PREDICTING MATERNAL HEALTH-RISK

Big Data Analytics Course

PRESENTED

Alessia De Liddo, Emanuela Pia Mastrorilli

Politecnico di Bari

Index

Introduction

Analysis

Objectives

Results

Introduction

Data Source

Data has been collected by **Daffodil International University** from different hospitals, community clinics, maternal health cares from the rural areas of Bangladesh through the **IoT based risk** monitoring system. The dataset collects significant risk factors for maternal mortality, that is one of the main concern of Sustainable Development Goals of ONU.

Data Set

The health factors collected can directly cause higher levels of complications during pregnancy resulting in loss of women's pregnancy and death of both woman and baby.

1.013 **INSTANCES**

6 **FEATURES**

Age	Age in years when a woman is pregnant.
SystolicBP	Upper value of Blood Pressure in mmHg, another significant attribute during pregnancy.
DyastolicBP	Lower value of Blood Pressure in mmHg, another significant attribute during pregnancy.
BS	Blood glucose levels is in terms of a molar concentration, mmol/L
BodyTemp	Pregnant women's body temperature.
HeartRate	A normal resting heart rate in beats per minute.
RiskLevel	<High risk, Mid risk, Low risk> Predicted risk intensity level during pregnancy.

Objectives

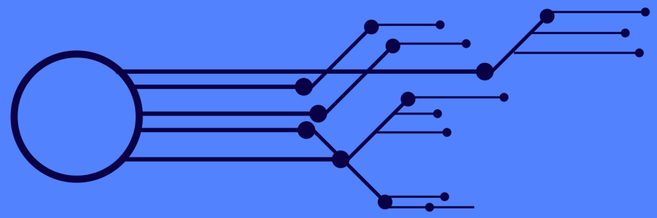
Exploratory analysis

Performed to **study health conditions** that act as stronger **indicators** to predict several maternal health risks during pregnancy.

Model classifier

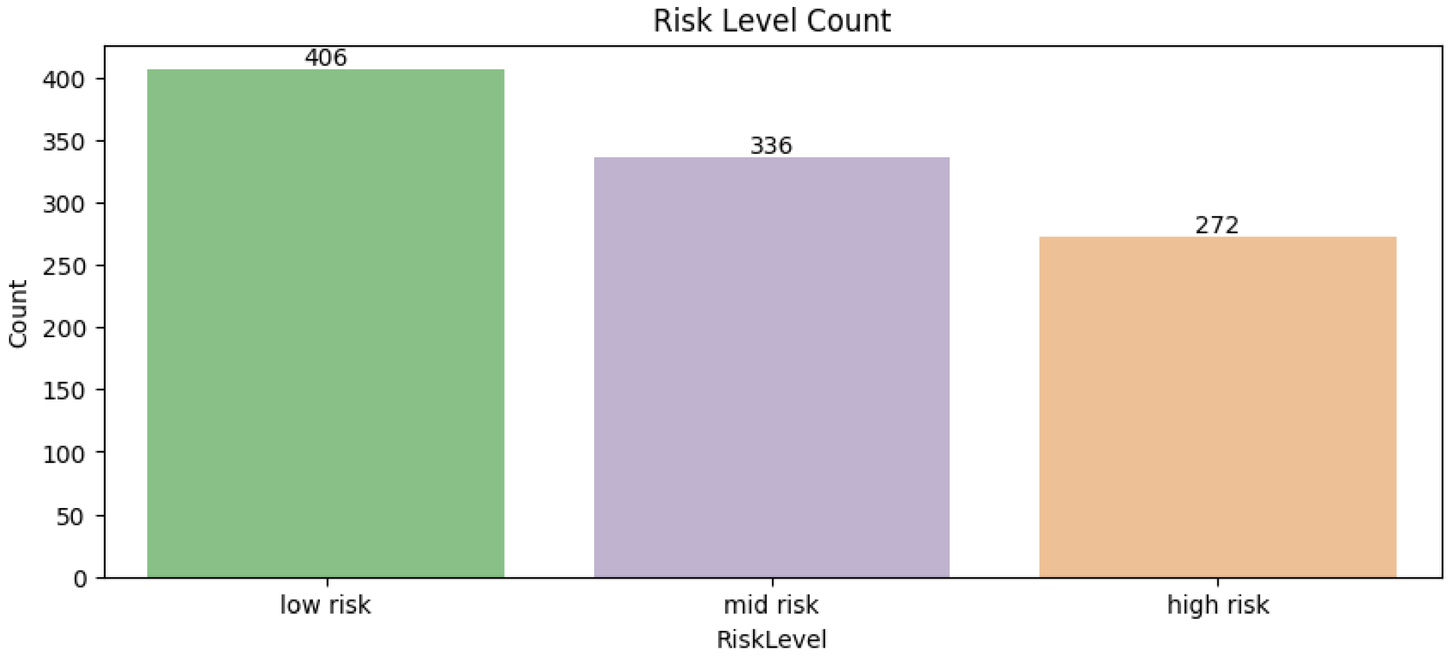
Models and approaches for the **timely prediction** of health risks during pregnancy using machine learning techniques.

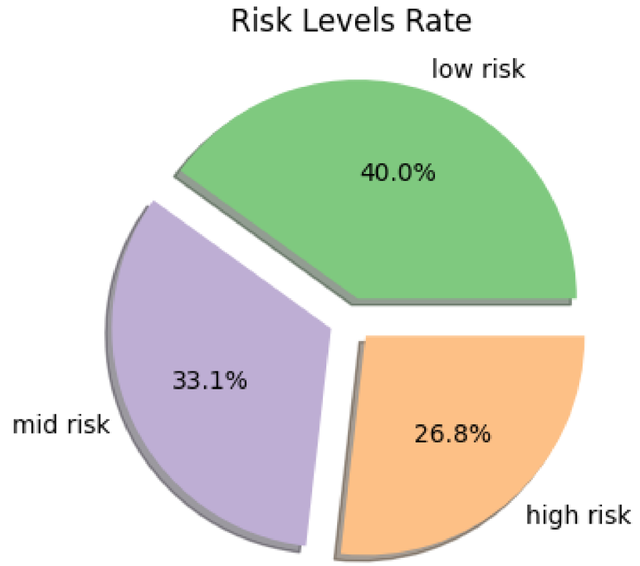
Analysis



Exploratory Data Analysis → Data Preprocessing → Modelling → Interpretability

Target Variable: Risk Level





Most instances are categorized as "Low Risk" (406 cases), suggesting typical health characteristics and a reduced probability of complications during pregnancy.

A considerable number of cases are designated as "Mid Risk" (336 cases), indicating a moderate level of risk that necessitates extra monitoring.

A total of 272 cases are identified as "High Risk," indicating an elevated risk level that requires vigilant monitoring and potential intervention.

Feature Variables

General Age distribution
20:40 years

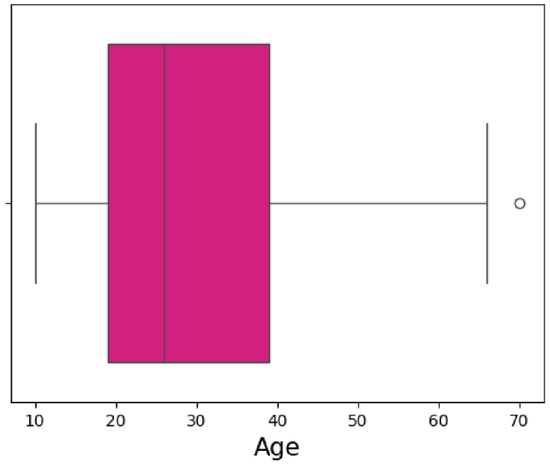
SystolicBP
100:120 mmHg

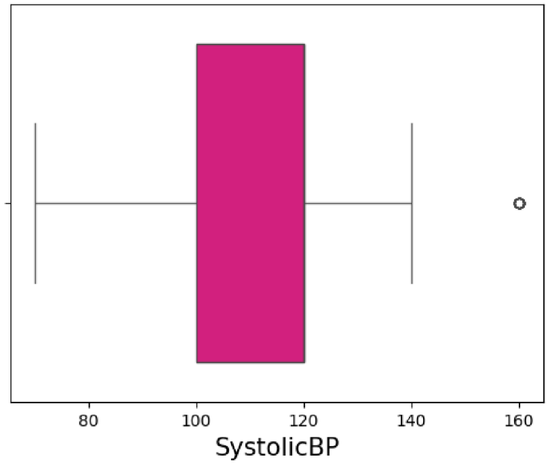
DiastolicBP
60:90 mmHg

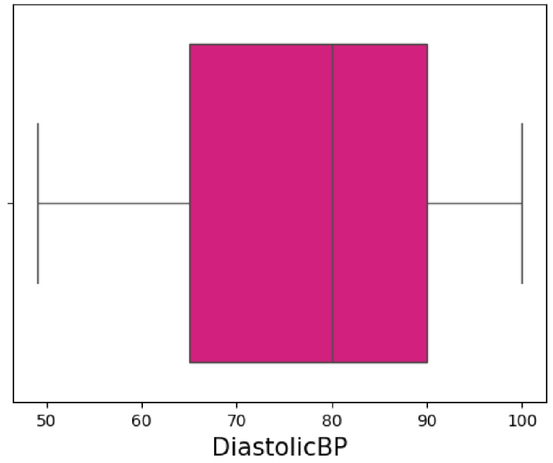
BS
7-8 mmol/l

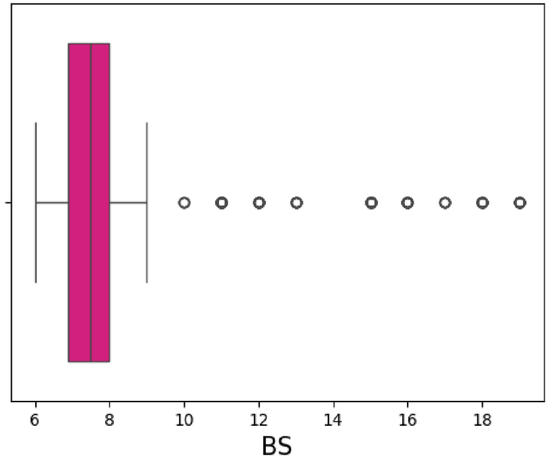
BodyTemp
98 F (36.6°)

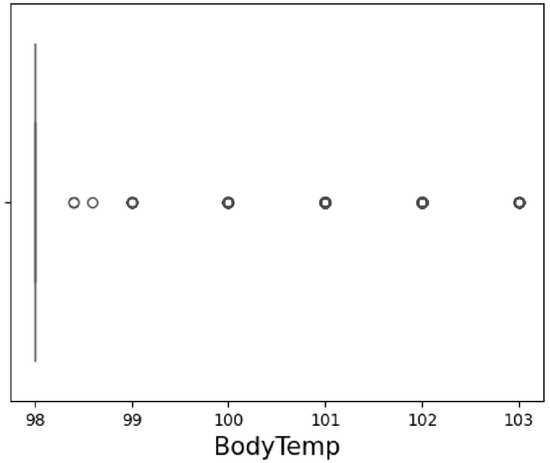
HeartRate
70-80 bpm

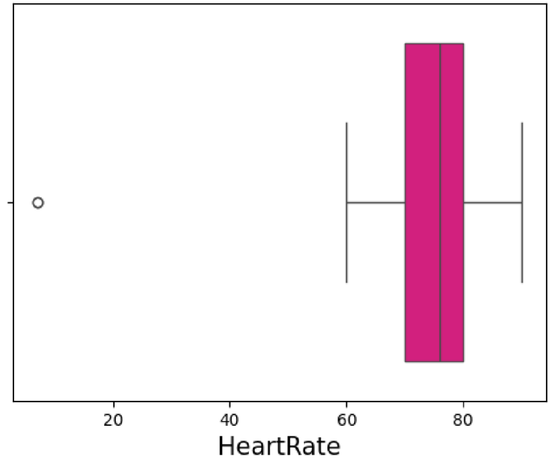




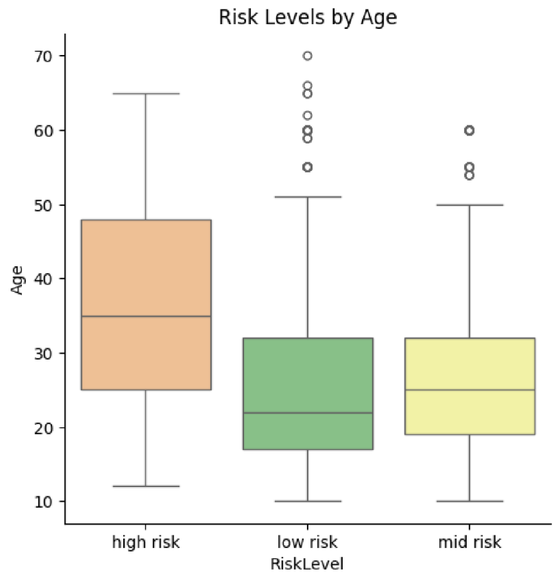


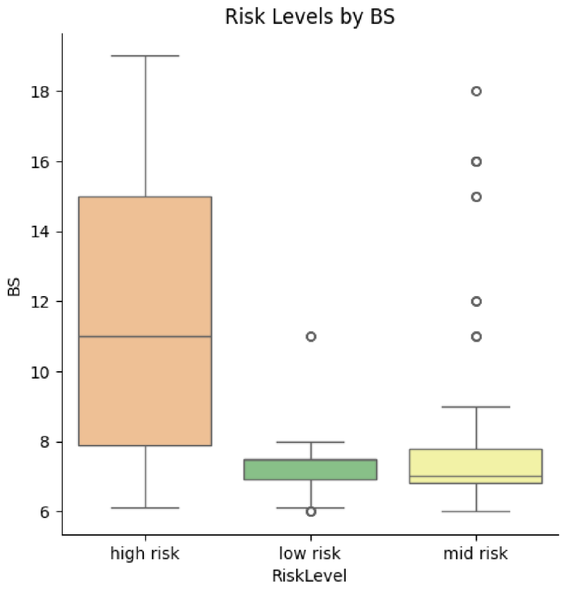


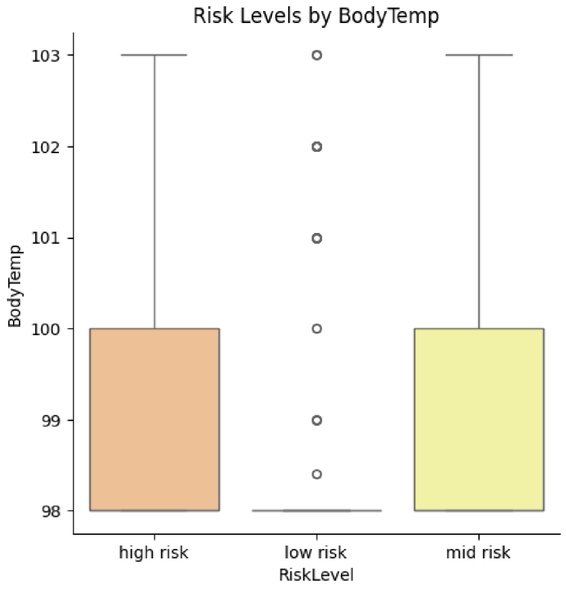


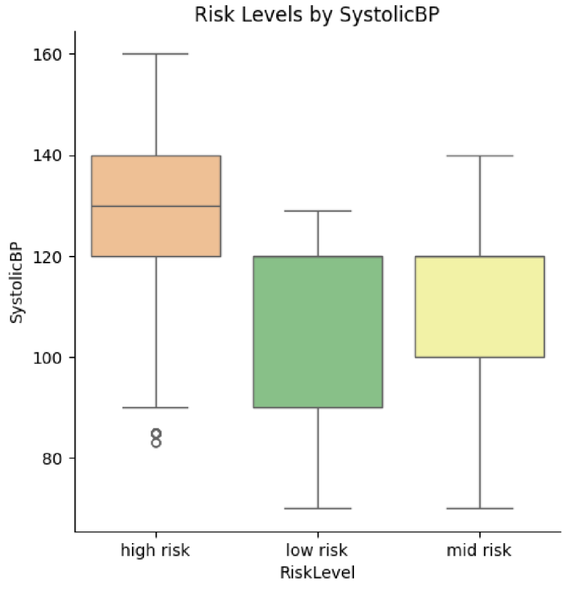


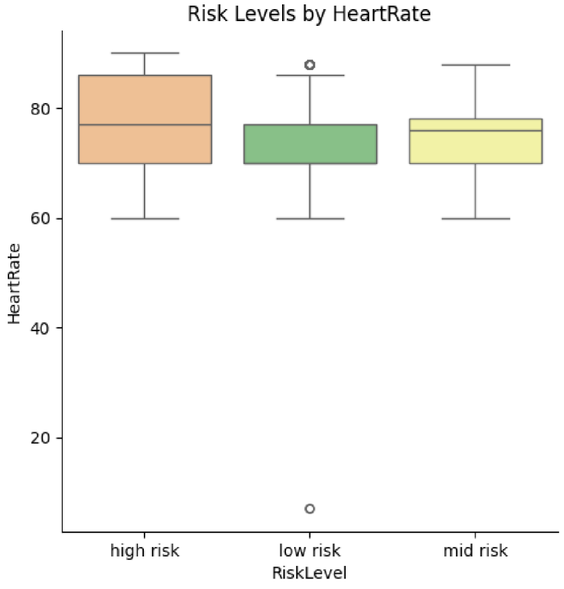
Detection of Outliers!

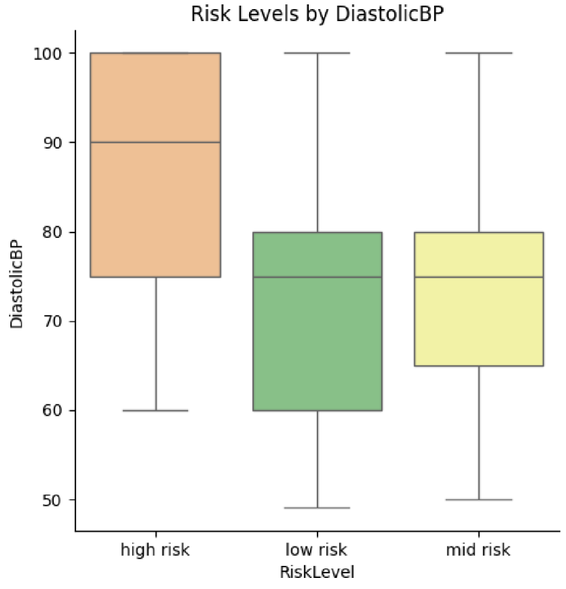












Qualitatively it seems there is a rising linear correlation between values and risk levels, except for body temperature. Notably, the heart rate appears to be decisive for values surpassing a specific threshold.

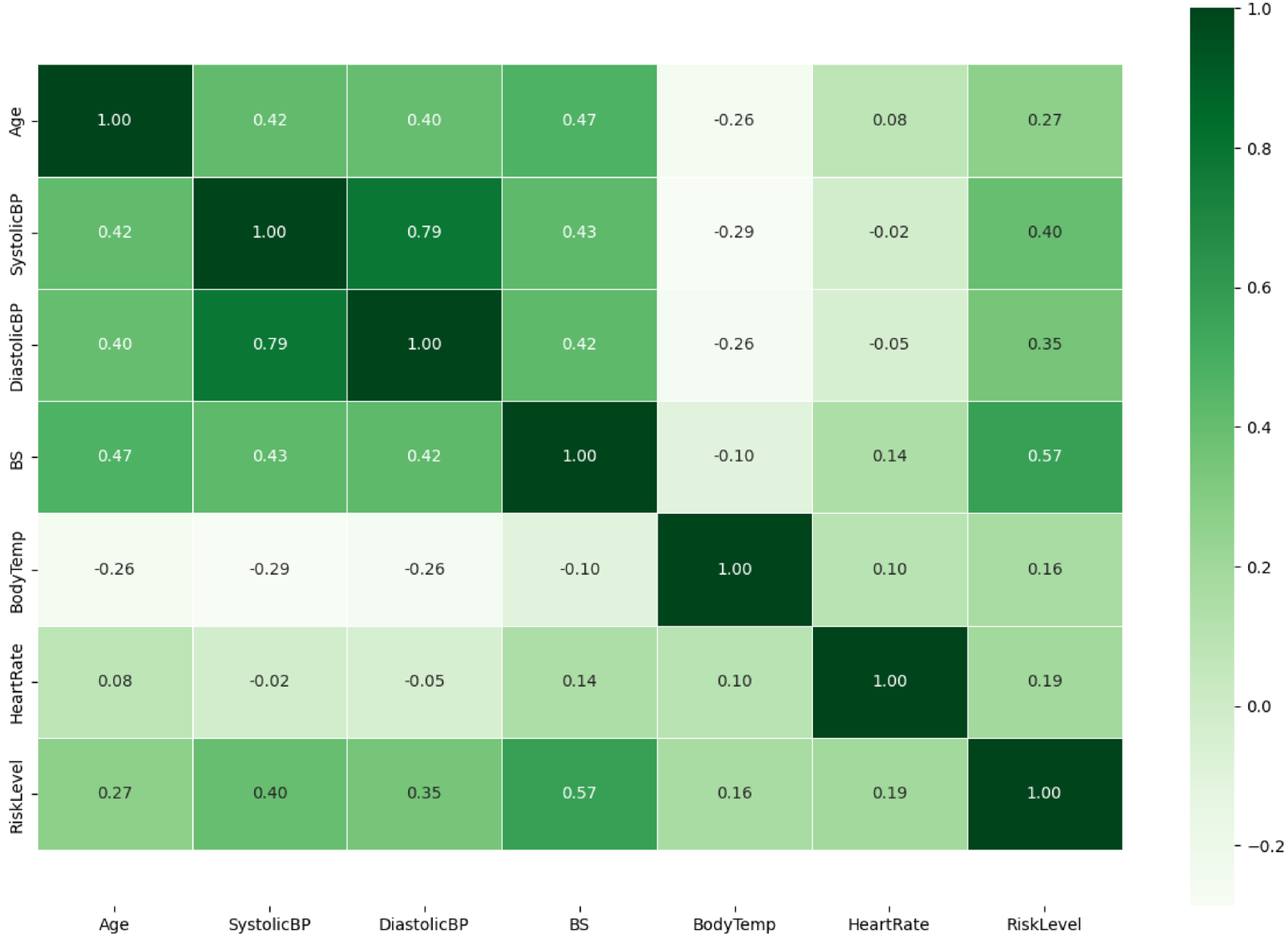
Correlation Analysis

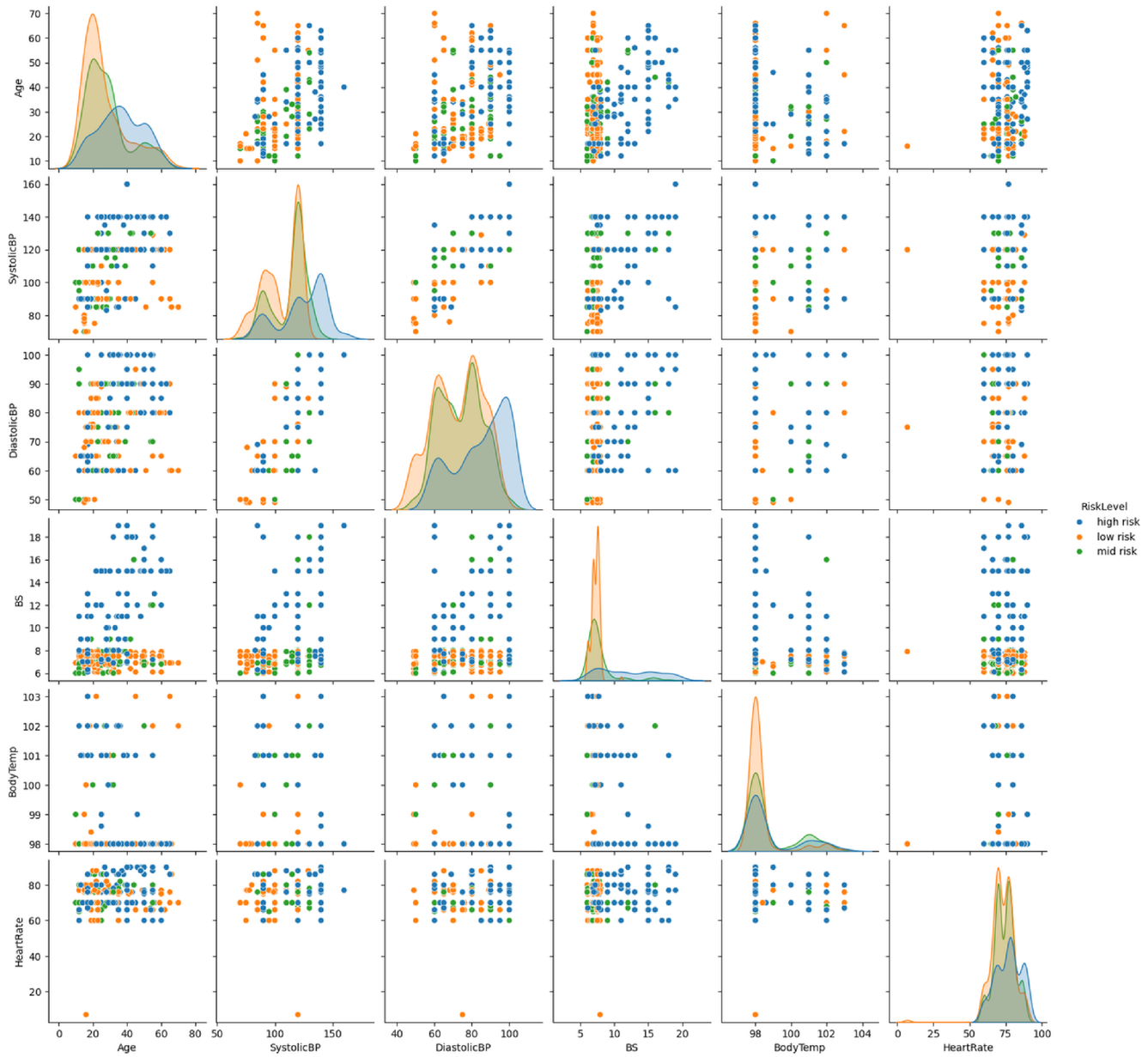
The correlation matrix says that DiastolicBP and SystolicBP are highly correlated.

Blood pressure and blood sugar are age-related.

Also the level of risk depends mainly on blood sugar, then blood pressure and age.

Heart rate, body temperature and age are less effective at risk here.





Age

In general, pregnant women below 24 years typically show a lower health risk, with risks tending to increase from the age of 25 onward. However, there is a decline in risk for women aged 59 and above.

Systolic&Diastolic BP

A significant correlation exists between Systolic Blood Pressure (SystolicBP) and Diastolic Blood Pressure (DiastolicBP). Elevated values for both indicators are associated with increased health risks. It's important to note that low SystolicBP and DiastolicBP at a young age do not necessarily guarantee a low health risk.

Blood Glucose (BS)

The data suggests that pregnant women with blood glucose levels equal to or exceeding 8 are more prone to elevated health risks, irrespective of other variables recorded in the dataset.

BodyTemp

The majority of pregnant women maintain a body temperature around 36.7°C, falling within the normal range. Elevated body temperatures (>=37.8°C) are associated with increased health risks. Younger women with normal body temperatures typically show lower health risks, whereas the influence of high body temperatures on health risks seems less prominent in older women. Additionally, a normal body temperature coupled with low Systolic Blood Pressure and Diastolic Blood Pressure is linked to a reduced health risk.

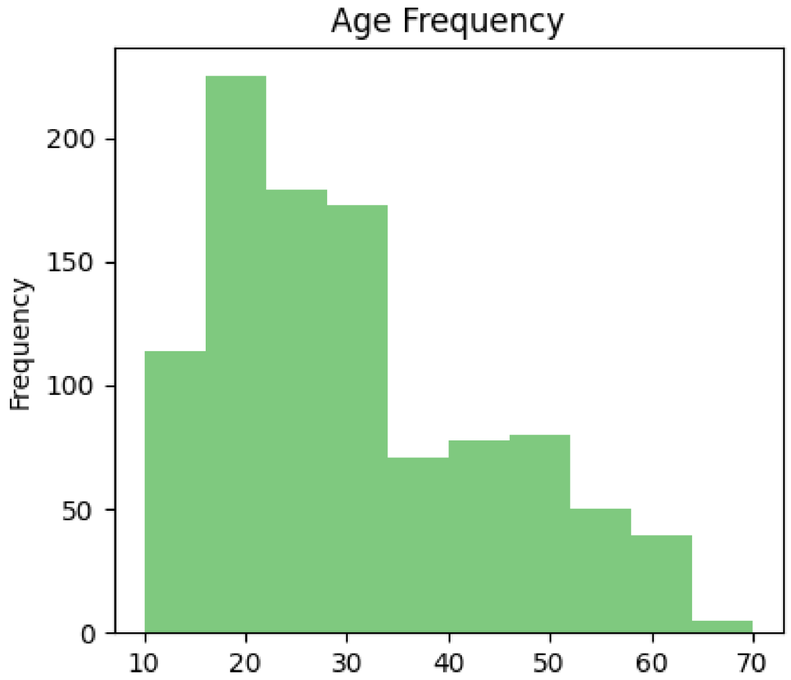
HeartRate

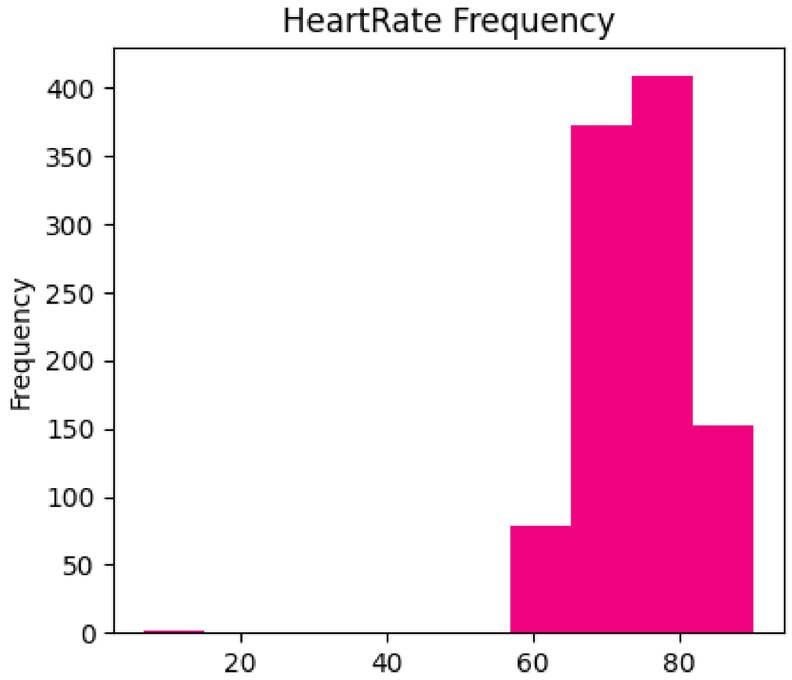
The analysis indicates a rising trend in health risks as heart rates increase. Nevertheless, when compared to other variables, it seems that heart rate has a relatively modest impact on the health risks of pregnant women.

Data Preparation

- Data cleaning
- Splitting sets
- Standardization

Data cleaning





Age Frequency

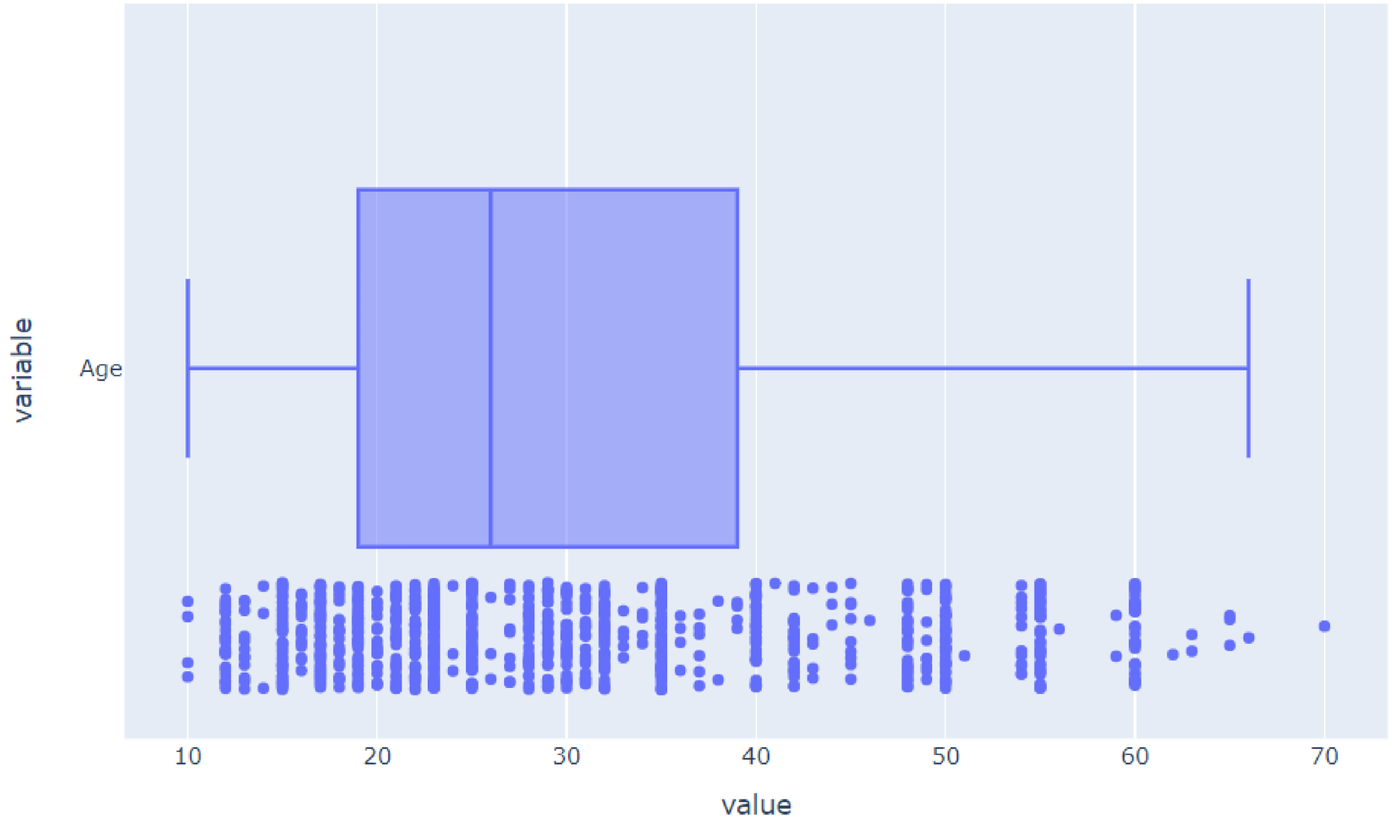
Some values, such as ages 10 or 70, appear to be irrelevant for pregnancy data. The majority of observations are concentrated in the 20 - 30 age range, followed by the 10 - 20 age range.

HeartRate Frequency

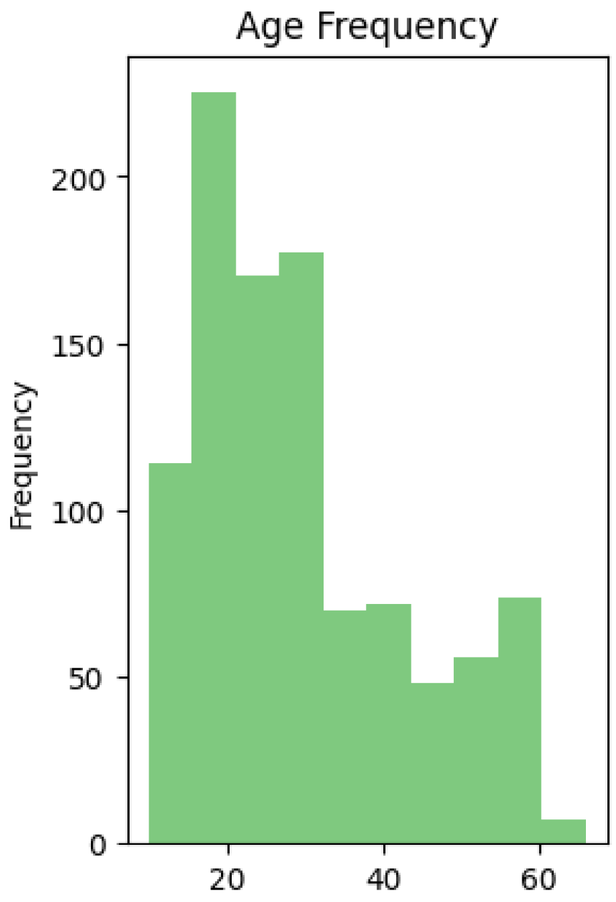
The heart rate value "7" is evidently an outlier, likely entered in error.

Age Frequency

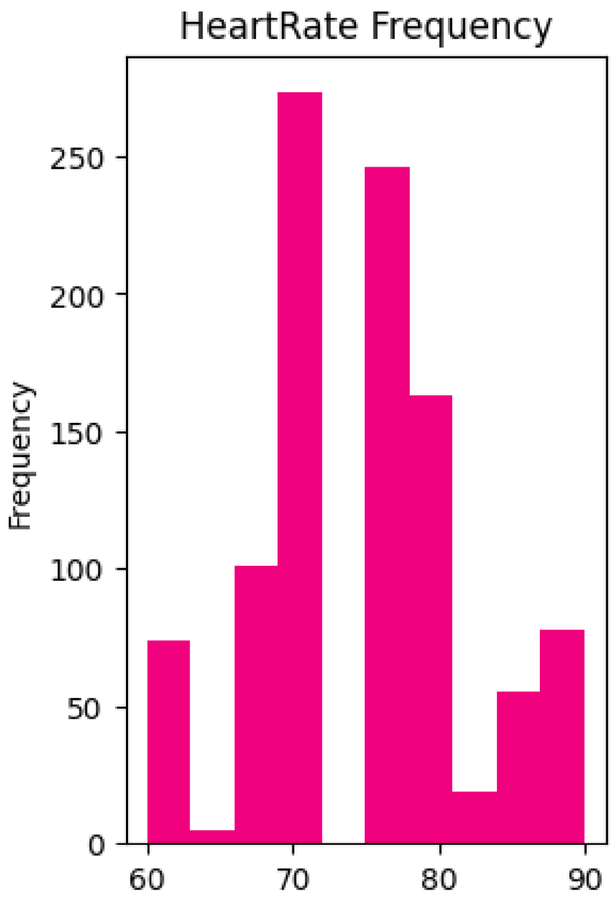
- The risk of maternal mortality for girls under 15 in low- and middle-income countries is double that of girls over 15 great.
- About 70,000 adolescents in developing countries die every year from causes that can be linked to pregnancy and childbirth: this is one of the first causes of death. Childhood marriages widespread in areas of Bangladesh justify the values found.



https://aidos.it/wp-content/uploads/2013/10/UNFPA2013-completo-def.pdf



Outlier removed



Outlier modified

Splitting sets

Standardization

Logistic Regression Classifier

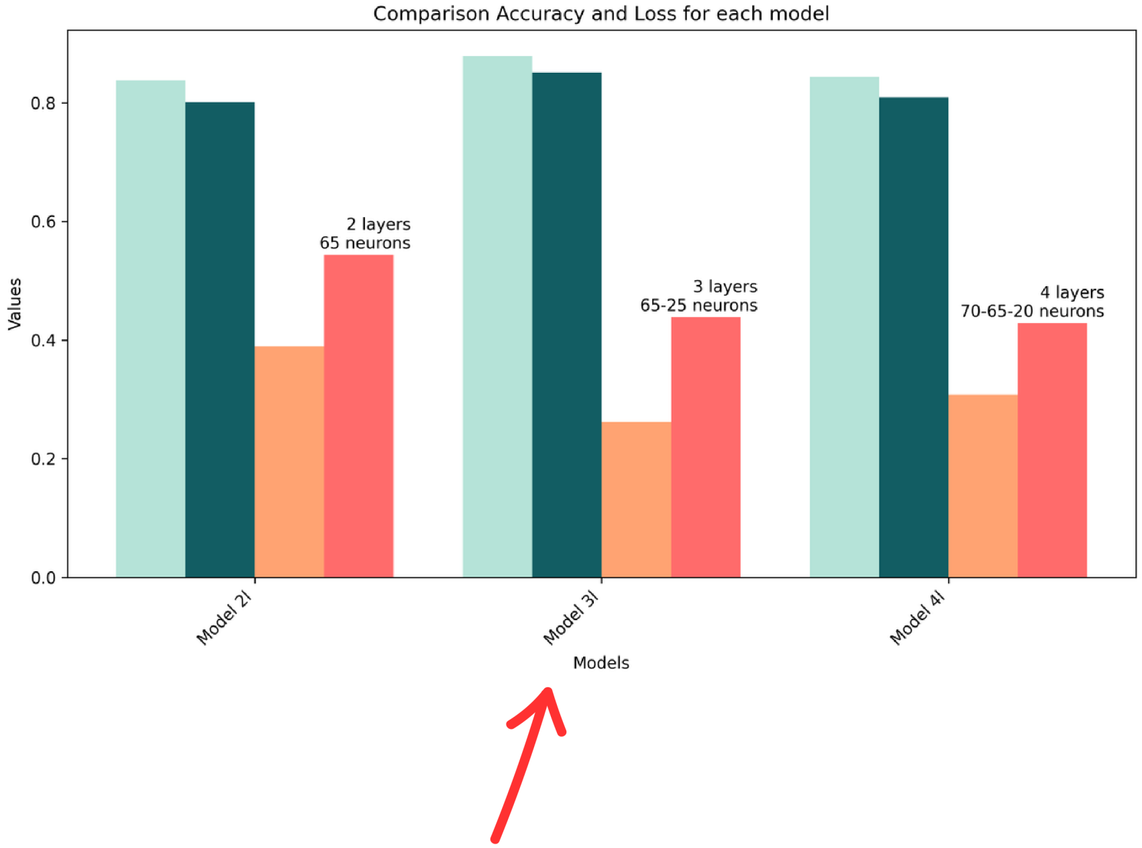
- Baseline Logistic Regression Model
- Grid Search for Logistic Regression Parameters

Deep Neural Network

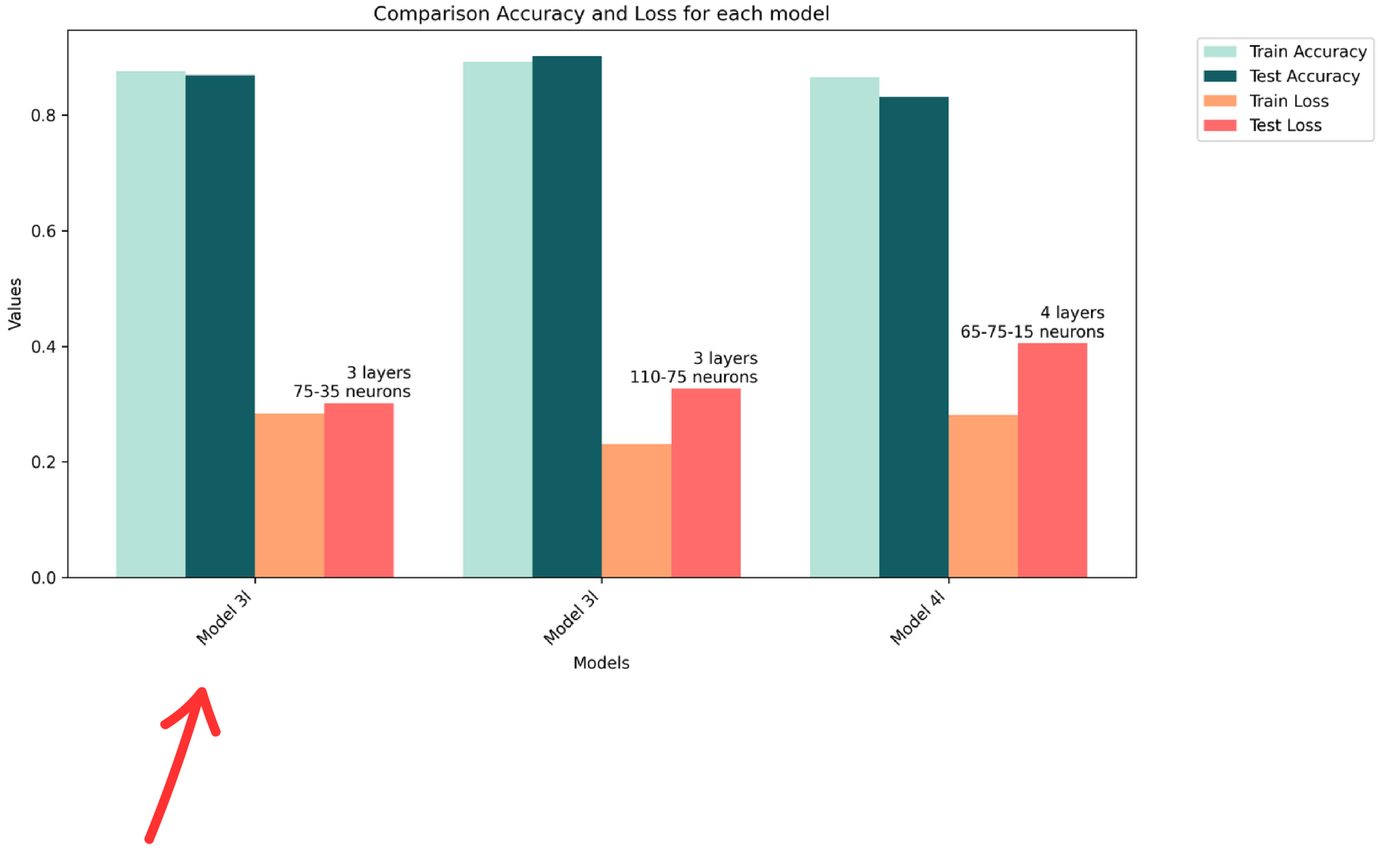
- Grid Search for Hyperparameters
- Definition ANN

ANN Model Selection

PRE balancing and data generation

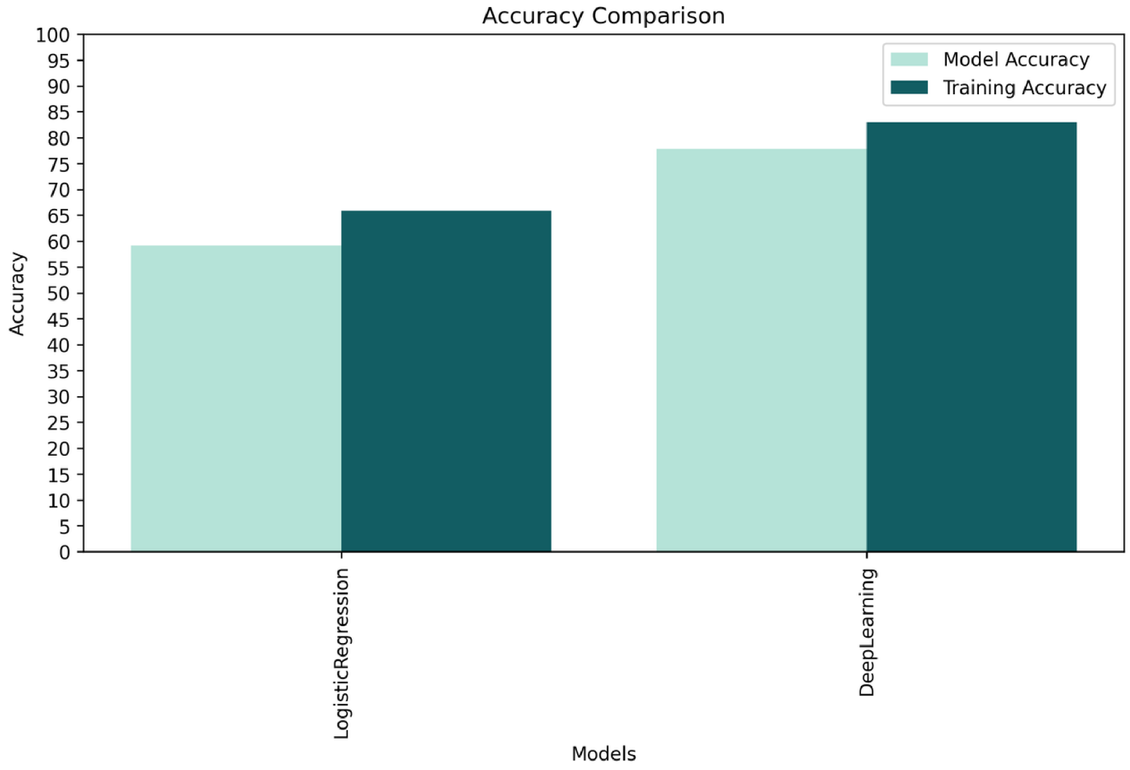


POST balancing and data generation

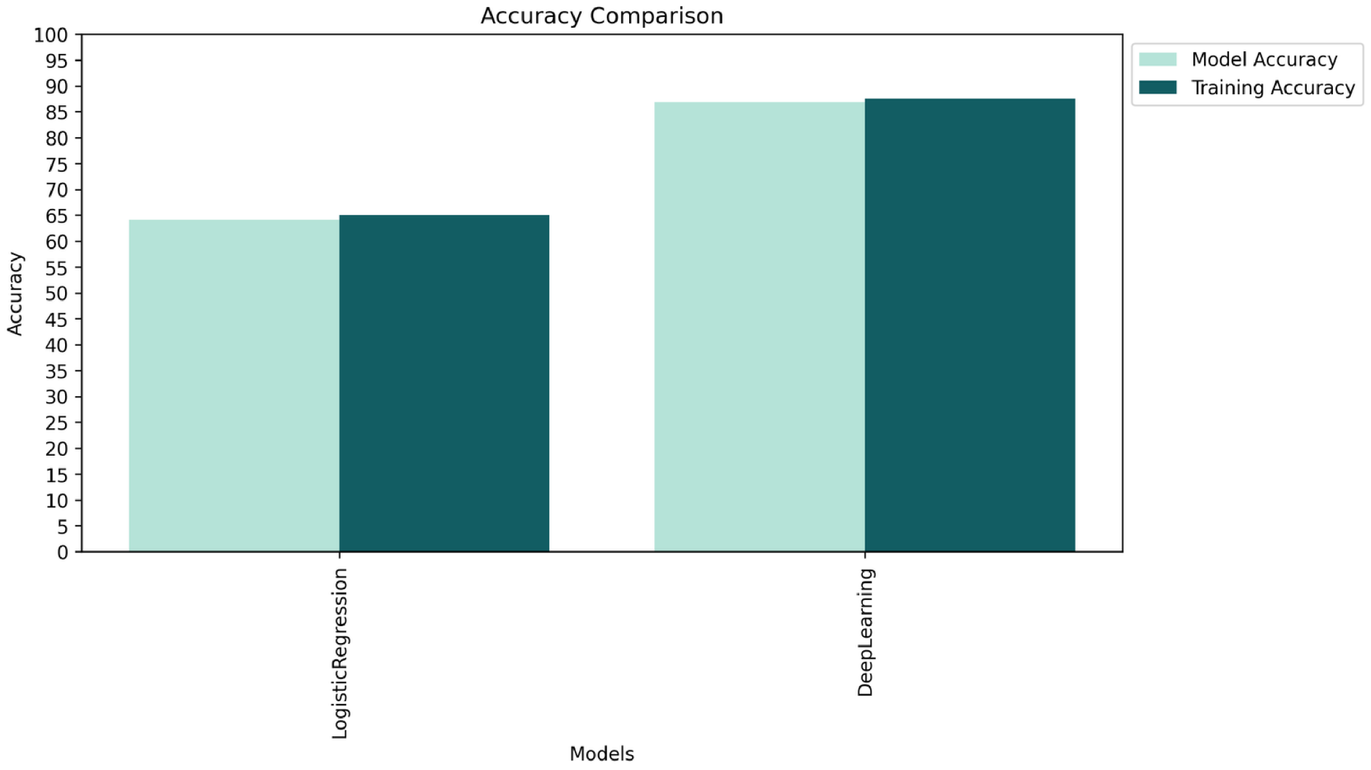


Accuracy Comparison

PRE balancing and data generation

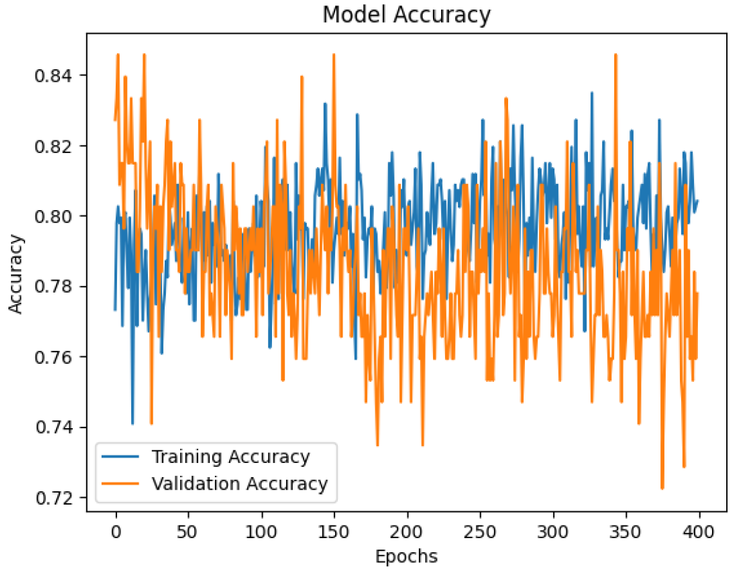


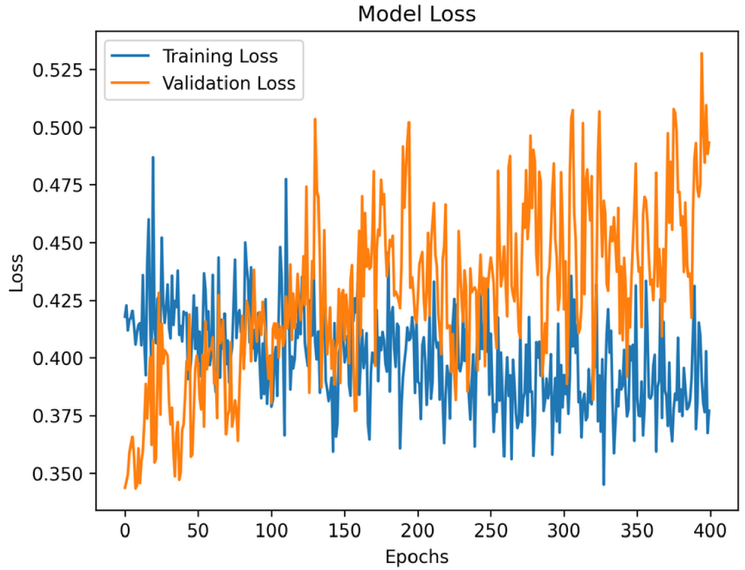
POST balancing and data generation



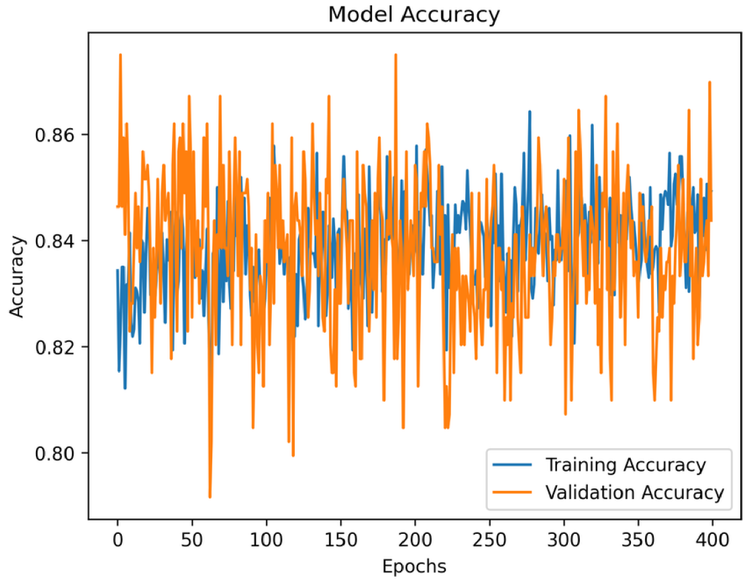
ANN History comparison

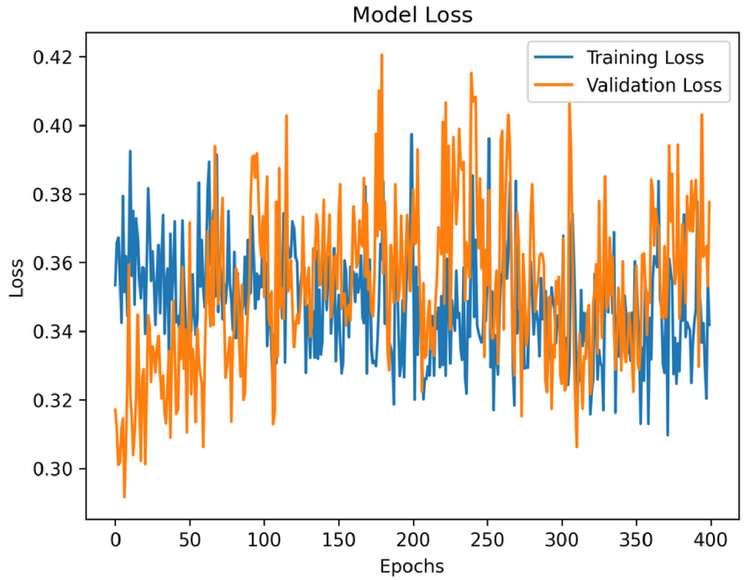
PRE balancing and data generation





POST balancing and data generation





Pre data generation and balancing

	LogReg Model	LogReg Model after GridSearch	DeepNeural Network
Accuracy	0.60	0.60	0.78

LR	Precision	Recall	f1-score	Support
Low risk	0.60	0.82	0.70	79
Mid Risk	0.39	0.31	0.35	64
High Risk	0.80	0.58	0.67	60

DNN	Precision	Recall	f1-score	Support
Low risk	0.76	0.76	0.76	78
Mid Risk	0.69	0.73	0.71	63
High Risk	0.91	0.85	0.88	62

Post data generation and balancing

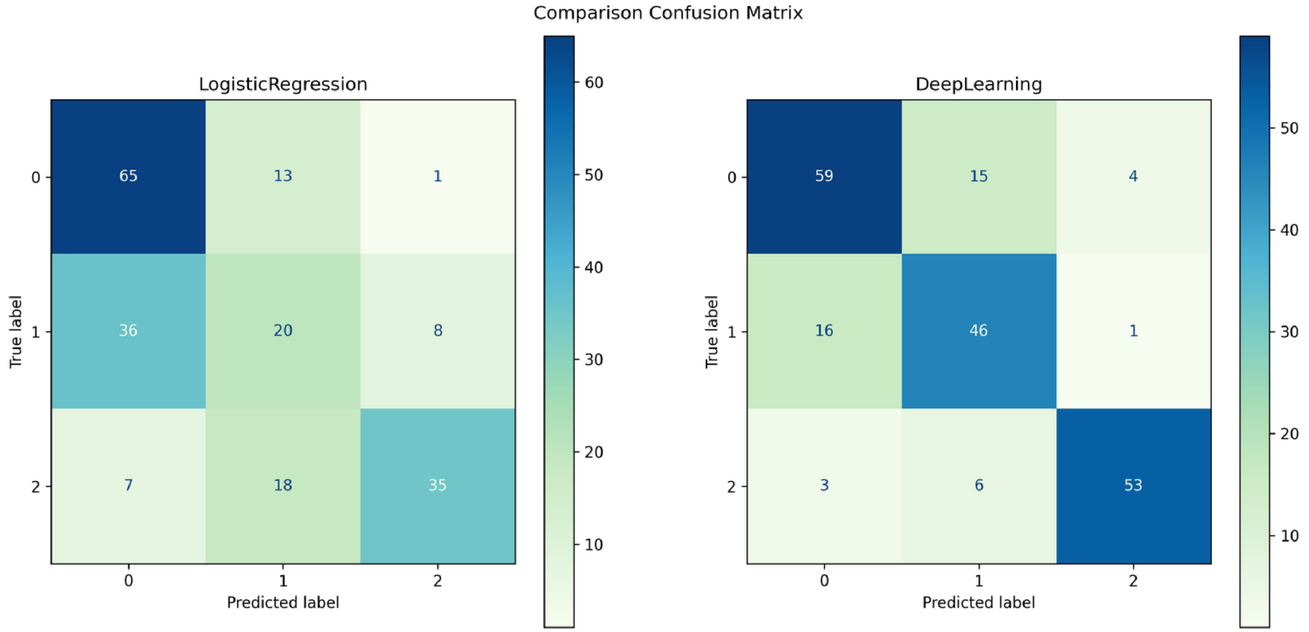
	LogReg Model	LogReg Model after GridSearch	DeepNeural Network
Accuracy	0.64	0.67	0.87

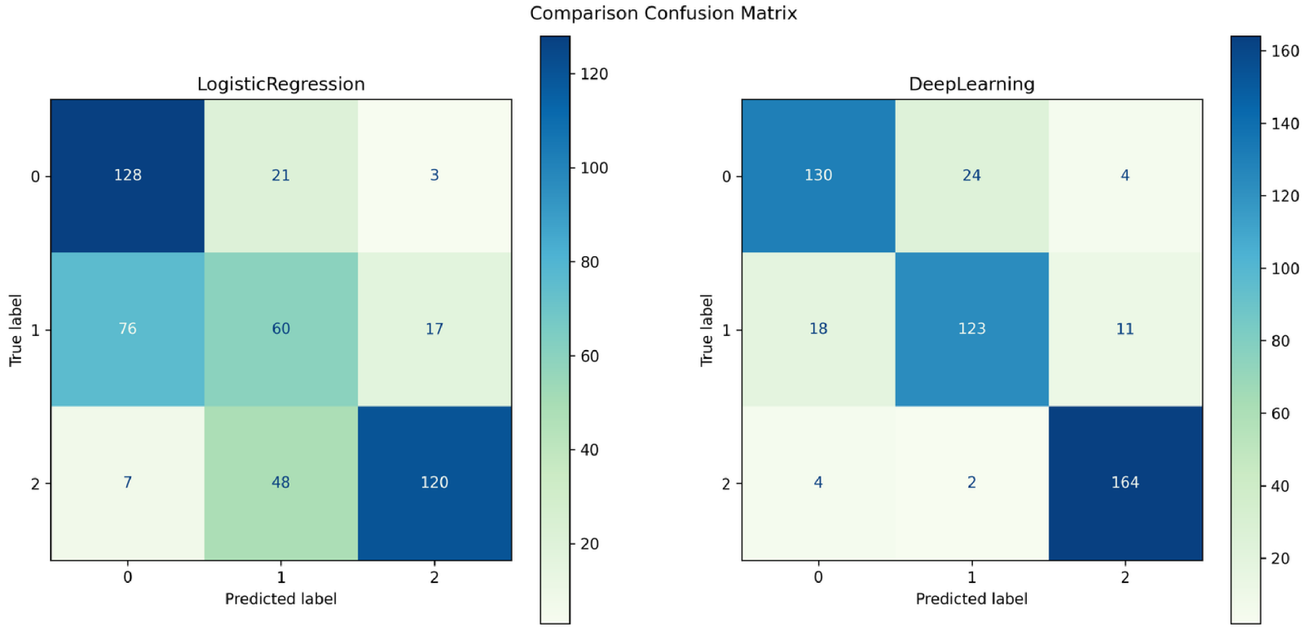
LR	Precision	Recall	f1-score	Support
Low risk	0.61	0.84	0.71	152
Mid Risk	0.47	0.39	0.43	153
High Risk	0.86	0.69	0.76	175

DNN	Precision	Recall	f1-score	Support
Low risk	0.86	0.82	0.84	158
Mid Risk	0.83	0.81	0.82	152
High Risk	0.92	0.96	0.94	170

Based on the confusion matrices:

- Logistic Regression more accurately predicts a lower level of risk.
- Overall, the Deep Learning model improves the prediction of risk, particularly high risk.
- Given the medical nature of the dataset, the most crucial predictions for us are those related to a high level of risk.





Conclusions

Logistic Regression

- Accuracy has slightly improved (0.60-0.64) since the implementation of SMOTE for balance and data generation.
- The confusion matrix indicates enhancements in the model's capacity to accurately classify high and mid risk levels following the introduction of balance through SMOTE and data generation.

Deep Neural Network

- Following the grid search, the accuracy has shown improvement, rising from 0.78 to 0.87, particularly with the incorporation of SMOTE (Synthetic Minority Over-sampling Technique).
- The confusion matrix reveals consistently positive outcomes both before and after the balancing process, indicating overall good performance in the model's predictions.

In both models, the application of SMOTE for balancing and increasing the dataset to positively influence the capability to accurately classify positive cases, as evidenced by the enhanced recall in particular for high-risk classes.

Thanks For The Attention

You can see the code here