Statistical Analysis in Maternal Health Conditions with Risk Factors

Group 22

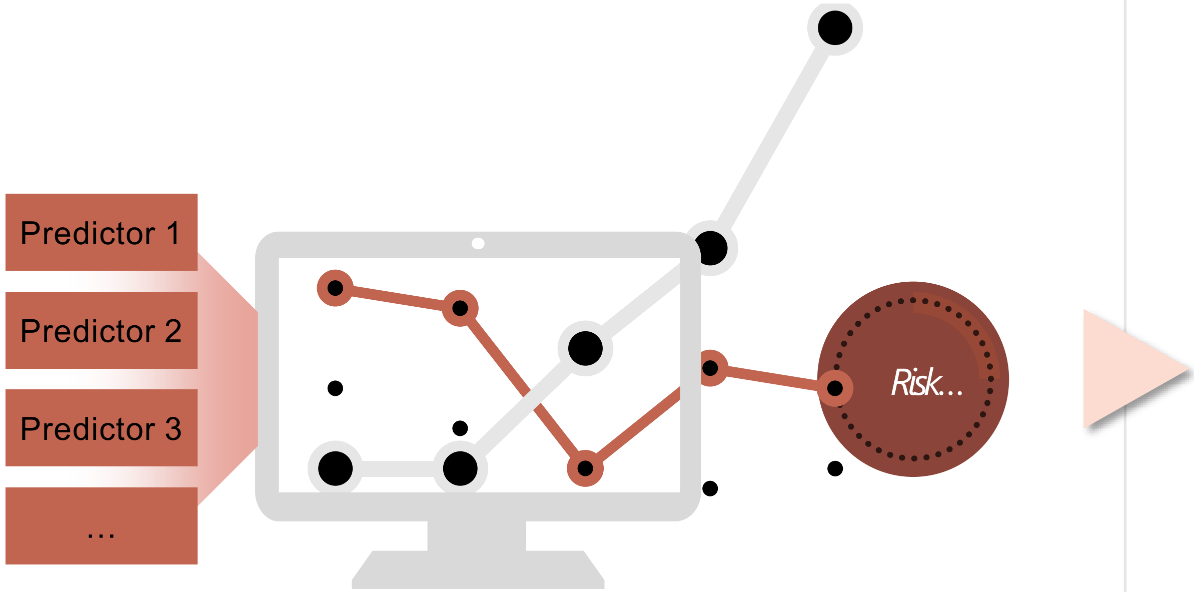
Ming Luo

Yingnan He

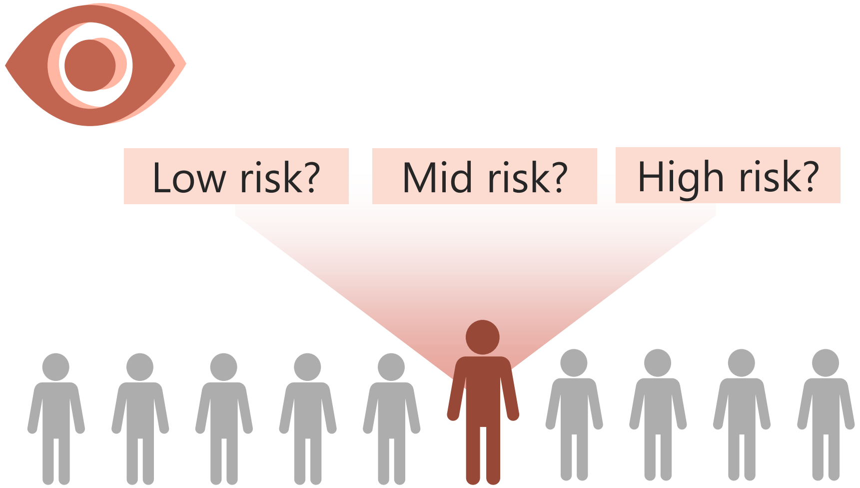
Introduction

Problem Background and Definition

- Explore the relationships between maternal risk level and medical parameters, then train the classification models



- Classify pregnant women into the maternal risk levels with medical risk factors



Overview

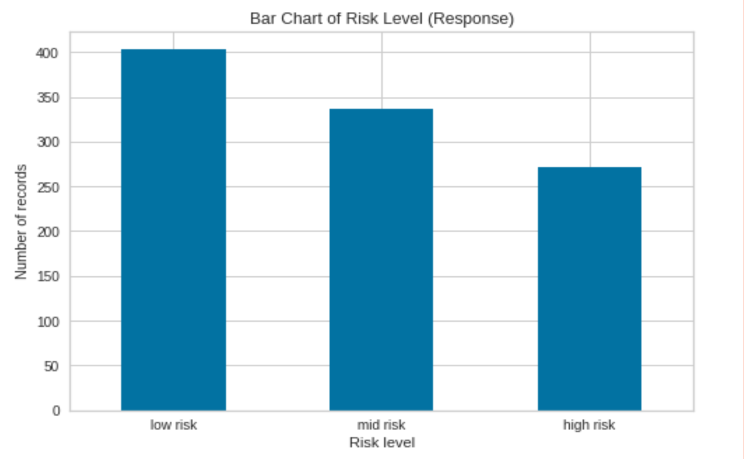
1. Data Collection and Exploration
2. Principal Component Analysis
3. Modeling and Model Selection
4. Model Performance Comparison

Dataset Overview

1014 Records, 6 Predictors, and 1 Response Variable

Input Variables	Response Variable
Numeric 1. Age (integer) 2. Systolic Blood Pressure (SystolicBP) (integer) 3. Diastolic Blood Pressure (DiastolicBP) (integer) 4. Blood Sugar (BS) (float) 5. Body Temperature (BodyTemp) (float) 6. Heart Rate (integer)	**Categorical** 1. Risk Level: Low, Mid, High (string)

Balanced Dataset

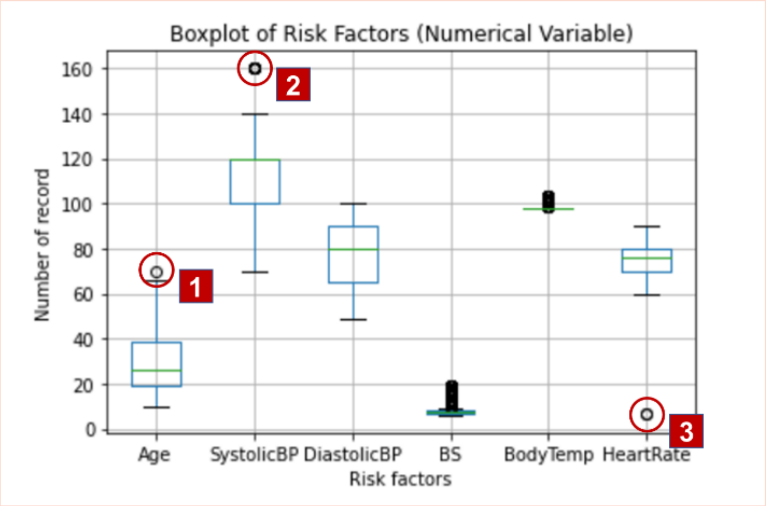


No Missing Values

	Number of Missing Values
Age	0
SystolicBP	0
DiastolicBP	0
BS	0
BodyTemp	0
HeartRate	0
RiskLevel	0

Data Processing

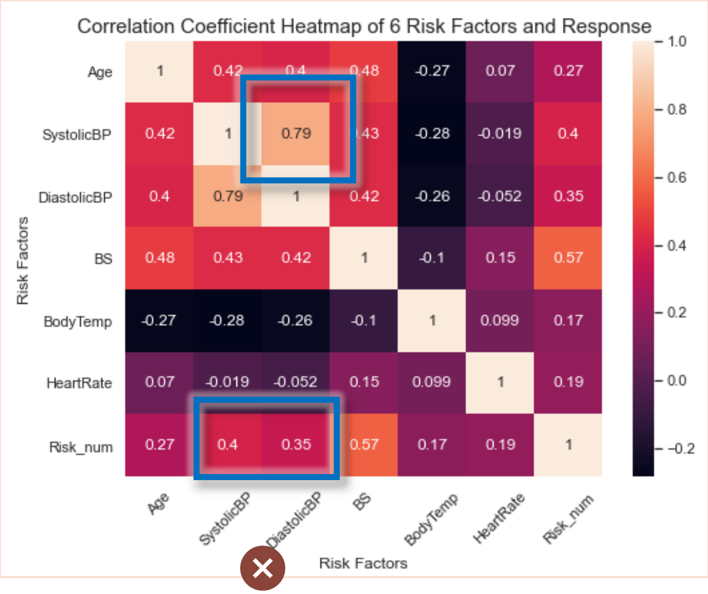
Drop Three Outliers



1. Age 70 and low risk level (impossible, 1 record)
2. SystolicBP with 160 and high risk (possible)
3. Heart Rate of 7 and low risk level (impossible, 2 records)

Therefore, we removed the 3 unreasonable records.

Drop One Highly Correlated Predictor



	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	Risk_num
Age	1	0.42	0.4	0.48	-0.27	0.07	0.27
SystolicBP	0.42	1	0.79	0.43	-0.28	-0.019	0.4
DiastolicBP	0.4	0.79	1	0.42	-0.26	-0.052	0.35
BS	0.48	0.43	0.42	1	-0.1	0.15	0.57
BodyTemp	-0.27	-0.28	-0.26	-0.1	1	0.099	0.17
HeartRate	0.07	-0.019	-0.052	0.15	0.099	1	0.19
Risk_num	0.27	0.4	0.35	0.57	0.17	0.19	1

1. Systolic BP and Diastolic BP are highly correlated
2. Systolic BP is more related to the response

Therefore, Diastolic BP is dropped.

PCA Variance & Weights

Variance

	PC1	PC2	PC3	PC4	PC5	PC6
Explained variance	2.618969	1.145773	0.840060	0.703358	0.486060	0.211720
Proportion of variance	0.436063	0.190773	0.139872	0.117110	0.080930	0.035252
Cumulative proportion	0.436063	0.626836	0.766708	0.883818	0.964748	1.000000

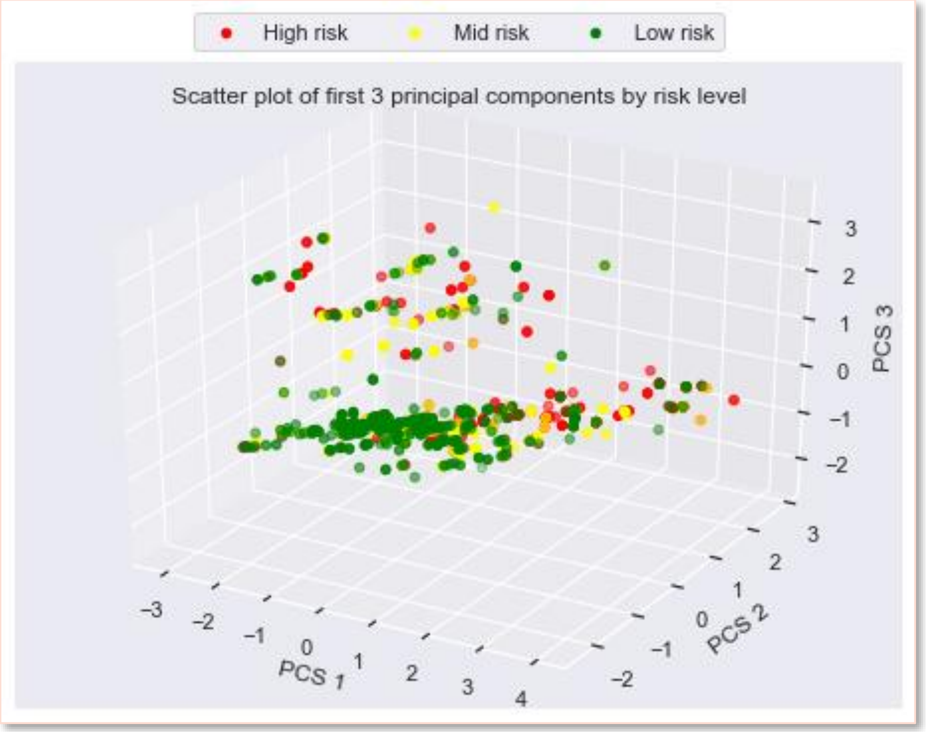
- First 2 PCs only capture 62.7% variance of predictors
- Even **first 3 PCs capture 76.7% variance**

Weights

	PC1	PC2	PC3	PC4	PC5	PC6
Age	0.439966	0.151309	-0.247566	0.548527	0.648796	-0.020696
SystolicBP	0.528636	-0.102061	0.248389	-0.365657	0.091943	0.711528
DiastolicBP	0.521171	-0.121386	0.310785	-0.348674	0.065910	-0.700815
BS	0.424525	0.361570	0.099046	0.433734	-0.700625	0.015314
BodyTemp	-0.273502	0.429063	0.804470	0.152421	0.264403	0.028074
HeartRate	0.018165	0.798202	-0.351347	-0.482164	0.074035	-0.033703

- PC1 is dominated by variables **age, blood pressure, as well as blood sugar.**
- PC 2 is dominated by variables **heartrate and body temperature**.

PCA Plotting



Plot data points on a 3D plane defined by the first 3 PCs

Plot data points on a 2D plane defined by the first 2 PCs

Because none of the first several components capture majority of variance in 6 predictors, **PCA might not be a very helpful tool** to predict the risk level of a pregnant woman.

Task specification and model selection

In this project, we are doing supervised learning.

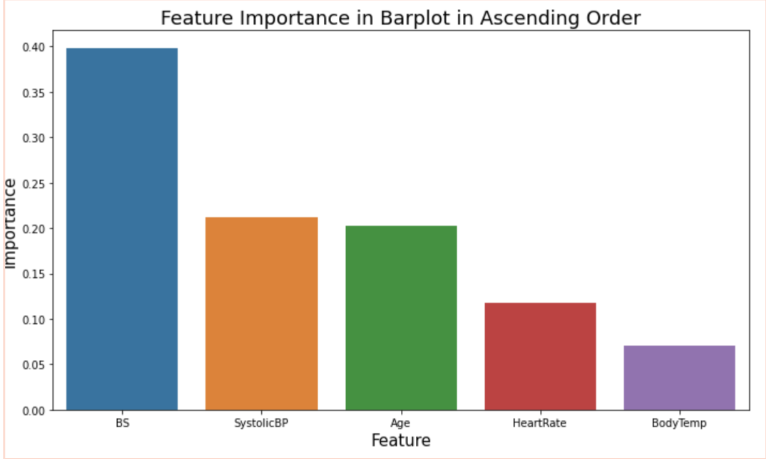
In addition, since all the predictors are numeric and the output variable is categorical, we applied the following classification models to the Maternal Health Risk dataset.

- **K-Nearest Neighbors with Random Forest**
- **Multinominal Logistic Regression**
- **Gaussian Naïve Bayes**
- **Decision Tree**
- **Artificial Neural Networks**

K-Nearest Neighbors with Random Forest

Apply the random forest model to do feature selection

- Avoid curse of dimensionality



Classification performance is pretty good.

- Naïve benchmark: 0.399
- Accuracy score: 0.74
- As a balanced dataset, macro average of recall (sensitivity) value: 0.73

	precision	recall	f1-score	support
high risk	0.84	0.83	0.83	76
low risk	0.69	0.83	0.75	93
mid risk	0.69	0.55	0.61	84
accuracy			0.74	253
macro avg	0.74	0.73	0.73	253
weighted avg	0.74	0.74	0.73	253

Confusion Matrix (Accuracy 0.7352)

	Prediction		
Actual	high risk	low risk	mid risk
high risk	63	7	6
low risk	1	77	15
mid risk	11	27	46

Multinominal Logistic Regression

Apply the multinominal logistic regression

- The outcome has three classes
- Classes have no meaning order

	Age	SystolicBP	BS	BodyTemp	HeartRate	Intercepts
First set of coefficients	-0.135527	0.680812	1.386398	0.565501	0.295627	-0.485296
Second set of coefficients	0.097373	-0.725530	-1.038431	-0.622962	-0.214778	0.120546
Third set of coefficients	0.038153	0.044718	-0.347968	0.057461	-0.080849	0.364751

Classification performance is slightly better than random guess

- Naïve benchmark: 0.399
- Accuracy score: 0.58
- As a balanced dataset, macro average of recall (sensitivity) value: 0.57

	precision	recall	f1-score	support
high risk	0.68	0.57	0.62	76
low risk	0.64	0.84	0.73	93
mid risk	0.38	0.31	0.34	84
accuracy			0.58	253
macro avg	0.57	0.57	0.56	253
weighted avg	0.57	0.58	0.57	253

Confusion Matrix (Accuracy 0.5810)

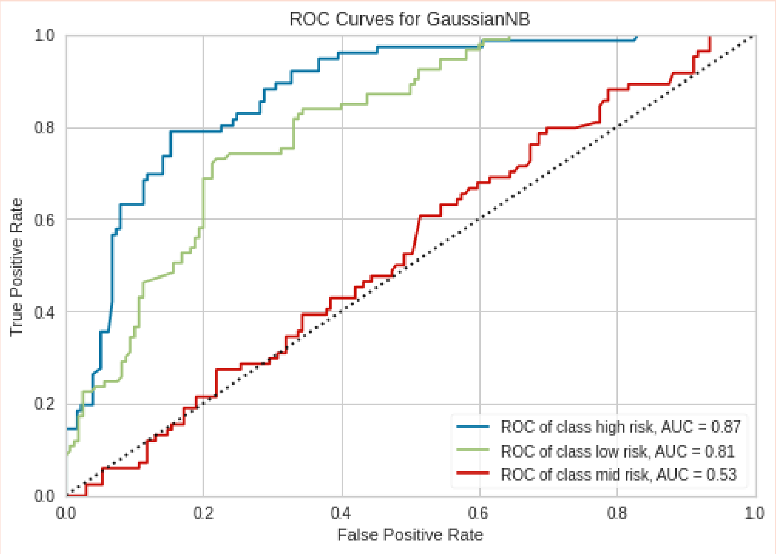
	Prediction		
Actual	high risk	low risk	mid risk
high risk	43	3	30
low risk	2	78	13
mid risk	18	40	26

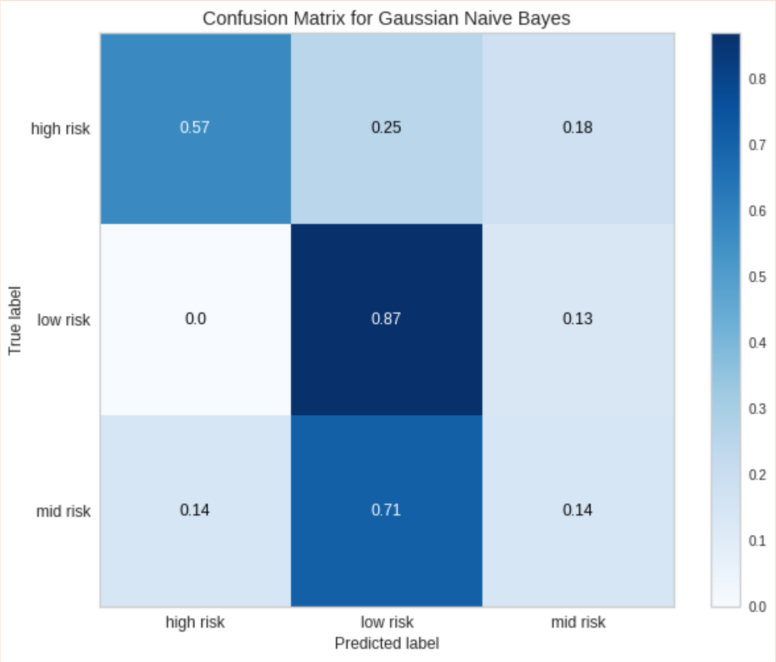
Gaussian Naïve Bayes

Assume that the predictors follow Gaussian distribution

- Accuracy score: 0.54 (naïve benchmark: 0.399)
- Precisions (positive predictive value) and recalls are not outstanding
- AUC of mid risk group is close to 0.5 (randomly guessing)

	precision	recall	f1-score	support
high risk	0.78	0.57	0.66	76
low risk	0.51	0.87	0.64	93
mid risk	0.32	0.14	0.20	84
accuracy			0.54	253
macro avg	0.53	0.53	0.50	253
weighted avg	0.53	0.54	0.50	253

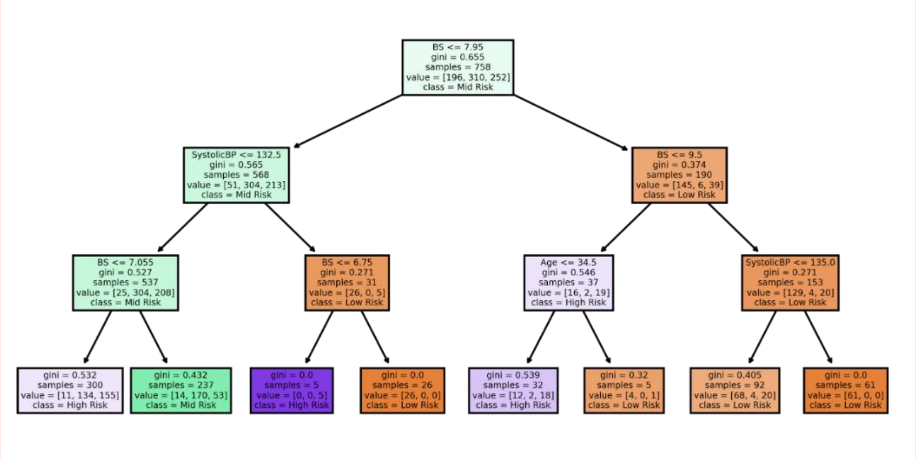




Decision Tree

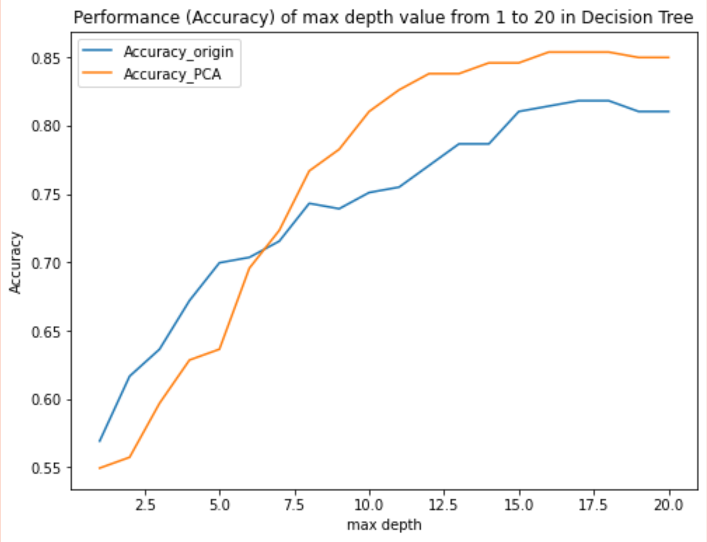
Part 1 Grow an ideal tree

1. Three predictors, **blood sugar, systolic blood pressure, and age** are strong measures



Grow a tree with max depth of 3

2. PCA scores perform better as the max depth increases



Accuracy with different max depth

3. Best combination of parameters

	Value
criterion	gini
max_depth	17
min_impurity_decrease	0.0
min_samples_leaf	1
splitter	random

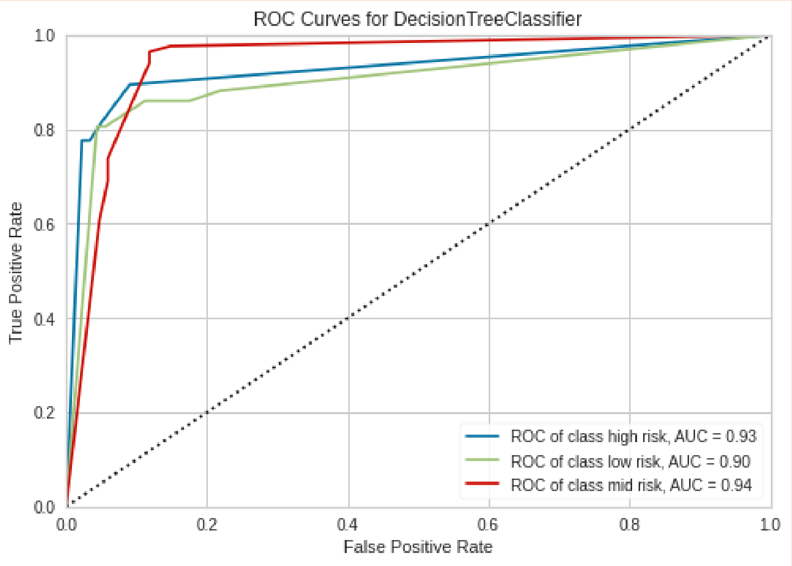
Find parameters by GridSearch

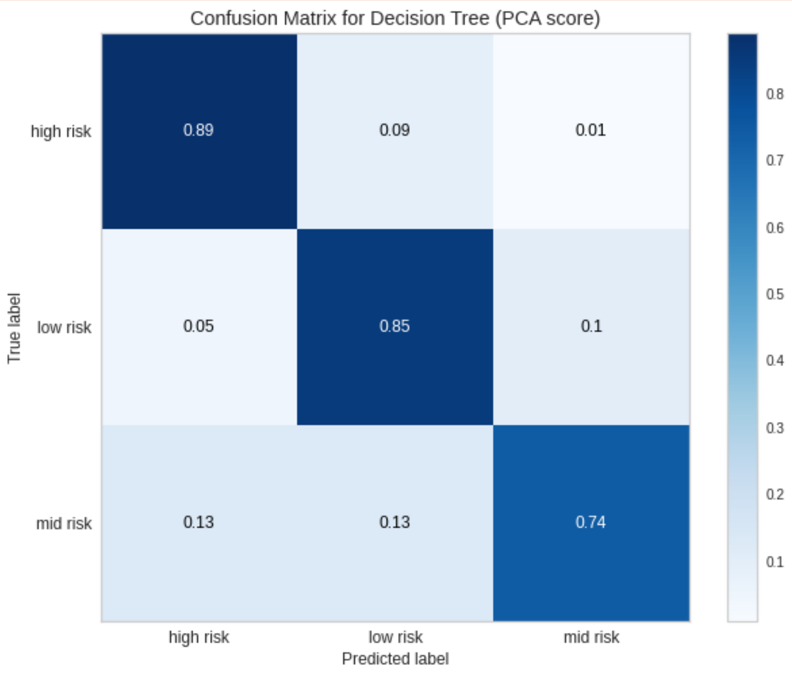
Decision Tree

Part 2 Modeling with PCA scores

- Accuracy score: 0.83 (naïve benchmark: 0.399)
- Precisions (positive predictive value): greater than 0.8
- Recall rate (sensitivity) of high risk is 0.89
- AUCs are higher than 0.9

	precision	recall	f1-score	support
high risk	0.81	0.89	0.85	76
low risk	0.81	0.85	0.83	93
mid risk	0.86	0.74	0.79	84
accuracy			0.83	253
macro avg	0.83	0.83	0.83	253
weighted avg	0.83	0.83	0.82	253



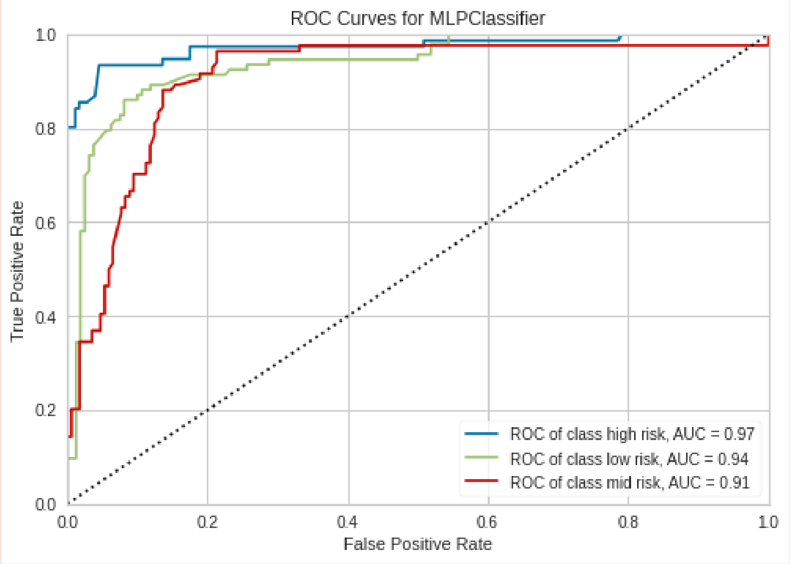


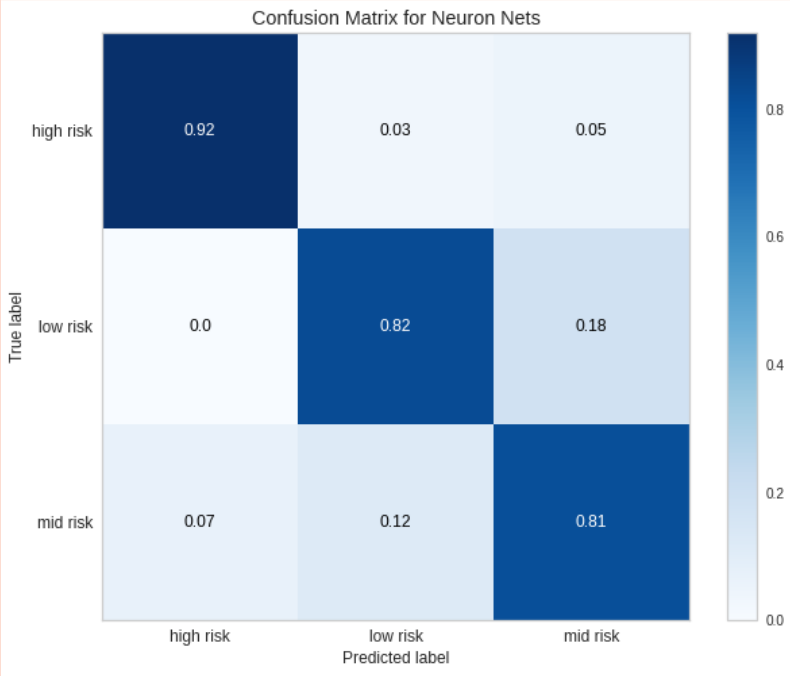
Artificial Neural Networks

ANN of 3 hidden layers with 10 nodes

- Accuracy score: 0.85 (naïve benchmark: 0.399)
- Precision (positive predictive value) of high risk: 0.92
- Recall (sensitivity) of 3 groups: higher than 0.8
- AUCs are higher than 0.9

	precision	recall	f1-score	support
high risk	0.92	0.92	0.92	76
low risk	0.86	0.82	0.84	93
mid risk	0.76	0.81	0.79	84
accuracy			0.85	253
macro avg	0.85	0.85	0.85	253
weighted avg	0.85	0.85	0.85	253



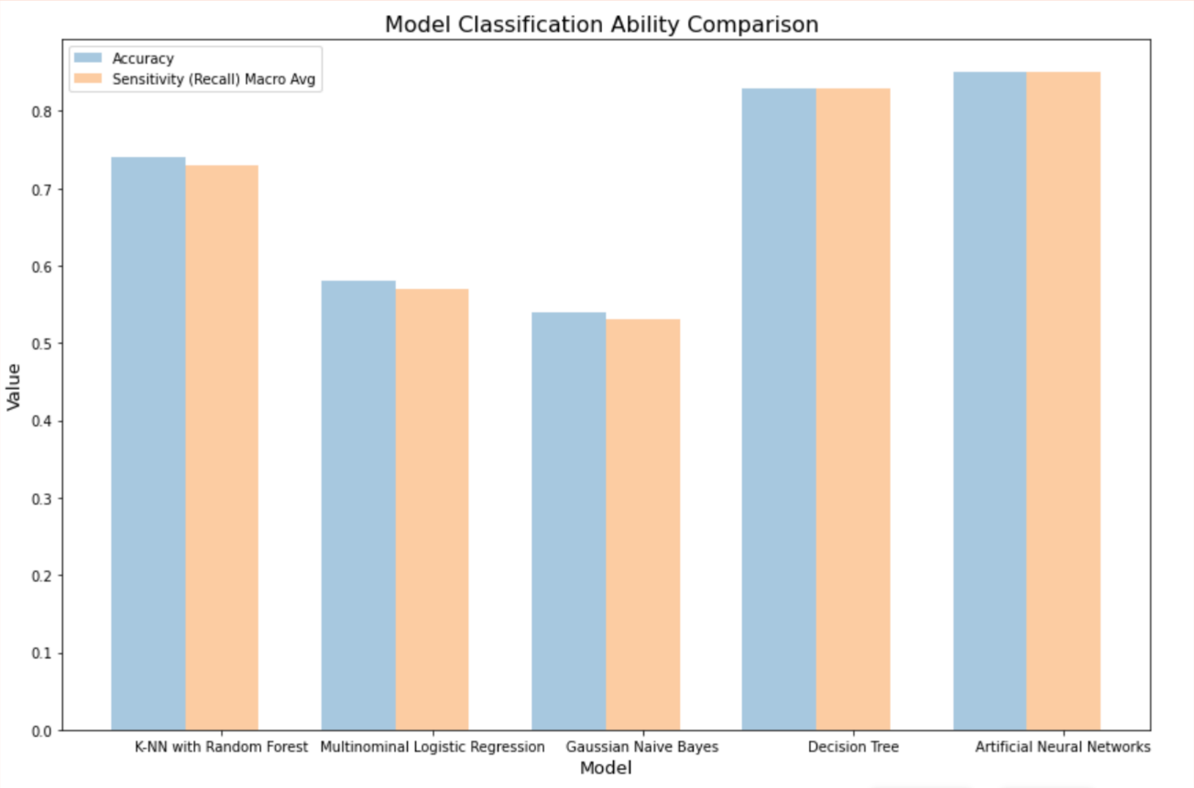


Models Performance Comparison

The best model is Artificial Neural Networks

- Highest Accuracy
- Highest Sensitivity (Recall) Macro Average value

	Model	Accuracy	Sensitivity (Recall) Macro Avg
0	K-NN with Random Forest	0.74	0.73
1	Multinominal Logistic Regression	0.58	0.57
2	Gaussian Naive Bayes	0.54	0.53
3	Decision Tree	0.83	0.83
4	Artificial Neural Networks	0.85	0.85



Conclusion

Physical Information

- Age
- Systolic Blood Pressure
- Blood Sugar
- Body Temperature
- Heart Rate

Model

- Artificial Neural Networks

Classification Output

- Low Risk
- Mid Risk
- High Risk

- a quick and reliable reference for medical experts
- reduction on diagnostic cost required for patients

Resources

- United Nations. (n.d.). *Goal 3 | Department of Economic and Social Affairs*. United Nations. Retrieved January 29, 2022, from https://sdgs.un.org/goals/goal3
- UCI Machine Learning Repository: Maternal Health Risk Data Set Data Set. (n.d.). Retrieved January 29, 2022, from http://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set#

Thank You!

For Your Attention