Statistical Analysis in Maternal Health Conditions with Risk Factors

Milestone: Project Report

Group 22
Yingnan He
Ming Luo

he.yingn@northeastern.edu
luo.ming1@northeastern.edu

Percentage of Effort Contributed by Yingnan He: _____50%_____

Percentage of Effort Contributed by Ming Luo: _____50%_____

Signature of Student 1: _____ YINGNAN HE_____

Signature of Student 2: _____MING LUO_____

Submission Data: _____April 24th, 2022_____

Contents

Project Background 3
Problem Definition 3
Data Collection and Processing 3
Checking Missing Values 4
Data Processing 4
Identifying Outliers 5
Checking Class Balance 6
Data Exploration and Visualization 7
Descriptive Statistics of the Predictors by Risk Level 7
Heatmap and Pair Plot of Predictors 9
Principal Component Analysis 11
Variance, Proportion Variance, and Cumulative Proportion of Variance 11
PCA Weights 11
PCA Visualization 12
Model Exploration and Model Selection 13
Data Preparation for Data Mining Task 13
Naive Benchmark 14
Task Specification and Model Selection 15
Data Partitioning 15
Data Standardization 15
Model Selection 16
K-Nearest Neighbors (K-NN) With Random Forest 16
Multinominal Logistic Regression 21
Gaussian Naïve Bayes 22
Decision Tree 24
Artificial Neural Networks 27
Models Performance Comparison 29
Conclusion 30

Project Background

It is highly important to monitor the maternal health conditions of women and maintain a lower risk level of pregnancy. An existing relationship between risk levels have been thoroughly studied by the literary material, reviewed by medical experts, and related to the risk factors, which are pregnancy-related medical parameters monitored.

With the insight of the deterministic factors of maternal health, medical experts could support patients with specific treatments in an efficient manner. With such encouraging techniques, we believe the newborn mortality rate will decrease significantly, not to mention the improvement of mothers' health conditions.

Problem Definition

After exploring and preprocessing the database, we will build models to classify the maternal health risk level with risk factors, such as blood sugar and age. Based on the model with best performance on the test set, we can further predict the health risk level during pregnancy for new records.

Data Collection and Processing

The original data set is called Maternal Health Risk Dataset, which was collected from various healthcare agencies in rural areas of Bangladesh through the IoT-based risk monitoring system and donated in 2020. There are 7 attributes in this dataset: Age, Systolic Blood Pressure, Diastolic Blood, Blood Sugar, Body Temperature, Heart Rate, and Risk Level. There are three kinds of data type in the dataset: the data type of four input variables Age, Systolic Blood Pressure (SystolicBP), Diastolic Blood Pressure (DiastolicBP), and Heart Rate is integer, the data type of rest two input variables Blood Sugar (BS) and Body Temperature (BodyTemp) is integer, and the data type of the output variable Risk Level is string. The total number of observations is 1014. Therefore, this data has 1014 rows and 7 columns.

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)

Checking Missing Values

We checked missing values, and fortunately, there is no missing values in the data set.

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

Fig1. Number of missing values

Data Processing

From the description table, we can see that there are 1014 records in this dataset. The mean and median (50% from above table) are close to each other for all the variables. For example, the mean of age is 29 and median of age is 26. Although the mean is slightly larger than median, they are relatively the same. Therefore, generally speaking, the distribution of each of the variable has no skew.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
count	1014	1014	1014	1014.00	1014.00	1014
mean	29	113	76	8.73	98.67	74
std	13	18	13	3.29	1.37	8
min	10	70	49	6.00	98.00	7
25%	19	100	65	6.90	98.00	70
50%	26	120	80	7.50	98.00	76
75%	39	120	90	8.00	98.00	80
max	70	160	100	19.00	103.00	90

Fig2. Statistical description of variables

Identifying Outliers

From the boxplot below, we can see that there are several unreasonable outliers that we can omit to increase the accuracy of classification models.

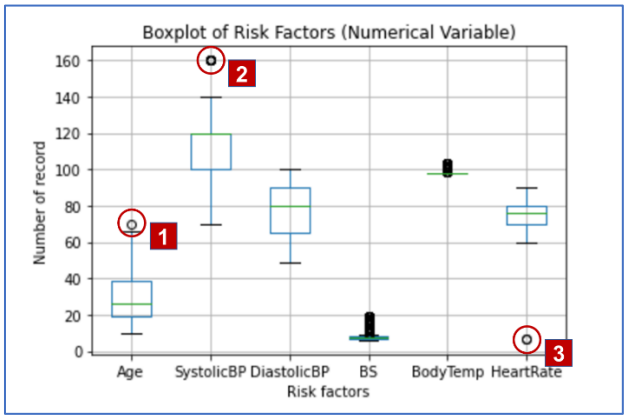

Fig3-1. Boxplot of 6 variables with outliers identified

1) The maximum value of Age is an outlier, which has a value of 70. According to the rule of thumb, it's impossible to get pregnant at age 70, and still maintain at a low risk level. Thus, we removed the record.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
338	70	85	60	6.9	102.0	70	low risk

Fig3-2. Records with age outliers of 70

2) The maximum value of SystolicBP is an outlier, which has a value of 160. According to the domain knowledge in medication, if a pregnant woman with SystolicBP of 160, she is

in danger. However, since it's possible to have SystolicBP with 160 due to getting pregnant at age 40, we will leave these records.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
123	40	160	100	19.0	98.0	77	high risk
130	40	160	100	19.0	98.0	77	high risk
166	40	160	100	19.0	98.0	77	high risk
262	40	160	100	19.0	98.0	77	high risk
362	40	160	100	19.0	98.0	77	high risk
538	40	160	100	19.0	98.0	77	high risk
583	40	160	100	19.0	98.0	77	high risk
689	40	160	100	19.0	98.0	77	high risk
961	40	160	100	19.0	98.0	77	high risk
994	40	160	100	19.0	98.0	77	high risk

Fig3-3. Records with SystolicBP outliers of 160

3) The minimum value of Heart Rate is an outlier, which has a value of 7. According to the rule of thumb, if a person with a heart rate of 7, she almost dies, and they cannot be assessed as at low risk. Thus, these two records are also unreasonable outliers.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
499	16	120	75	7.9	98.0	7	low risk
908	16	120	75	7.9	98.0	7	low risk

Fig3-4. Records with heart rate outliers of 7

Therefore, we removed 3 unreasonable records, and there are 1011 records for analysis.

Checking Class Balance

There is one categorical variable in the dataset, and it is the output variable, Risk Level.

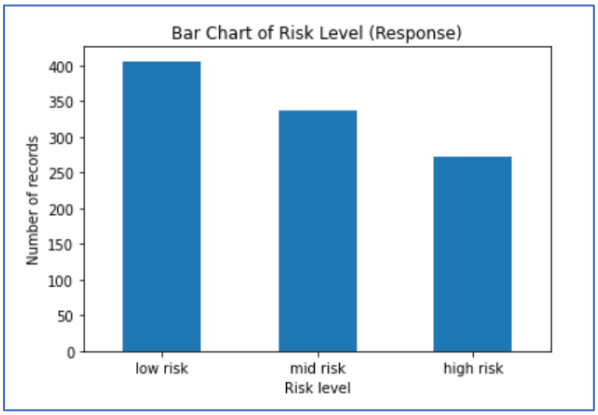

Fig4. Bar chart of response, Risk Level

From the plot, we can see that with the increase of risk level, the number of records in each risk level is decreasing. Since none of the numbers are extremely large or small, the classes are relatively evenly distributed.

Data Exploration and Visualization

Descriptive Statistics of the Predictors by Risk Level

	Risk_num	0	1	2
Age	count	403.00	336.00	272.00
	mean	26.82	28.36	36.22
	std	12.97	12.55	13.03
	min	10.00	10.00	12.00
	25%	17.50	19.00	25.00
	50%	22.00	25.00	35.00
	75%	32.00	32.00	48.00
	max	66.00	60.00	65.00
SystolicBP	count	403.00	336.00	272.00
	mean	105.85	113.15	124.19
	std	15.89	14.98	20.23
	min	70.00	70.00	83.00
	25%	90.00	100.00	120.00
	50%	120.00	120.00	130.00
	75%	120.00	120.00	140.00
	max	129.00	140.00	160.00
DiastolicBP	count	403.00	336.00	272.00
	mean	72.55	74.23	85.07
	std	13.09	11.49	14.11
	min	49.00	50.00	60.00
	25%	60.00	65.00	75.00
	50%	75.00	75.00	90.00
	75%	80.00	80.00	100.00
	max	100.00	100.00	100.00

	Risk_num	0	1	2
BS	count	403.00	336.00	272.00
	mean	7.22	7.80	12.12
	std	0.65	2.29	4.17
	min	6.00	6.00	6.10
	25%	6.90	6.80	7.90
	50%	7.50	7.00	11.00
	75%	7.50	7.80	15.00
	max	11.00	18.00	19.00
BodyTemp	count	403.00	336.00	272.00
	mean	98.36	98.83	98.90
	std	1.10	1.43	1.56
	min	98.00	98.00	98.00
	25%	98.00	98.00	98.00
	50%	98.00	98.00	98.00
	75%	98.00	100.00	100.00
	max	103.00	103.00	103.00
HeartRate	count	403.00	336.00	272.00
	mean	73.10	74.18	76.74
	std	6.90	6.77	8.70
	min	60.00	60.00	60.00
	25%	70.00	70.00	70.00
	50%	70.00	76.00	77.00
	75%	77.00	78.00	86.00
	max	88.00	88.00	90.00

Fig5. Descriptive statistics of the predictors by risk level

To be more readable, we replace the three risk levels with numbers. 0, 1, and 2 represent low-risk level, mid-risk level, and high-risk level respectively. After summarizing the data at each risk level, we have the initial recognition for each variable. At the first look of mean value, the

variables of age, blood pressure, and blood sugar give larger scales from low-risk group to high-risk group.

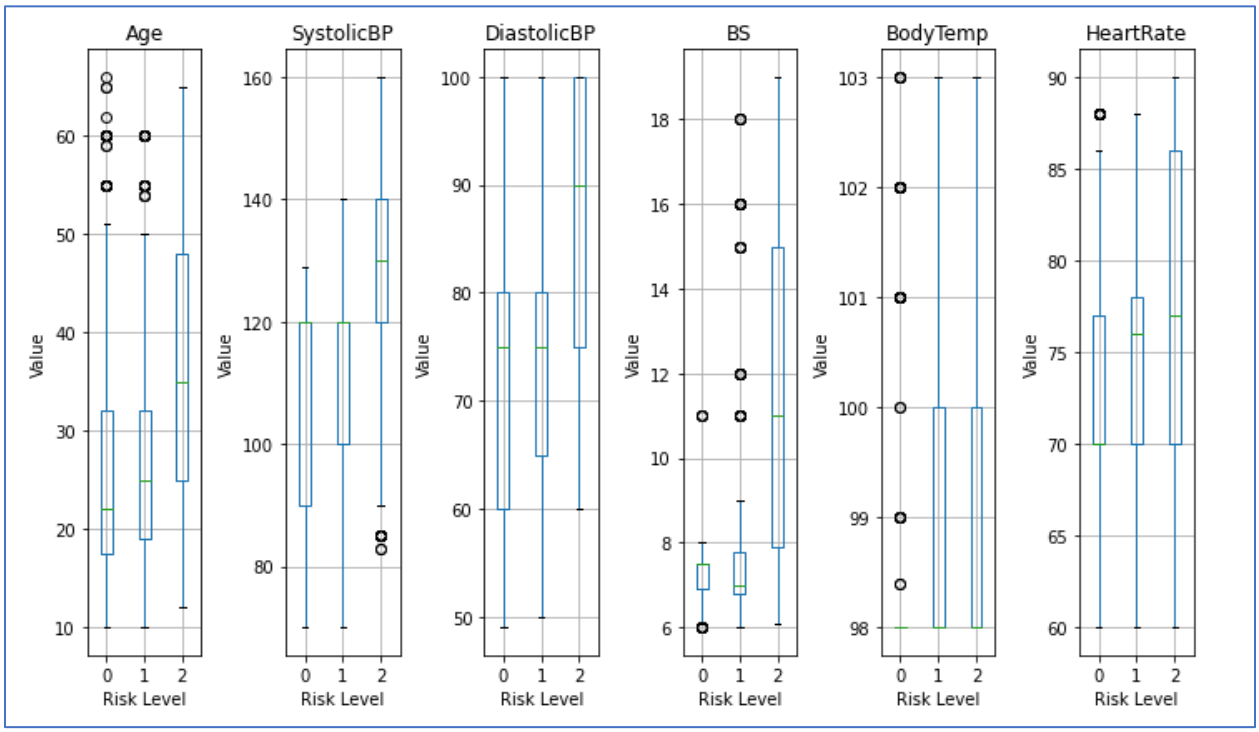

Fig6. Distributions of risk factors by risk level

These side-by-side boxplots are for exploring the risk level output variable by different numerical predictors. Every boxplot compares one of the six risk factors' contributions across the three risk levels.

1) **Age and Risk Level:** The first plot shows the relationship between output variable Risk Level and input variable Age. We can see that the median age in different risk levels is following the rule that the larger the median age is, the higher the risk level is.
2) **Systolic Blood Pressure and Risk Level:** The second plot shows the distribution of SystolicBP among different risk levels. we can tell that if a person with SystolicBP is higher than 120, she is highly possible to be classified as high-risk level.
3) **Diastolic Blood Pressure and Risk Level:** The third plot shows the distribution of DiastolicBP among different risk levels. If a person with DiastolicBP is higher than 80, it is highly possible to be classified as a high-risk level.

4) **Blood Sugar and Risk Level:** Based on the fourth boxplot, we can tell that if a person with BS is below 8, it is highly possible to be classified as low-risk level or medium risk level.
5) **Body Temperature and Risk Level:** In the fifth boxplot, the medians of BodyTemp in different risk levels are the same. The distributions of BodyTemp in medium and high-risk levels are similar. It's hard to use BodyTemp to tell the risk level.
6) **Heart Rate and Risk Level:** In the sixth boxplot, the distributions of heart rate in all risk levels overlap between heart rate 70 to 75. If a lady's heart rate is 70, it's impossible to tell which risk level she will possibly be.

Heatmap and Pair Plot of Predictors

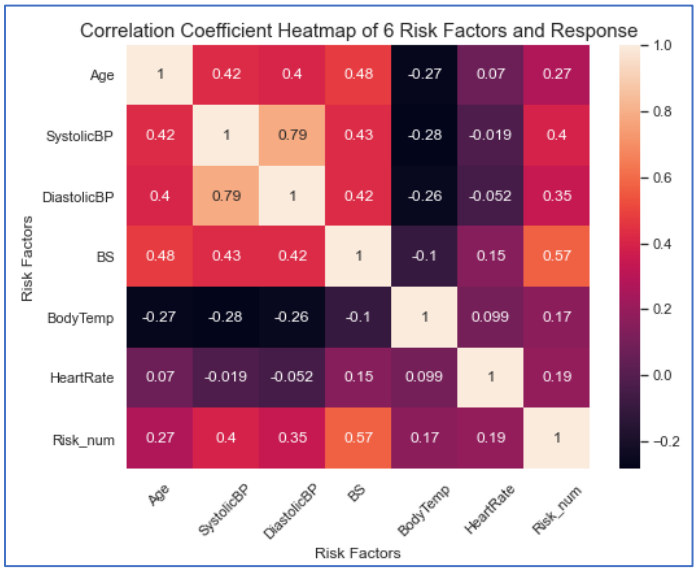

Fig7. Correlation Coefficient Heatmap of 6 Risk Factors and Response

Based on this correlation coefficient heatmap of 6 risk factors, we can tell age, Systolic BP, Diastolic BP, and blood sugar are positive related with each other, since their coefficients are near 0.5. Systolic BP and Diastolic BP are showing a relatively strong positive relationship especially.

Body temperature and heart rate, however, are showing relatively weak correlations with other risk factors since the absolute values of coefficient with others are low.

Comparing correlation coefficients of response (risk level) and 6 risk factors, blood sugar has the highest correlation with risk level, following by systolic blood pressure and Diastolic blood pressure.

SystolicBP and Diastolic BP are strongly positive related. The correlation coefficient between them is nearly 0.8. We plan to drop one of them in the later analysis. The correlation coefficients of Systolic BP and Diastolic BP with the output variable Risk Level are 0.4 and 0.35, respectively. Thus, we will keep Systolic BP, which has a stronger correlation with response.

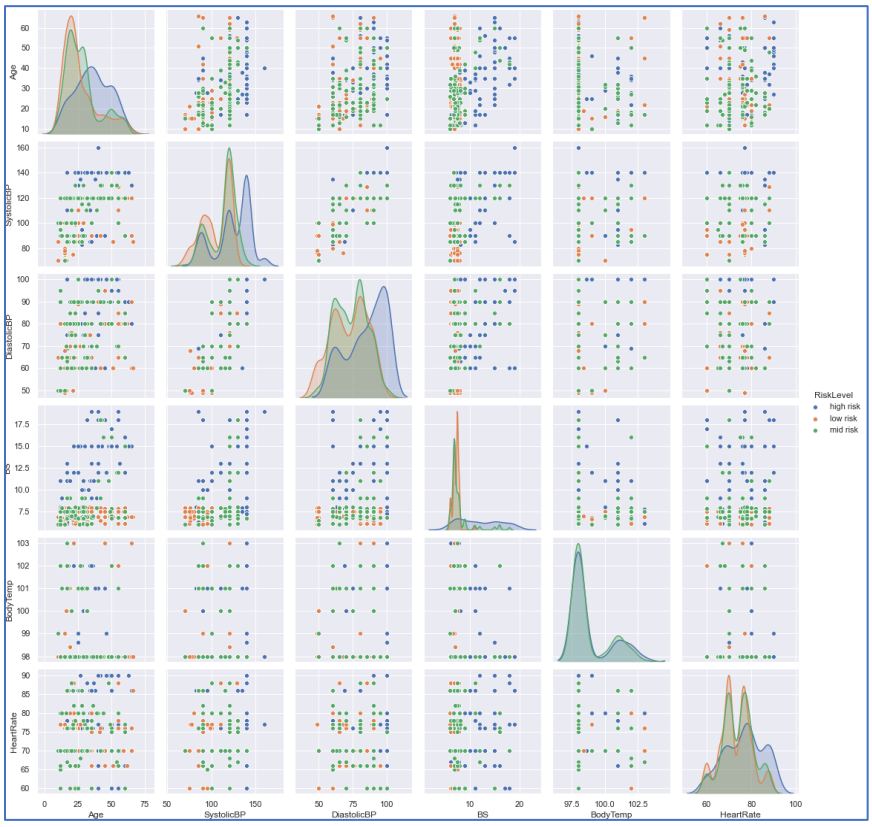

Fig8. Pair plot of 6 risk factors by risk level

This big pair plot of 6 risk factors gives us comprehensive information, supporting some insights we acquired from other charts above.

According to the plots along the diagonal, we can tell the distributions of age, blood pressure, blood sugar (BS) among high-risk women is different from the other two groups. High-risk women tend to be elder, with higher blood pressure and blood sugar. This was concluded from the chart 'Distributions of risk factors by risk level' as well.

Moreover, from the scatter plots, it is hard to find an extremely strong positive/negative relationship among these risk factors. But in the scatter plot of systolic blood pressure and diastolic blood pressure, the points seem to move in the same direction, which verifies the result from our correlation coefficient heatmap.

Also, when looking at the horizontal plots of blood sugar, points in blue (high-risk women) are more likely located at a higher place compared to other groups. We can guess blood sugar is a relatively strong indicator to classify if a pregnant woman is facing a higher risk of maternity. This is also similar to one of our observations generated from the heatmap above.

Principal Component Analysis

Variance, Proportion Variance, and Cumulative Proportion of 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

Fig9. Variance, proportion variance, and cumulative proportion of variance of 6 PCs

Observing cumulative proportions, we can tell that the first 2 principal components only capture 62.7% variance of predictors. Even first 3 components capture 76.7% variance. This can be seen as a good sign to some extent. On one hand, 6 risk factors provide relatively less information in common, which means their relationships are not linearly dependent.

This means these features might give us different criteria to classify pregnant women by risk level. We think that decision tree may be a better algorithm if we conduct later classification analysis. (It is a temporary assumption here. The performance of models should be evaluated by sensitivity and positive predictive rate in practice)

PCA 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

Fig10. PCA weights for 6 risk factors

Based on the weighting table above, PC1 is dominated by variables age, systolic blood pressure, diastolic blood pressure, as well as blood sugar. PC2 is dominated by variables heartrate and body temperature.

PCA Visualization

Two scatter plots above use first 3 and first 2 principal components as axis respectively. We colored points in different color by risk level (Green: low risk, orange: mid risk, red: high risk).

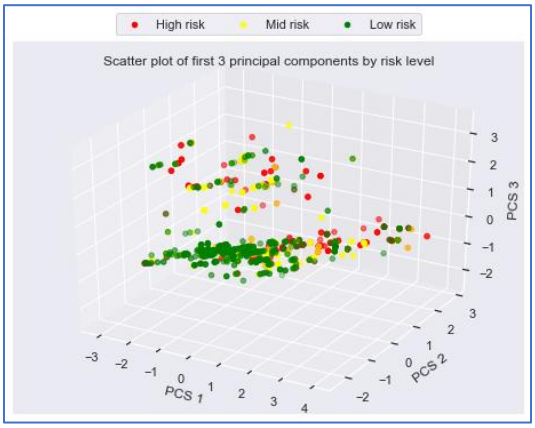

Fig11. Plot data points on a 3D plane defined by the first 3 components

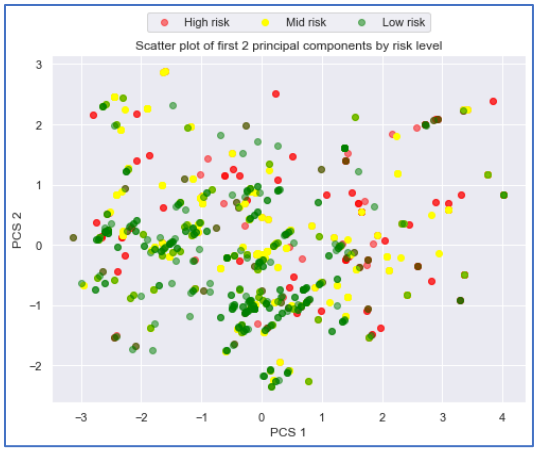

Fig12. Plot data points on a 2D plane defined by the first 2 components

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

In following data mining process, we will explore other algorithms for classification and evaluate their performances.

Model Exploration and Model Selection

Data Preparation for Data Mining Task

We remove three unreasonable outliers during the data exploration and one predictor: Diastolic BP, which is a high correlation with another predictor: SystolicBP, during the dimension reduction process from the original dataset. In addition, we add one column based on the output variable that three risk levels have been replaced by numbers. 0, 1, and 2 to represent low-risk level, mid-risk level, and high-risk level respectively. This is because the original output variable data type is string, and it limits the data mining model selection. Converting the string into discrete integers can make the data apply to more models, such as the logistic regression model. Therefore, the dataset for the following data mining tasks has 7 attributes in this dataset: Age, Systolic Blood Pressure, Blood Sugar, Body Temperature, Heart Rate, Risk Level, and Risk Number. There are three kinds of data types in the dataset: the data type of three input variables Age, Systolic Blood Pressure (SystolicBP), and Heart Rate is integer, the data type of rest two input variables and one output variable Blood Sugar (BS), Risk

Number, and Body Temperature (BodyTemp) is integer, and the data type of the output variable Risk Level is string. The total number of observations is 1011. Therefore, this data has 1011 rows and 7 columns.

	Age	SystolicBP	BS	BodyTemp	HeartRate	RiskLevel	Risk_num
0	25	130	15.0	98.0	86	high risk	2
1	35	140	13.0	98.0	70	high risk	2
2	29	90	8.0	100.0	80	high risk	2
3	30	140	7.0	98.0	70	high risk	2
4	35	120	6.1	98.0	76	low risk	0
...	...	...	...	...	...	...	...
1006	22	120	15.0	98.0	80	high risk	2
1007	55	120	18.0	98.0	60	high risk	2
1008	35	85	19.0	98.0	86	high risk	2
1009	43	120	18.0	98.0	70	high risk	2
1010	32	120	6.0	101.0	76	mid risk	1

1011 rows × 7 columns

Fig13. Dataset for data mining task

Naive Benchmark

The naive rule for the classification relies only on the output variables, which is risk level. The naive rule ignores the information provided by the predictors. In another word, the record will be classified as the most prevalent class in the dataset regardless of its profile values. A good-classification model should perform better than the naive rule. To get the naïve benchmark for the Maternal Health Risk dataset, we count the number of three risk levels and pick the risk level with the highest count number (the number is 403). Then divide this number by the sum of the number of three risk levels, which is the row number of the dataset (the number is 1011).

	Number
low risk	403
mid risk	336
high risk	272

Fig14. Number of records of three risk levels

The naive benchmark for the dataset is 0.399.

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 will apply the following classification models to the Maternal Health Risk dataset.

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

Data Partitioning

To avoid bias when using the same data to train the model and evaluate the model performance, we will partition the Maternal Health Risk dataset into 2 functional partitions randomly: the training dataset and the validation dataset. This step is done with Python scikit-learn package built-in function: train_test_split(). We set 75% of the dataset as training dataset and the rest of the 25% dataset as validation dataset. Therefore, 758 records are chosen randomly as training dataset and the rest 253 records belong to validation dataset. We will apply the same training dataset and the validation dataset mentioned above for all of our models.

Data Standardization

Below is descriptive statistics of the five predictors that are prepared to do data mining task. We can observe that the values of the five predictors are in different scale. For example, the range of predictor Age is between 10 and 66. However, the range of SystolicBP is between 70 and 160. If we don't standardize the data, Age and SystolicBP will not be given equal

importance in terms of variability. This will lead to a biased model and the classification result will be inaccurate.

	Age	SystolicBP	BS	BodyTemp	HeartRate
count	758.00	758.00	758.00	758.00	758.00
mean	29.56	112.89	8.68	98.65	74.65
std	13.33	18.35	3.27	1.36	7.58
min	10.00	70.00	6.00	98.00	60.00
25%	19.00	100.00	6.90	98.00	70.00
50%	26.00	120.00	7.50	98.00	76.00
75%	36.00	120.00	7.98	98.00	80.00
max	66.00	160.00	19.00	103.00	90.00

Fig15. Descriptive statistics of the 5 predictors

Therefore, we will standardize the dataset. Here are the steps we used to standardize the dataset. First, we calculate the mean and standard deviation of predictors of the training dataset. Second, use the above mean and standard deviation to standardize both the predictors of training data and validation data.

Model Selection

So far, we are going to apply 5 models: K-Nearest Neighbors with Random Forest, Multinominal Logistic Regression, Gaussian Naïve Bayes, Decision Tree, and Artificial Neural Networks to the Maternal Health Risk dataset.

K-Nearest Neighbors (K-NN) With Random Forest

K-Nearest Neighbors (K-NN)

We pick 15 different numbers of neighbors to do the classification computation. The 10 different numbers of neighbors are in the range of [1, 15] with the step of 1. We use k to indicate the number of neighbors. After comparing the accuracy score for different k values on the validation set, we pick the k with the maximum classification accuracy score. Below are the

K-NN model accuracy scores with different k values. When k = 1, the K-NN model accuracy score is the highest, which is nearly 0.79. Then, we will use k =1 to classify the new record.

	k Value	Accuracy Score
0	1	0.790514
1	2	0.731225
2	3	0.711462
3	4	0.691700
4	5	0.644269
5	6	0.695652
6	7	0.707510
7	8	0.687747
8	9	0.664032
9	10	0.656126
10	11	0.652174
11	12	0.679842
12	13	0.667984
13	14	0.679842
14	15	0.671937

Fig16. K-NN model accuracy scores with different k values

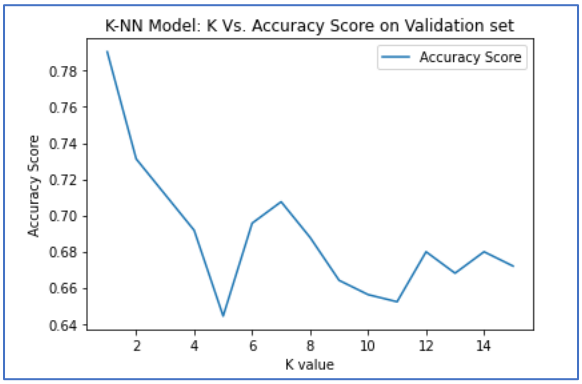

Fig17. K-NN model accuracy scores with different k values line plot

K-NN Classification Performance Evaluation

After applying the model to the validation set, the highest classification accuracy score for the model is nearly 0.79.

However, when k =1, the classification model is very sensitive to the local structure of the training dataset and may cause overfitting problem. Therefore, we will pick the k = 6 and corresponding accuracy score is 0.707510. This accuracy is larger than the naïve benchmark we get above, which is 0.399. Therefore, the K-NN model with k = 7 is a good classification model.

K-Nearest Neighbors (K-NN) With Random Forest

In addition, 1000 records with 5 predictors may not be sufficient to make accurate classification model. Therefore, we will apply the random forest method to do feature selection. With the subset of predictors, one of the disadvantages of K-NN model, the curse of dimensionality phenomenon will be overcome. The accuracy of K-NN model on validation dataset will increase as well.

Feature	Importance
BS	0.397863
SystolicBP	0.211505
Age	0.202407
HeartRate	0.117347
BodyTemp	0.070878

Fig18. The importance value of each predictor

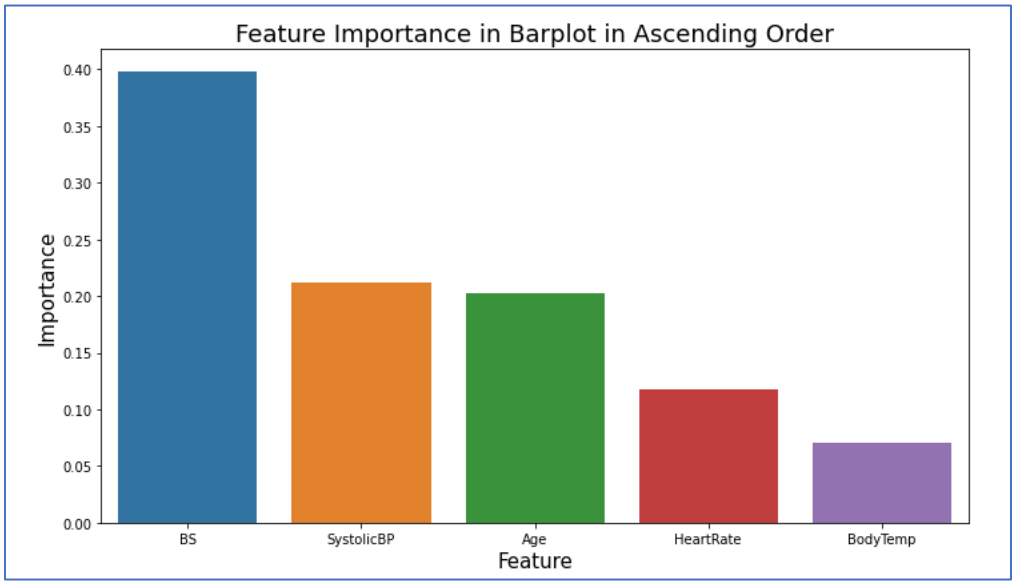

Fig19. The bar plot of the importance value of each predictor

From the above Fig18 and Fig19, we get the importance value of each predictor. Then we pick the top three most important predictors to make a new K-NN classification model, which are BS, SystolicBP, and Age.

	K	Accuracy_FeatureSelected
0	1	0.794466
1	2	0.739130
2	3	0.731225
3	4	0.675889
4	5	0.731225
5	6	0.731225
6	7	0.735178
7	8	0.727273
8	9	0.695652
9	10	0.707510
10	11	0.695652
11	12	0.707510
12	13	0.687747
13	14	0.707510
14	15	0.691700

Fig20. Accuracy scores with different k values for K-NN model with selected predictors

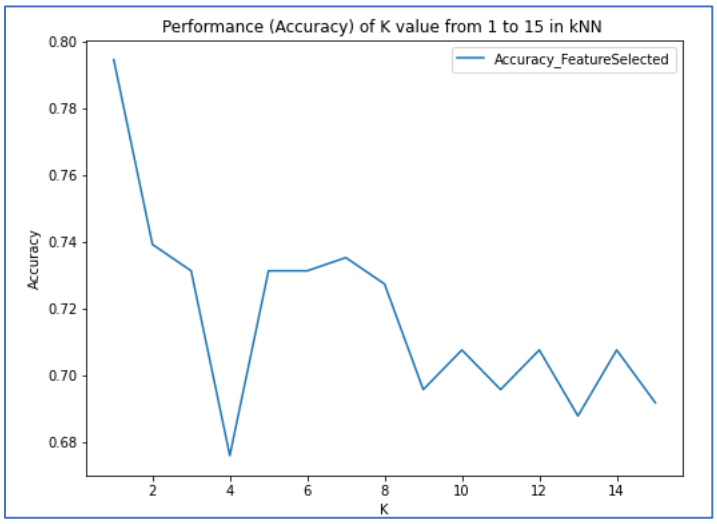

Fig21. Line plot of accuracy scores with different k values for K-NN model with selected predictors

K-NN with Selected Features Classification Performance Evaluation

Above Fig20 and Fig21 are table and line plot information about the accuracy scores with different k values for K-NN model with selected predictors: BS, SystolicBP, and Age. We can observe that when k = 7, the K-NN model has a high accuracy score, which is nearly 0.74. In addition, recall the accuracy score with K-NN model with k = 7 without feature selection, 0.707510. The K-NN model accuracy score with selected features is improved a bit. Therefore, we will use K-NN model with k = 6 with selected features: BS, SystolicBP, and Age.

Below figures are confusion matrix and classification performance summary of K-NN model on validation dataset. Since the low risk, mid risk, and high risk classes are balanced, the accuracy score is reliable to check the classification performance. However, we will check sensitivity (recall) to detect the ability of predicting each risk level. The sensitivity scores for low risk, mid risk, and high risk levels are 0.83, 0.55, and 0.83 respectively. In addition, since the data is in balance, the macro average value of recall, which is the arithmetic mean of the three recall values, is similar to weighted average recall score. Therefore, the K-NN model predicts the high risk and low risk accurately, when it comes to predict the mid risk level, the performance is slightly better than random guess.

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

Fig22. Confusion Matrix of K-NN model on validation dataset

	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

Fig23. Classification Performance Summary of K-NN model

Multinominal Logistic Regression

Logistic regression can classify a new record into a certain class based on the information from predictors. In addition, since the output variable has three classes: high risk, mid risk, and low risk and the classes haven't meaningful order, we will apply nominal (multi-class) logistic regression for multi-class classification. We choose the l2 penalty to regularize the dataset and apply a very large penalty parameter C = 1e42. In this nominal logistic model. there will be three estimated intercepts since the dataset output variable has three classes. In addition, the model will also have three separate sets of coefficients for the predictors for each level of the output variable. We will use all the information to predict new data and evaluate classification performance on the validation set.

	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

Fig24. Three Intercepts and three sets of coefficients for each level of risk

Logistic Regression Classification Performance Evaluation

After applying the model to the validation set, the classification accuracy score for the model is 0.581. This accuracy is larger than the naïve benchmark we get above, which is 0.399. Therefore, the model is a good classification model.

The following figures are confusion matrix and classification performance summary of Multinomial Logistic Regression model on validation dataset. As we mentioned before, since the low risk, mid risk, and high risk classes are balanced, the accuracy score is a good way to evaluate the classification performance. However, we will check sensitivity (recall) to detect the ability of predicting each risk level. The sensitivity scores for low risk, mid risk, and high risk levels are 0.84, 0.31, and 0.57 respectively. Therefore, the Multinomial Logistic Regression model predicts the low risk accurately, when it comes to predict the high risk level, the

performance is slightly better than random guess. However, when it comes to predict the mid risk level, the performance is worse than random guess.

Confusion Matrix (Accuracy 0.5810)

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

Fig25. Confusion Matrix of Multinomial Logistic Regression model

	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

Fig26. Classification Performance Summary of Multinomial Logistic Regression model

Gaussian Naïve Bayes

This time we are using the Gaussian Naive Bayes classification model to train our data for triple response prediction. Bayes' theorem is a generative approach to computing posterior probabilities $P(C_k|X)$ through modeling class-conditional likelihood $P(X|C_k)$, as well as the class priors $P(C_k)$. The predictors, age, systolic blood pressure, body temperature, blood sugar, and heart rate are continuous variables. Thus, we assume they are following the Gaussian Distribution and the class-conditional densities are Gaussian. After we fit the model, we predicted validation data and get the confusion matrix.

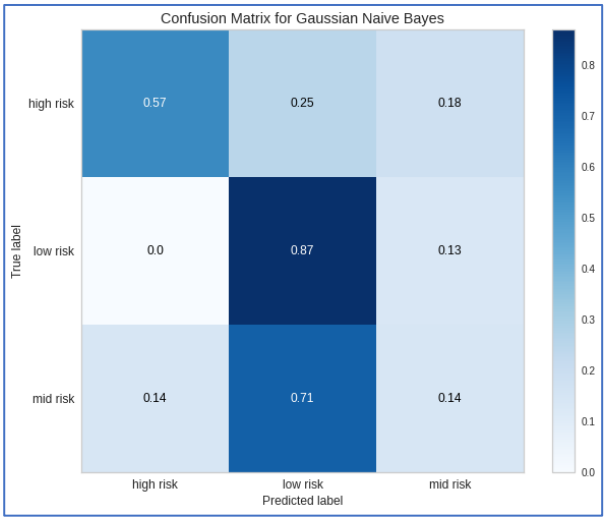

Fig27. Confusion Matrix of Gaussian Naive Bayes model

	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

Fig28. Summary of Gaussian Naive Bayes model

Gaussian Naïve Bayes Classification Performance Evaluation

The accuracy is 0.54, which is higher than naive benchmark of 0.399. But the recall rates (Sensitivity) of 3 classes are not very ideal. That's partially due to the assumption that predictors are Gaussian is not that trustful, and distributions of predictors in 3 classes are overlapped with each other. That can be easily observed from the diagonal plots of Fig8.

Below is the ROC graph of Gaussian Naive Bayes model. By visualizing this classifier performance at all cutoff values, we can tell the model does not perform very well on predicting the mid risk group, because its curve falls too close to the diagonal line, representing that it performs like a random classifier when recognizing women in the mid risk. Combined with the confusion matrix, 71% of mid risk women were classified as low risk. Thus, in the later analysis, we are going to explore a model which can predict all of 3 response classes.

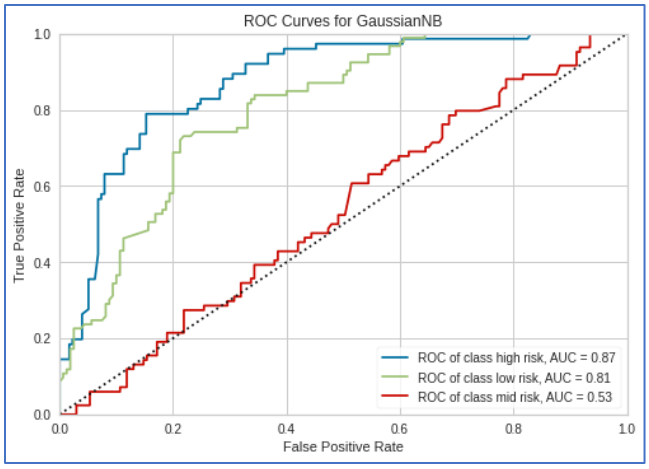

Fig29. ROC curve of Gaussian Naive Bayes model

Decision Tree

The fourth algorithm is the Decision Tree. Unlike the Gaussian Naive Bayes model, it makes few assumptions about data distribution and can fit complex datasets. Initially, we pruned a tree model with a max depth of 3 to see which ones of predictors perform best splits.

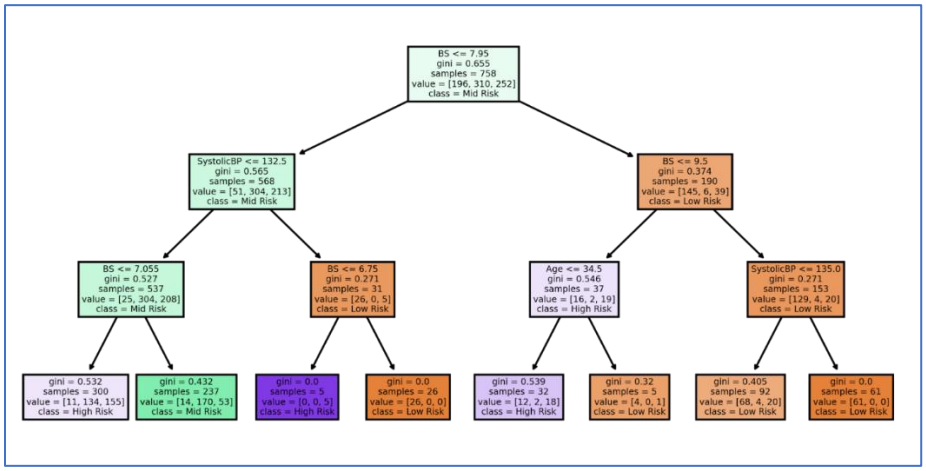

Fig30. Decision tree plot with a max depth equals 3

This inspired us that 3 predictors, blood sugar, systolic blood pressure, and age are strong measures that can better split records to produce child nodes with homogeneous class distribution when growing a tree with limited depth.

However, observing the Gini index of leaves, we did not grow trees that successfully classify the 3 classes. Therefore, in later analysis, we firstly explored and selected the parameter of max depth in decision tree model which can lead an optimal performance. We conducted decision tree algorithm both on original Data as well as PCA score. The accuracy trend of max depth on 2 dataset is shown as blow.

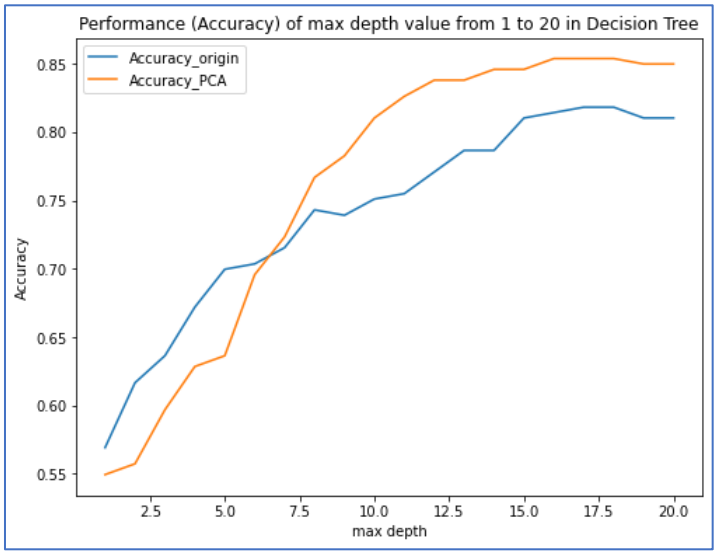

Fig31. Decision Tree models' accuracies with different max depth

Decision Tree algorithm has better performance on PCA scores as the max depth increases. As the max depth increased to 16, it reached the highest accuracy around 0.85.

Next, we optimized the classifier by cross-validation on PCA score dataset. We imported GridSearchCV object from Sklearn to help us look for and evaluate the possible combinations of parameter, including splitter, impurity measures, minimum impurity decrease, minimum sample leaf, along with max depth. As the result of classifier optimization by 5-fold cross-validation. The best combination of parameters is shown as below:

	Value
criterion	gini
max_depth	17
min_impurity_decrease	0.0
min_samples_leaf	1
splitter	random

Fig32. Optimized combination of parameters by cross-validation in decision tree

We input these values of parameters into decision tree classifier with PCA score, the performance summary is shown as below:

	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

Fig33. Summary of Decision Tree Model with optimized combination of parameters

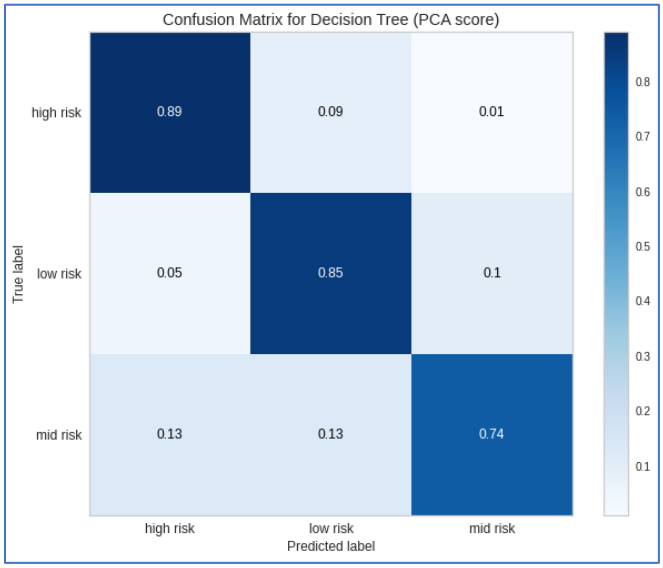

Fig34. Confusion Matrix of Decision Tree Classifier on PCA score

Decision Tree Classification Performance Evaluation

After applying the model to the validation set, the classification accuracy score for the model is 0.83. It is much larger than the naïve benchmark of 0.399. And the precisions (positive predictive value) in 3 groups are greater than 0.8. Surprisingly, the recall rate (sensitivity) of high risk is 0.89, which means that our model can recognize 9 pregnant women out of a high-risk group of 10. This time, the decision tree algorithm on PCA score is good performing.

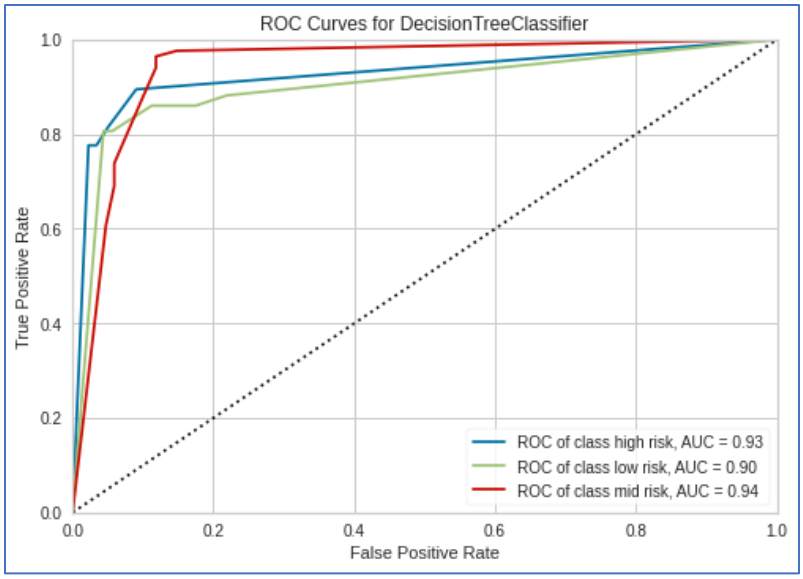

Fig 35. ROC curve for Decision Tree Classifier on PCA score

To better present the model performance, we constructed ROC curves of 3 groups by plotting sensitivity (True Positive Rate) vs (1-Specificity) as cutoff values vary from 0 to 1. This time, 3 ROC curves are all closer to the top left, with area under the curve higher than 0.9. According to these metrics, we believe that Decision Tree Classifier on the PCA score is performing very well.

Artificial Neural Networks

The fifth classifier is Artificial Neural Networks (ANN). It simulates the structure of human brain as it learns from experience. Here we adopted a basic neural networks architecture for the training and prediction process with error back propagation algorithm. The activation function in our ANN classifier is sigmoid function.

In our practice, we simply trained an ANN model of 3 hidden layers with 10 nodes in each layer to get a best performance.

Below are the ANN model evaluation results based on validation dataset.

	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

Fig36. Summary of Artificial Neural Networks Classifier

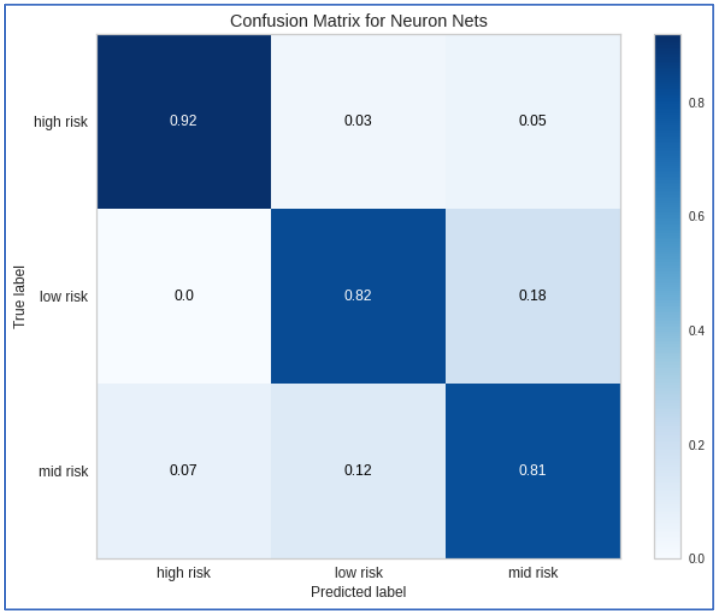

Fig37. Confusion Matrix of ANN Classifier

Artificial Neural Networks Classification Performance Evaluation

After applying the model to the validation set, the classification accuracy score for the model is 0.85. It is much larger than the naïve benchmark of 0.399. And the recall rates (sensitivity) of predicting 3 response groups are all greater than 0.8. Surprisingly, the recall rate (sensitivity) of high risk is 0.92, which is best performing among all models we currently have.

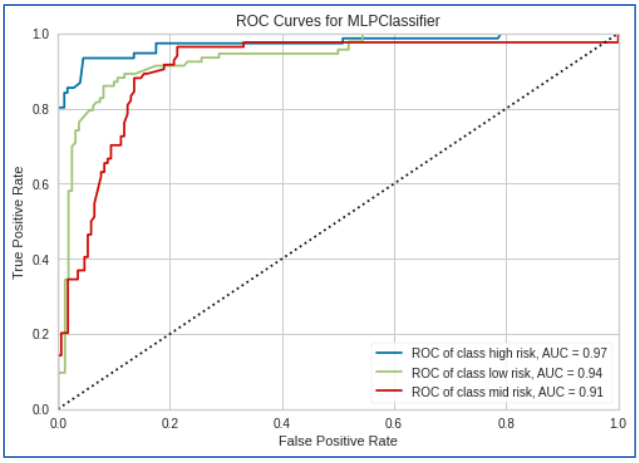

Fig 38. ROC curve for ANN Classifier

We constructed ROC curves of ANN Classifier for 3 risk groups as well. This time, 3 ROC curves are all much closer to the top left compared to other models. And the areas under curves (AUCs) are all higher than 0.9. Even though we cannot have 100% sensitivity along with 100% specificity, observing the ROC of high-risk group prediction (the blue ROC curve), this ANN model provides a good balance between this two metrics. Thus, this ANN Classifier can detect the women in high risk with high true positive rate without sacrificing its specificity.

Models Performance Comparison

We have explored six different models on the previous section. Below is the table summary of the model classification performance evaluation on the same validation dataset. The dataset is balanced. Therefore, we will compare the accuracy score and Sensitivity (Recall) Macro Average. The accuracy score will show the overall ability of predicting the new record into right class. The Recall (Sensitivity) Macro Average is the arithmetic mean of recall values of three risk levels, which also indicates the ability of classification. All the Accuracy and Recall Macro Average values are from each model's classification summary report.

	Model	Accuracy	Sensitivity (Recall) Macro Average
1	**K-Nearest Neighbors with Random Forest**	0.74	0.73
2	**Multinominal Logistic Regression**	0.58	0.57
3	**Gaussian Naïve Bayes**	0.54	0.53

4	**Decision Tree**	0.83	0.83
5	**Artificial Neural Networks**	0.85	0.85

Fig 39. Model Performance Comparison Table

Here is the bar chart of the above information.

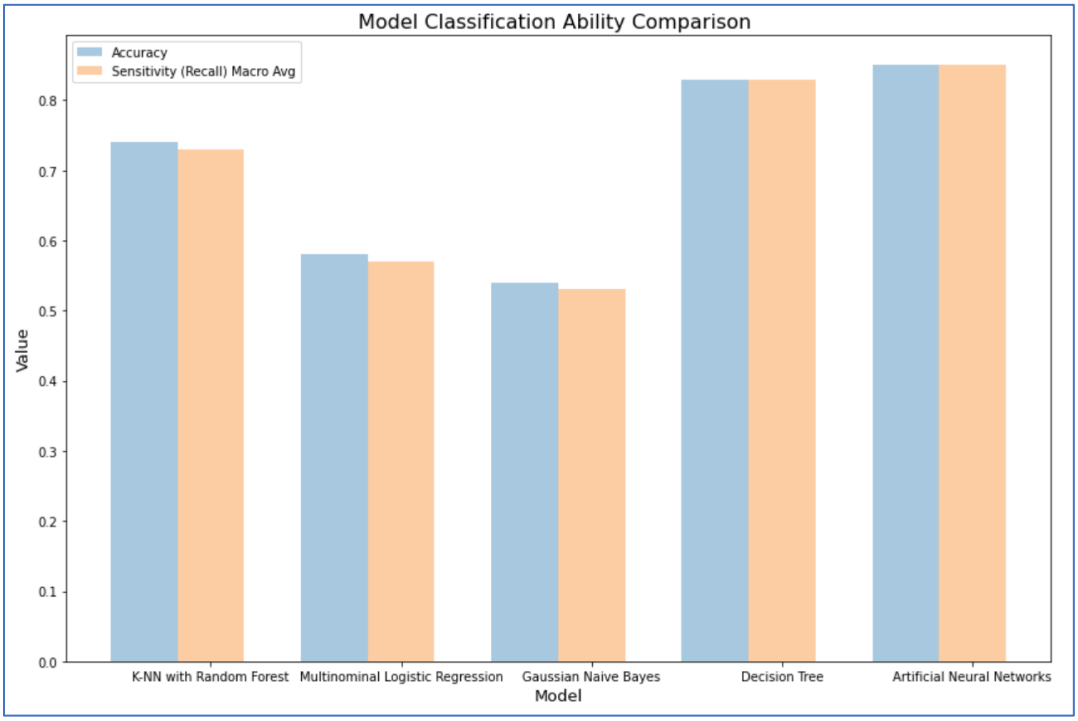

Fig 40. Model Performance Comparison Bar chart

From the above plot and table, we can observe that Artificial Neural Networks model gives the highest accuracy score and Sensitivity (Recall) Macro Average value. In addition, accuracy score and Sensitivity (Recall) Macro Average value are very close for each model. This is because the database is balanced, and each value indicates the model ability on classify new data as the right class. the Therefore, Artificial Neural Networks model is the best model to do classification job for the Maternal Health Risk dataset.

Conclusion

In the future, once we collected information about age, systolic blood pressure, blood sugar, body temperature, and heart rate from pregnant women, we could put the information in our artificial neural networks model. Their maternal health condition of them could be predicted

with high accuracy. Therefore, medical experts can do a quick initial judgment about the future treatment. With timely and proper treatment, the newborn mortality rate will decrease. In addition, in some rural areas where medical support is not sufficient, this model can be a great assistant for doctors, which can give them some reliable information on maternal health conditions without using expensive medical machines. It also saves money for the clients. All in all, this model will be a great source to reduce the pressure on the current national medical system.