 33.1% mid risk and 40.0% low risk. Percentage is rounded to second decimal point.
* No missing values.



Robust PCA

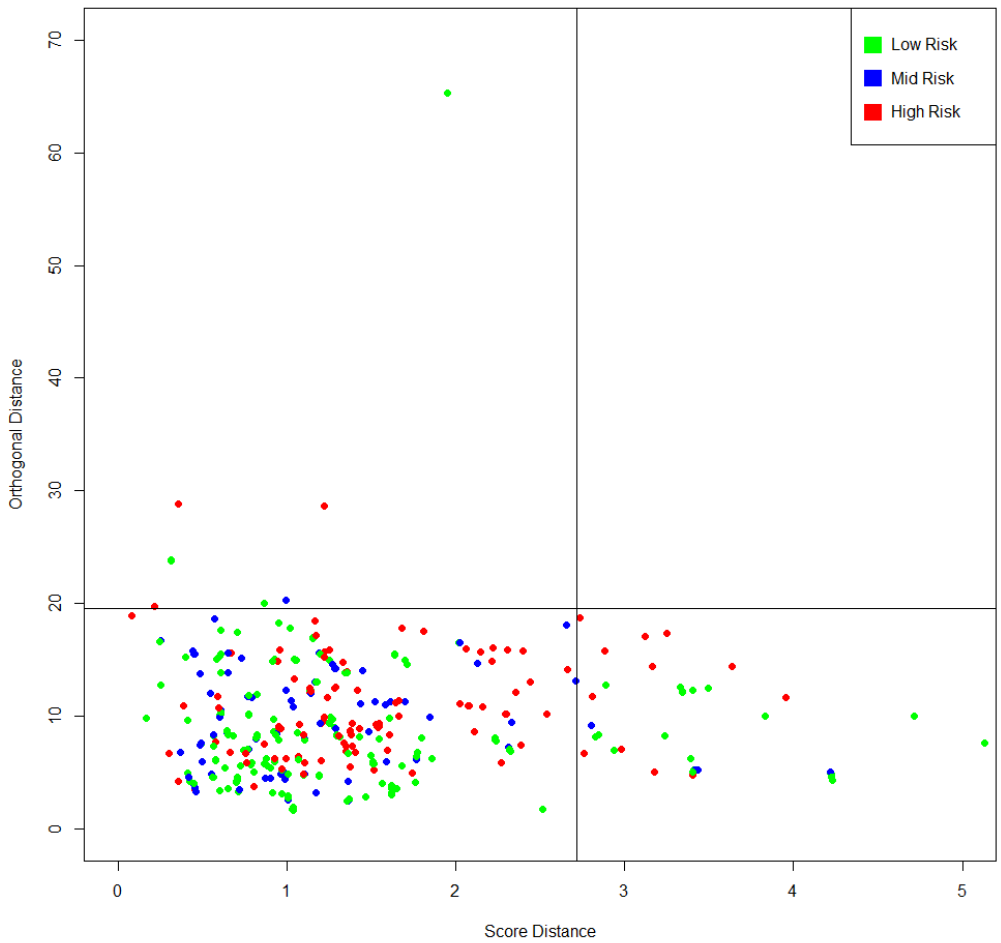
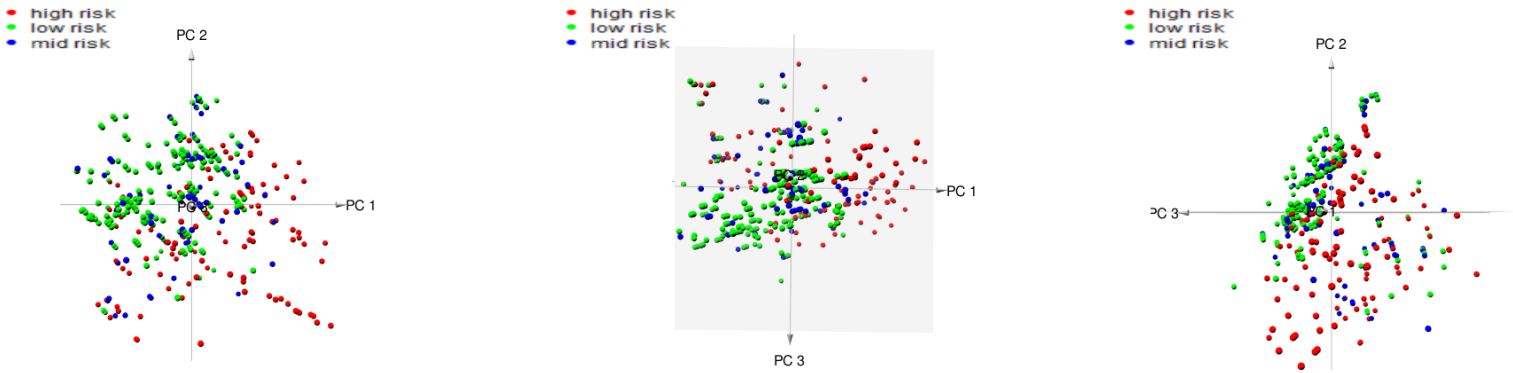


Figure 1: Robust PCA results

Data Transformation and Visualization



(a) PC1 Vs. PC2 **(b)** PC1 Vs. PC3 **(c)** PC3 Vs. PC2

Figure 2: Visualization of a PCA results using three different perspectives

Data Visualization and Transformation

	PC1	PC2	PC3
Age	0,436	-0,157	0,233
SystolicBP	0,53	0,104	-0,24
DiastolicBP	0,522	0,123	-0,304
BS	0,425	-0,362	-0,104
BodyTemp	-0,274	-0,428	-0,81
HeartRate	0,018	-0,797	0,358

Figure 3: Contribution of the original variables to the first three principal components

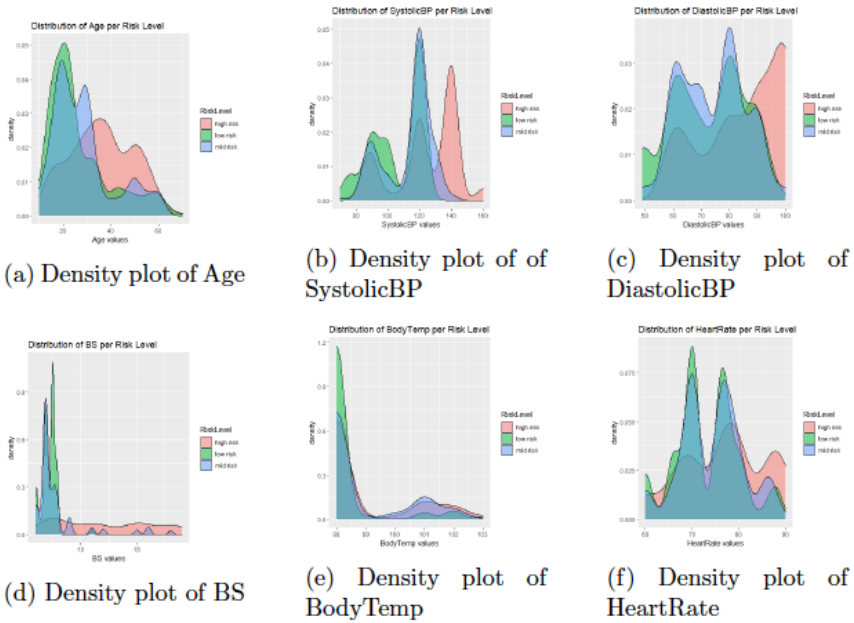
Data Visualization and Transformation

From the table and graphs of the PCA we can make the following inferences:

- There is some separability amongst the different risk groups;
- Large portion of high risk individuals tend to have higher values in the PC1 direction;
- Low risk individuals tend to have lower PC1 values and higher PC2 and PC3 values.
- High risk individuals tend to have higher PC1 values.
- PC1 has high weights in Age, SystolicBP, DiastolicBP and BS, which will lead to elevated values in this direction.
- PC2 is mainly influenced by the variable HeartRate.
- PC3 is essentially influenced by BodyTemp.

Hypotheses Testing

- For our hypothesis testing we considered a threshold of 0.05 for the p-value.
- Even though the Anova is quite robust against violations of the normality assumption, looking at the density plots we can see that within each Risk Level, the distributions don't have a normal shape.



(a) Density plot of Age

(b) Density plot of of SystolicBP

(c) Density plot of DiastolicBP

(d) Density plot of BS

(e) Density plot of BodyTemp

(f) Density plot of HeartRate

Figure 4: Density plot of the variables according to Risk Level

Hypothesis Testing

- Shapiro-Wilk (normality) tests, where the highest p-value was $< 10^{-}5$.

	Shapiro-Wilk (Normality)Test					
	Age	**SBP**	**DBP**	**BS**	**BT**	**HR**
Low Risk	$< 2{,}2 * 10^{-16}$	$< 2{,}2 * 10^{-16}$	$1{,}94 * 10^{-12}$	$< 2{,}2 * 10^{-16}$	$< 2{,}2 * 10^{-16}$	$2{,}11 * 10^{-11}$
Mid Risk	$1{,}97 * 10^{-15}$	$< 2{,}2 * 10^{-16}$	$1{,}27 * 10^{-10}$	$< 2{,}2 * 10^{-16}$	$< 2{,}2 * 10^{-16}$	$1{,}07 * 10^{-9}$
High Risk	$1{,}92 * 10^{-5}$	$1{,}05 * 10^{-15}$	$5{,}36 * 10^{-15}$	$2{,}89 * 10^{-12}$	$< 2{,}2 * 10^{-16}$	$4{,}18 * 10^{-9}$

Figure 5: Shapiro-Wilk (Normality) Test Results

Hypothesis Testing

- Discard both the F-test and the Bartlett's test.
- Levene's test and Fligner-Killeen's test, which are non-parametric tests.
- Homogeneity of variance across groups was rejected for all variables, except for Age.
- The Kruskal-Wallis test was used, therefore discarding an ANOVA Test.

	Levene's and Flingner-Kileen's Test					
	Age	**SBP**	**DBP**	**BS**	**BT**	**HR**
p-value (Levene's Test)	0,104	$2{,}144 * 10^{-7}$	0,005	$4{,}543 * 10^{-67}$	$1{,}134 * 10^{-7}$	$6{,}536 * 10^{-5}$
p-value (F.Killeen's Test)	0,029	$1{,}206 * 10^{-9}$	0,002	$1{,}364 * 10^{-96}$	$1{,}470 * 10^{-6}$	$1{,}753 * 10^{-7}$

Figure 6: Levene's and Flingner-Killeen's Test Restults

	Kruskal-Wallis Test					
	Age	**SBP**	**DBP**	**BS**	**BT**	**HR**
p-value	$6{,}56 * 10^{-22}$	$6{,}78 * 10^{-37}$	$9{,}66 * 10^{-30}$	$9{,}68 * 10^{-67}$	$8{,}65 * 10^{-8}$	$1{,}21 * 10^{-8}$

Figure 7: Kruskal-Wallis Test Results

Hypothesis Testing

- Parwise Welch's t-tests (with Bonferroni correction).

	Pairwise t tests with non-pooled SD-Age	
	High Risk	**Low Risk**
Low Risk	$< 2 * 10^{-16}$	
Mid Risk	$6{,}8 * 10^{-13}$	0,39

	Pairwise t tests with non-pooled SD-DiastolicBP	
	High Risk	**Low Risk**
Low Risk	$< 2 * 10^{-16}$	
Mid Risk	$< 2 * 10^{-16}$	0,18

	Pairwise t tests with non-pooled SD-SystolicBP	
	High Risk	**Low Risk**
Low Risk	$< 2 * 10^{-16}$	
Mid Risk	$9{,}7 * 10^{-13}$	$5{,}5 * 10^{-10}$

	Pairwise t tests with non-pooled SD-BS	
	High Risk	**Low Risk**
Low Risk	$< 2 * 10^{-16}$	
Mid Risk	$< 2 * 10^{-16}$	$2{,}8 * 10^{-5}$

	Pairwise t tests with non-pooled SD-BodyTemp	
	High Risk	**Low Risk**
Low Risk	$< 5{,}7 * 10^{-6}$	
Mid Risk	1	$5{,}2 * 10^{-6}$

	Pairwise t tests with non-pooled SD-HeartRate	
	High Risk	**Low Risk**
Low Risk	$3{,}7 * 10^{-8}$	
Mid Risk	0,00023	0,09781

Figure 8: Pairwise T-tests for each variable

Hypothesis Testing

- *SystolicBP* and *BS* are seen to have means that are statistically different between all groups.They seem to be the ones who have more impact in the risk factor.
- For the remaining variables, there is at least one pair of groups where we don't reject the (null) hypothesis of the means being equal.

Information Theory

- Other method to assess the impact of our descriptive variables on the target variable is Information Theory.
- Impact is measured by entropy, which measures the "amount of information" present in a variable.
- The more certain a variable is about an event, the less information it will contain, and information here is the entropy (i.e. small value of entropy).

Entropy:

$$H(j) = \sum_{i=1}^{3} -p(x_i)\log_2(p(x_i)),$$

where $\mathrm{p}(\mathrm{x}_i) = \frac{\text{n. samples in class } i \text{ at node j}}{\text{n. total samples at node j}}$

Information Theory

With the formula for entropy we can now proceed in order to understand which variables give us less entropy. We will do this resorting to the following algorithm:

1. Calculate the initial entropy of the system.
2. Find the variable that further reduces the system's entropy.
3. Calculate the information gain (or entropy reduction).

Our initial entropy of the system is:
$H(\text{initial}) = -\frac{325}{896}\log_2(\frac{325}{896}) - \frac{276}{896}\log_2(\frac{276}{896}) - \frac{208}{896}\log_2(\frac{208}{896}) \simeq 1.54$

Information Theory

- This algorithm is the construction criterion of a well-known Machine Learning model: a Decision Tree

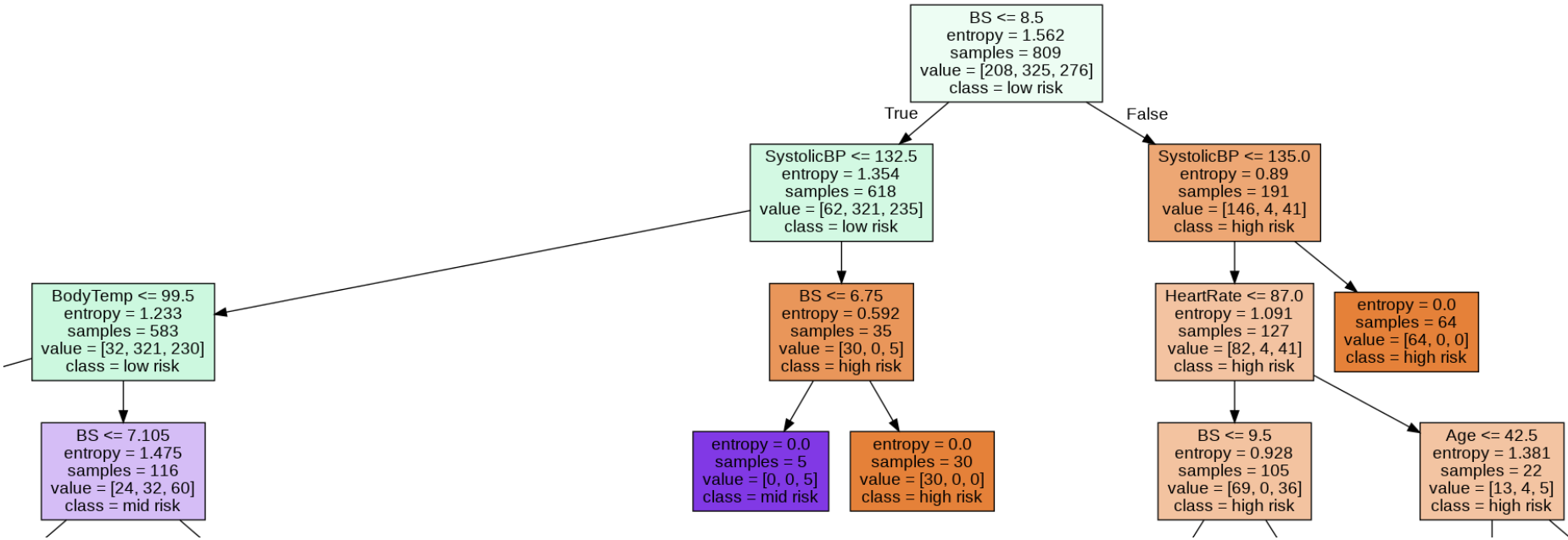


Figure 9: Top of the Decision Tree

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

Figure 10: 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)	0,8	0,82	0,81	
High Risk (n=60)	0,78	0,7	0,74	

Figure 11: Evaluation Metrics

Weighted Precision	Weighted Recall	Weighted F1-Score
0,83	0,83	0,83

Figure 12: Weighted Evaluation Metrics

Cumulative Link Models

- Predict and assess the importance of explanatory variables in a model based on prior observations.

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

. Where g is a transformation function, mapping probabilities to the real line and $P_j = P(Y \leq j)$.

Ordinal Logistic Regression

$$\text{logit}(P_j) = \log\left(\frac{P_j}{1-P_j}\right) = \beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{jp}X_p,$$

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:

1. No multi-colinearity;
2. proportional odds

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

with $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ and $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ column vectors.

Ordinal Logistic Regression

We can also define the odds of being less than or equal to a particular category j as $Pj/(1 - Pj)$.

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

where $\lambda_j = e^{\theta_j}$.

Ordered Probit Regression

- Results from modeling the probit of the cumulative probabilities as a linear function of the covariates.

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

where Φ is the standard normal cdf.

Modeling and results

- Pair correlation values and Variance Inflation Function (VIF) to check if the assumptions hold.
- Mathematically, the VIF equals the ratio of the overall model variance to the variance of a linear model that includes only that single independent variable.
- High values indicate that it is difficult to assess accurately the contribution of predictors to a model.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
VIF	1,43	2,83	2,76	1,45	1,15	1,06

Figure 13: VIF associated to each variable

Correlation Matrix

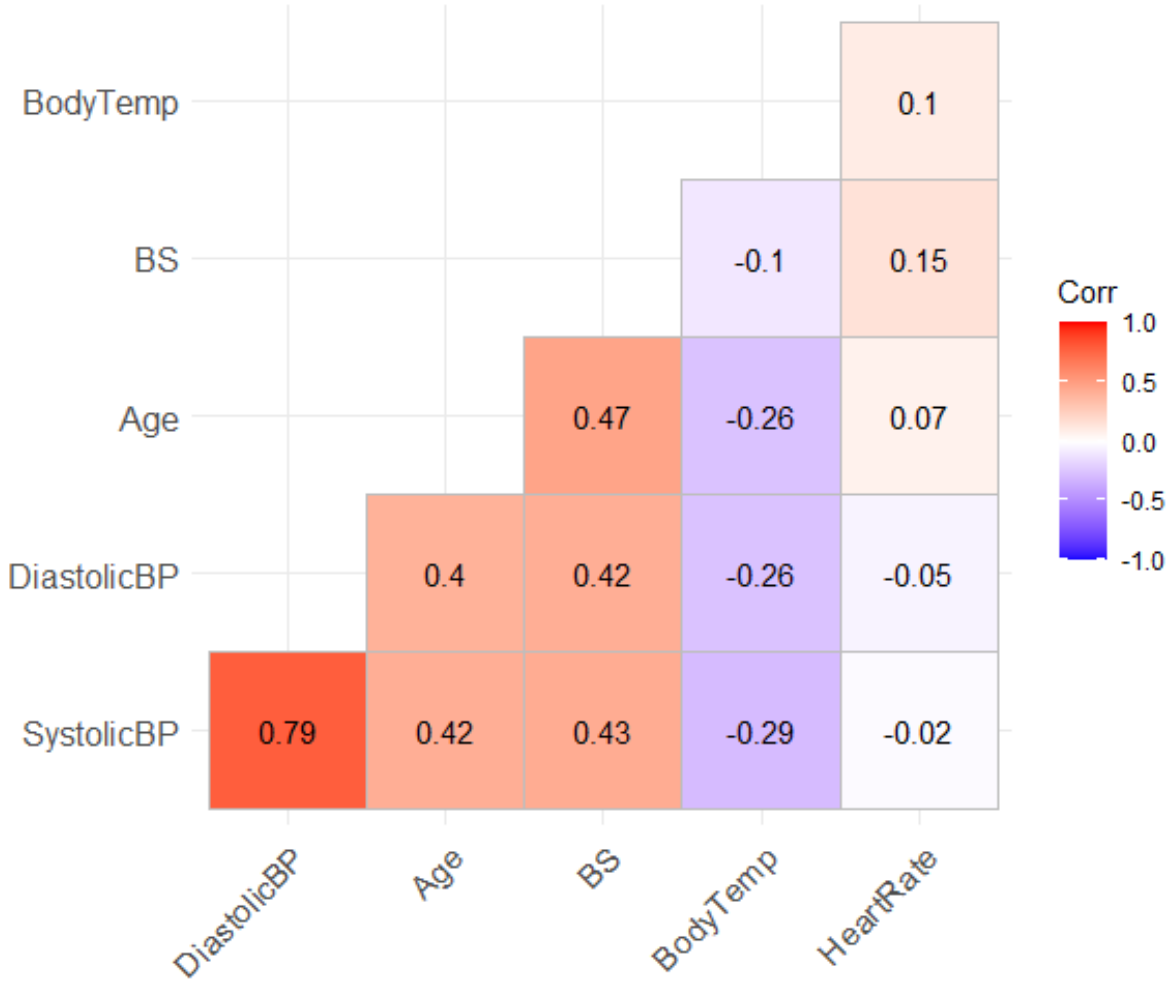


Figure 14: Correlation of the variables

Brant Hypothesis Test

	Brant Hypothesis test						
	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

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

OLR results

	Age	SystolicBP	BS	BodyTemp	HeartRate	Low/Mid	Mid/High
2,5% IC	-0,0258	0,0349	0,3883	NA	0,0179		
97,5% IC	0,0007	0,0557	0,5445	NA	0,0593		
Coef	-0,012	0,045	0,463	0,458	0,039	55,96	58,276

Figure 16: Confidence Intervals when profiling the likelihood function - OLR model

	Age	SystolicBP	BS	BodyTemp	HeartRate
2,5% IC	-0,0255	0,0355	0,3856	0,438	0,0178
97,5% IC	0.0007	0,0547	0,5413	0,4784	0,0592

Figure 17: Confidence Intervals when assuming normal distribution - OLR model

	Age	SystolicBP	BS	BodyTemp	HeartRate
OR	0,9877	1,0462	1,5896	1,5812	1,0393
2,5% IC	0,9749	1,0362	1,4704	1,549	1,018
97,5% IC	1,0007	1,0562	1,7183	1,613	1,061

Figure 18: Odds ratio and 95% confidence intervals - OLR model

OPR results

	Age	SystolicBP	BS	BodyTemp	HeartRate	Low/Mid	Mid/High
2,5% IC	-0,0147	0,0199	0,2251	NA	0,0104		
97,5% IC	0,0009	0,0316	0,3071	NA	0,0343		
Coef	-0,007	0,026	0,265	0,267	0,022	32,536	33,894

Figure 19: Confidence intervals when profiling the likelihood function - OPR model

	Age	SystolicBP	BS	BodyTemp	HeartRate
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

Figure 20: Confidence intervals assuming Normal distribution - OPR model

Predictions OLR

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

Figure 21: Prediction for Ordinary Logistic Regression

	Recall	**Precision**	**F1-Score**	**Global Accuracy**
Low Risk (n=64)	0,82	0,7	0,76	0,64
Mid Risk (n=79)	0,45	0,41	0,43	
High Risk (n=60)	0,58	0,84	0,69	

Figure 22: Evaluation Metrics of the Ordinal Logistic Regression model

Weight.Precision	**Weight.Recall**	**Weight.F1-Score**
0,63	0,61	0,61

Figure 23: Weighted Evaluation Metrics of the Logistic Regression model

Predictions OPR

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

Figure 24: 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	0,43	
High Risk (n=60)	0,58	0,84	0,69	

Figure 25: Confusion Matrix for the Ordered Probit Regression model

Weight.Precision	**Weight.Recall**	**Weight.F1-Score**
0,63	0,6	0,6

Figure 26: Weighted Evaluation Metrics of the Ordered Probit Regression model

Conclusions

- Our goal was to assess which variables played a major role in the dynamic of the risk level associated with pregnant women.
- Robust PCA $->$ 2 outliers were detected, which lead these two observations to be discarded.
- PCA $->$ some separability between the different risk groups. Corroborated the natural tendency of the human body.
- *Levene´s test* and *Fligner-Killeen´s test* $->$ low p-value indicated that there was no variance homogeneity across groups, except for *Age*.
- Non parametric Kruskal-Wallis test $->$ indicated that we should reject the null hypothesis, that for each level of Risk, the means are the same.

Conclusions

- Pairwise Welch's t-tests(with *bonferroni correction*) $->$ *SystolicBP* and *BS* turned out to be the most significant in the risk factor.
- Information Theory $->$ Trained a decision tree based on entropy criteria and verified that *BS* and *SystolicBP* have the highest impact on maternal mortality risk.
- Ordinal Logistic Regression and Ordered Probit Regression $->$ Again showed us that the variables *BS* and *BodyTemp* have a high influence in the Risk Level.
- Knowledge to apply in future endeavours.

Thank you for your time!