CSE422 LAB Project Report

Title: Maternal Health Risk Prediction

Prepared by:

Ishraq Kamal Adib - 24341160
Afia Abida Shohid - 22201457

Date: 08.01.2025

Table of Contents :

1. Introduction

The project aims to create a predictive model for maternal health risk levels, focusing on accurately classifying pregnant individuals into low, medium, or high-risk categories based on health indicators. This is crucial for early risk identification in maternal healthcare, enabling timely interventions and improved outcomes for both mothers and babies. Early and precise risk identification can help reduce complications and improve overall maternal health. The project uses machine learning classification algorithms to compare their performance on a dataset of maternal health indicators and risk levels, aiming to identify the most effective method for predicting risk and contributing to better preventative care.

2. Dataset Description:

- **Source:**

  Link: **maternal+health+risk**
  Reference: https://archive.ics.uci.edu/dataset/863/maternal+health+risk

- **Description:**

  Features: The dataset contains 6 features, including Age, SystolicBP, DiastolicBP, BS, BodyTemp, HeartRate and RiskLevel as the target.

  Problem Type: Multiclass Classification problem. The target variable RiskLevel has 3 discrete categories: low risk, mild risk and high risk.

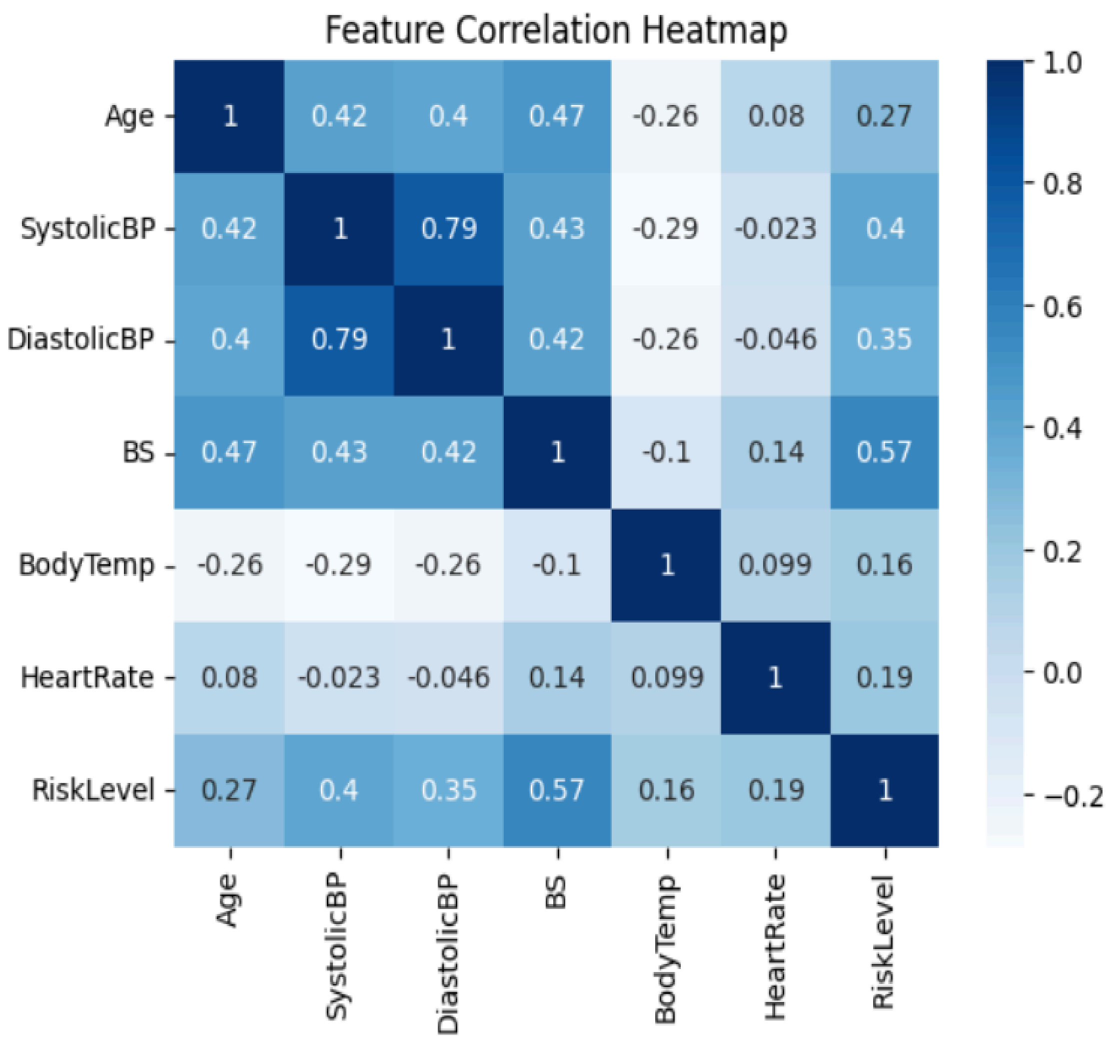
  Number of Data Points: 7098

  Feature Types:

- Quantitative: Age, SystolicBP, DiastolicBP, BS, BodyTemp,HeartRate
- Categorical: RiskLevel

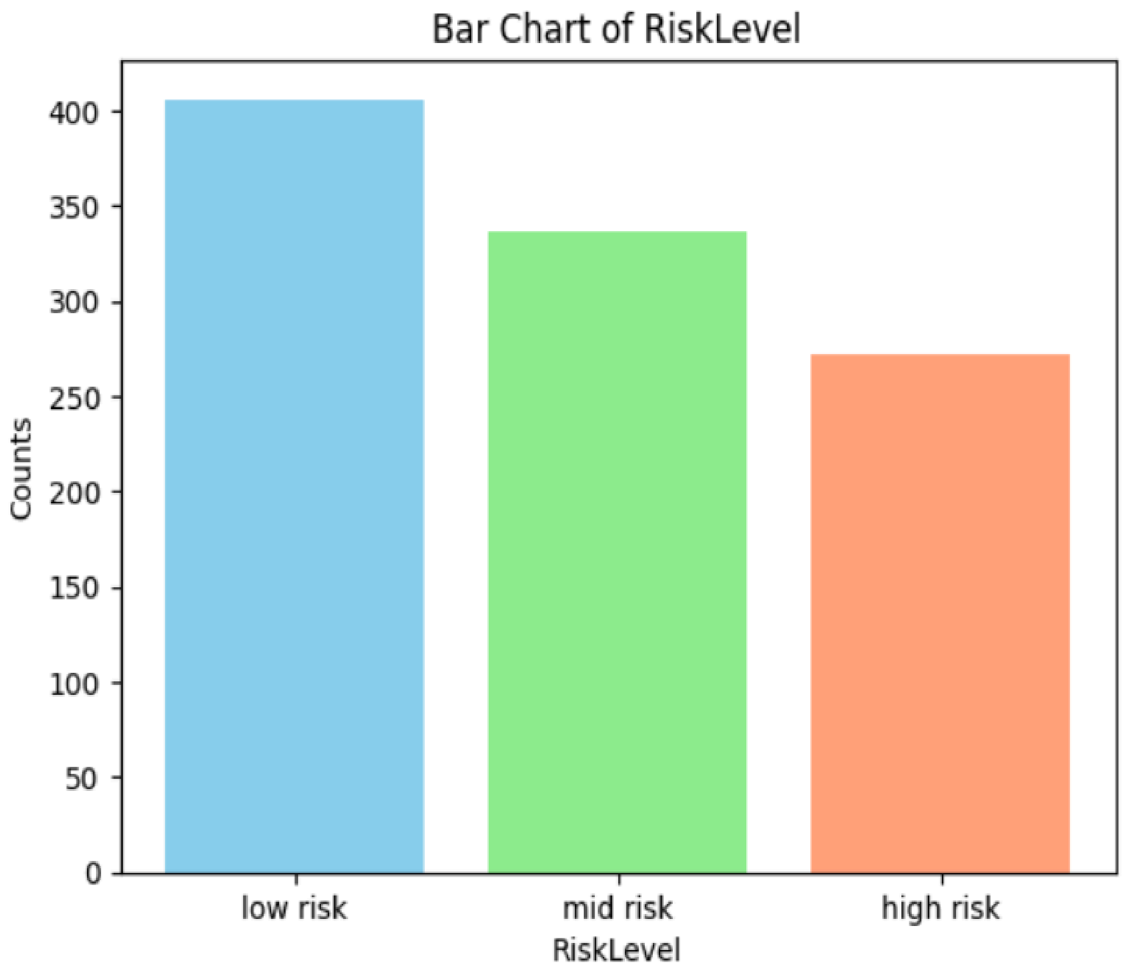
- **Correlation Analysis:**

  A heatmap was created to visualize the correlation between quantitative features.



- **Imbalanced dataset:**

  The classes of the target variable **RiskLevel** are imbalanced. A bar chart was plotted to show the distribution of instances across categories.



3. Dataset Pre-Processing:

Faults Identified:

- **Categorical values:** The target feature RiskLevel has non numeric categorical values which requires encoding.

Solutions:

- **Encoding:** Mapping was done to convert the categorical values of the RiskLevel column into numerical representations by assigning specific values to the classes.

4. Feature Scaling:

The input features of the dataset were scaled using MinMaxScaler to ensure that all the features contribute equally to the models. It also prevented the dominance of larger values of the columns of the data so that the result is not biased towards a certain feature.

5. Dataset Splitting:

The dataset was split into training and testing sets:

- The **training set** contains 70% of the data.
- The **testing set** contains 30% of the data.

6. Model Training & Testing:

The following models were applied to train the dataset:

1. Neural Networks
2. Naive Bayes
3. Logistic Regression
4. Decision Tree

7. Model Selection/Comparison Analysis:

- **Accuracy:** It measured the overall prediction accuracy for each model.
- **Precision:** It measured the proportion of correctly predicted positive observations out of all predicted positive observations.
- **Recall:** It measured the proportion of correctly predicted positive observations out of all actual positive observations.
- **F1-Score:** This combined Precision and Recall, and is defined as a sample weighted average (harmonic mean) of precision and recall.
- **Confusion Matrix:** It visualized the relation between the predicted values and true values for each model.

Evaluation Metrics:

1. Neural Network Results:

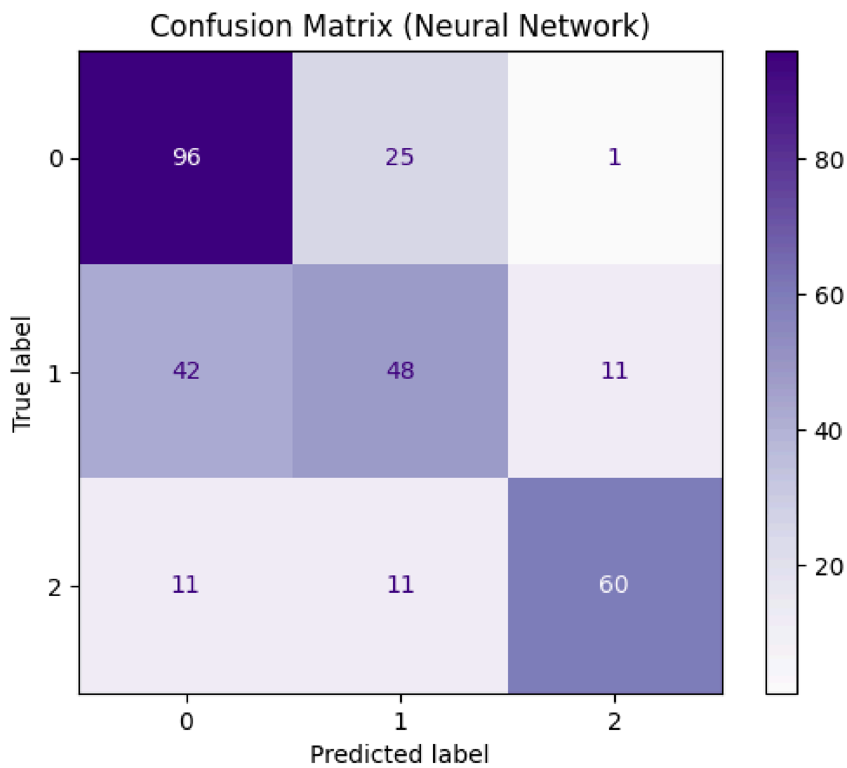
Accuracy (Neural Network): 0.6688524590163935

```
Classification Report (Neural Network):
              precision    recall  f1-score   support

           0       0.64      0.79      0.71       122
           1       0.57      0.48      0.52       101
           2       0.83      0.73      0.78        82

    accuracy                           0.67       305
   macro avg       0.68      0.66      0.67       305
weighted avg       0.67      0.67      0.66       305


Confusion Matrix (Neural Network):
 [[96 25 1]
  [42 48 11]
  [11 11 60]]
```



2. Naive Bayes Results:

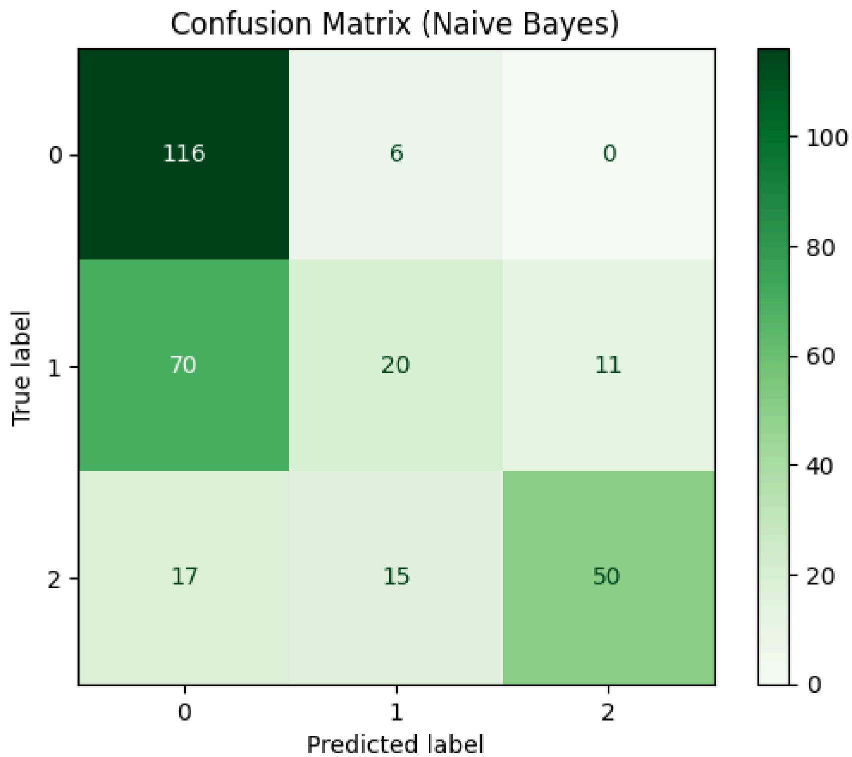
```
Accuracy (Naive Bayes): 0.6098360655737705

Classification Report (Naive Bayes):
              precision    recall  f1-score   support

           0       0.57      0.95      0.71       122
           1       0.49      0.20      0.28       101
           2       0.82      0.61      0.70        82

    accuracy                           0.61       305
   macro avg       0.63      0.59      0.56       305
weighted avg       0.61      0.61      0.57       305

Confusion Matrix (Naive Bayes):
 [[116 6  0]
  [70  20 11]
  [17  15 50]]
```



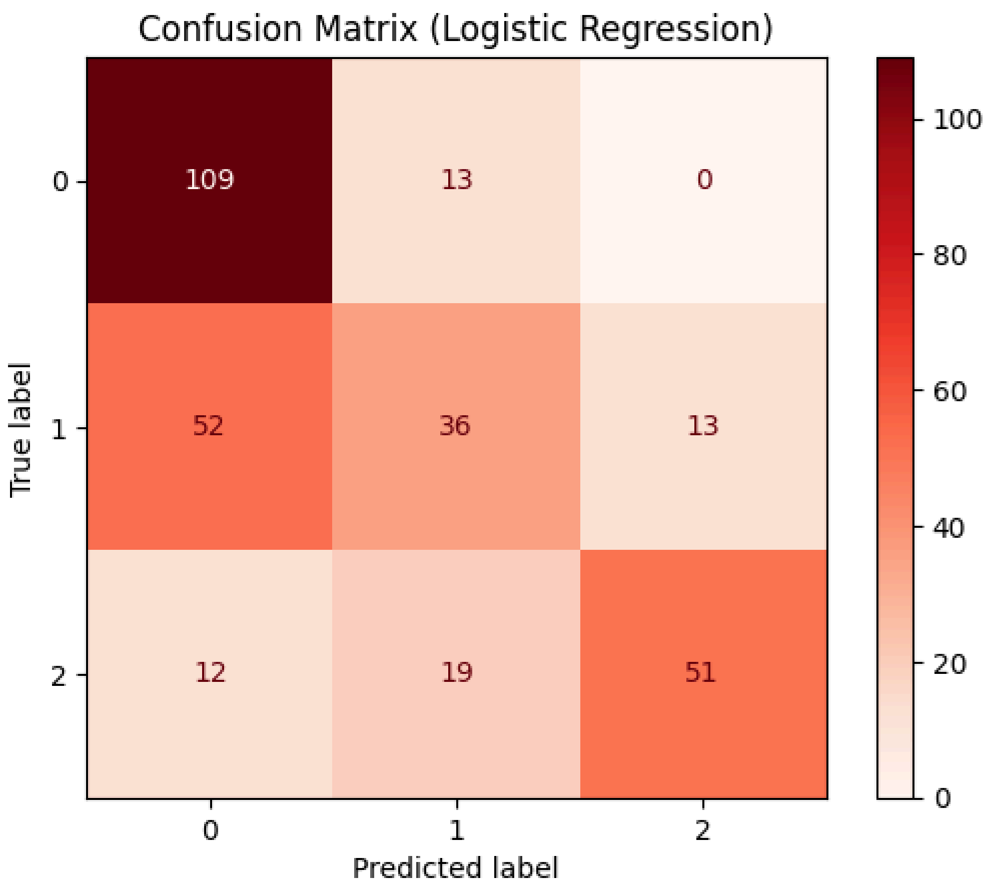
3. Logistic Regression Results:

```
Accuracy (Logistic Regression): 0.6426229508196721
Classification Report (Logistic Regression):
              precision    recall  f1-score   support

           0       0.63      0.89      0.74       122
           1       0.53      0.36      0.43       101
           2       0.80      0.62      0.70        82

    accuracy                           0.64       305
   macro avg       0.65      0.62      0.62       305
weighted avg       0.64      0.64      0.62       305

Confusion Matrix (Logistic Regression):
 [[109 13 0]
  [52  36  13]
  [12  19  51]]
```



4. Decision Tree Results:

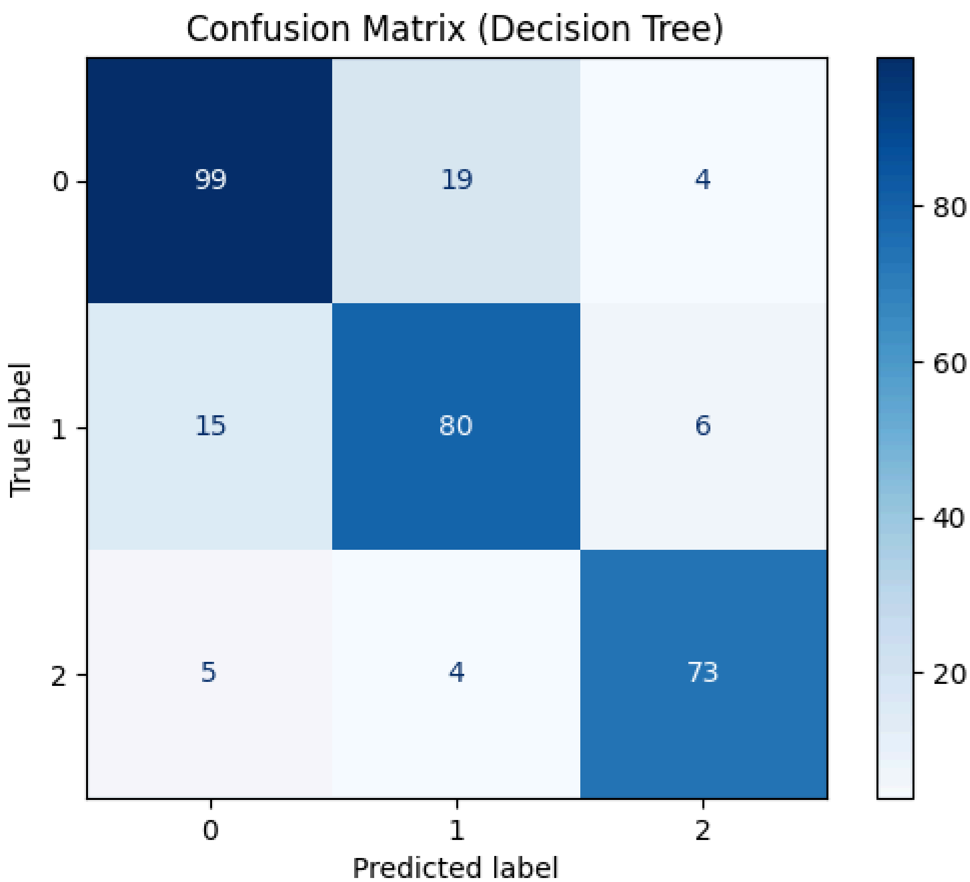
Accuracy (Decision Tree): 0.8262295081967214

Classification Report (Decision Tree):

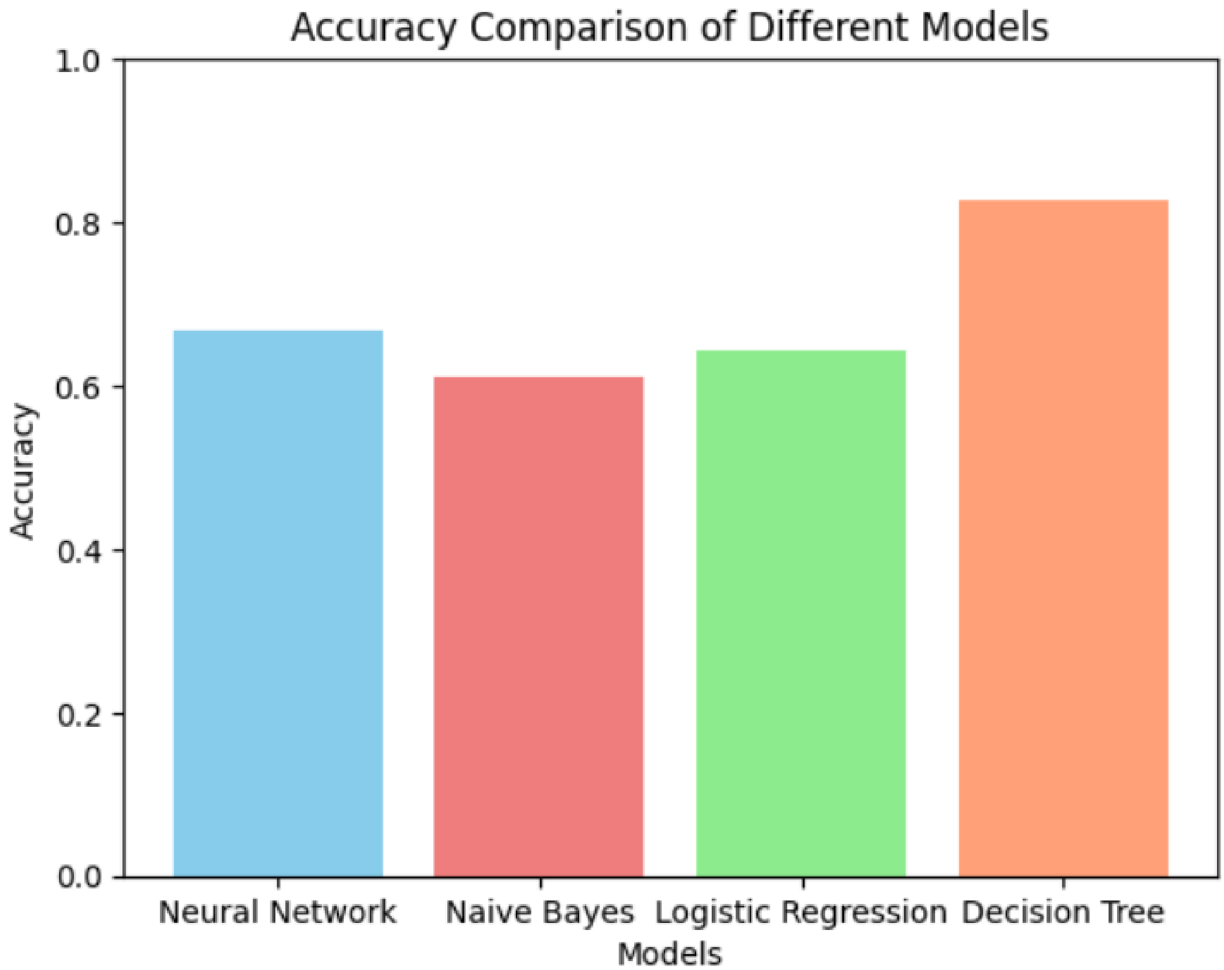
	precision	recall	f1-score	support
0	0.83	0.81	0.82	122
1	0.78	0.79	0.78	101
2	0.88	0.89	0.88	82
accuracy			0.83	305
macro avg	0.83	0.83	0.83	305
weighted avg	0.83	0.83	0.83	305

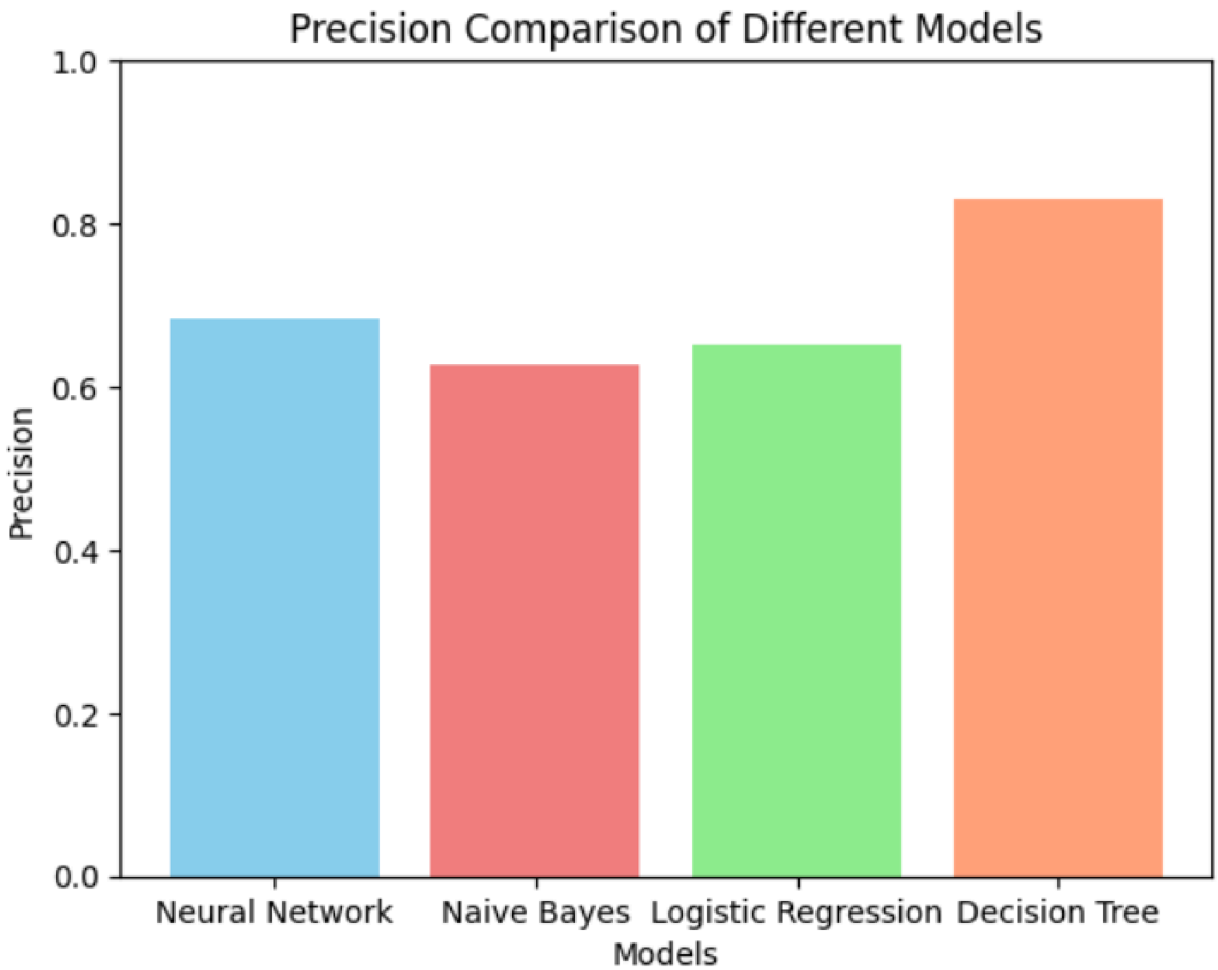
Confusion Matrix (Decision Tree):

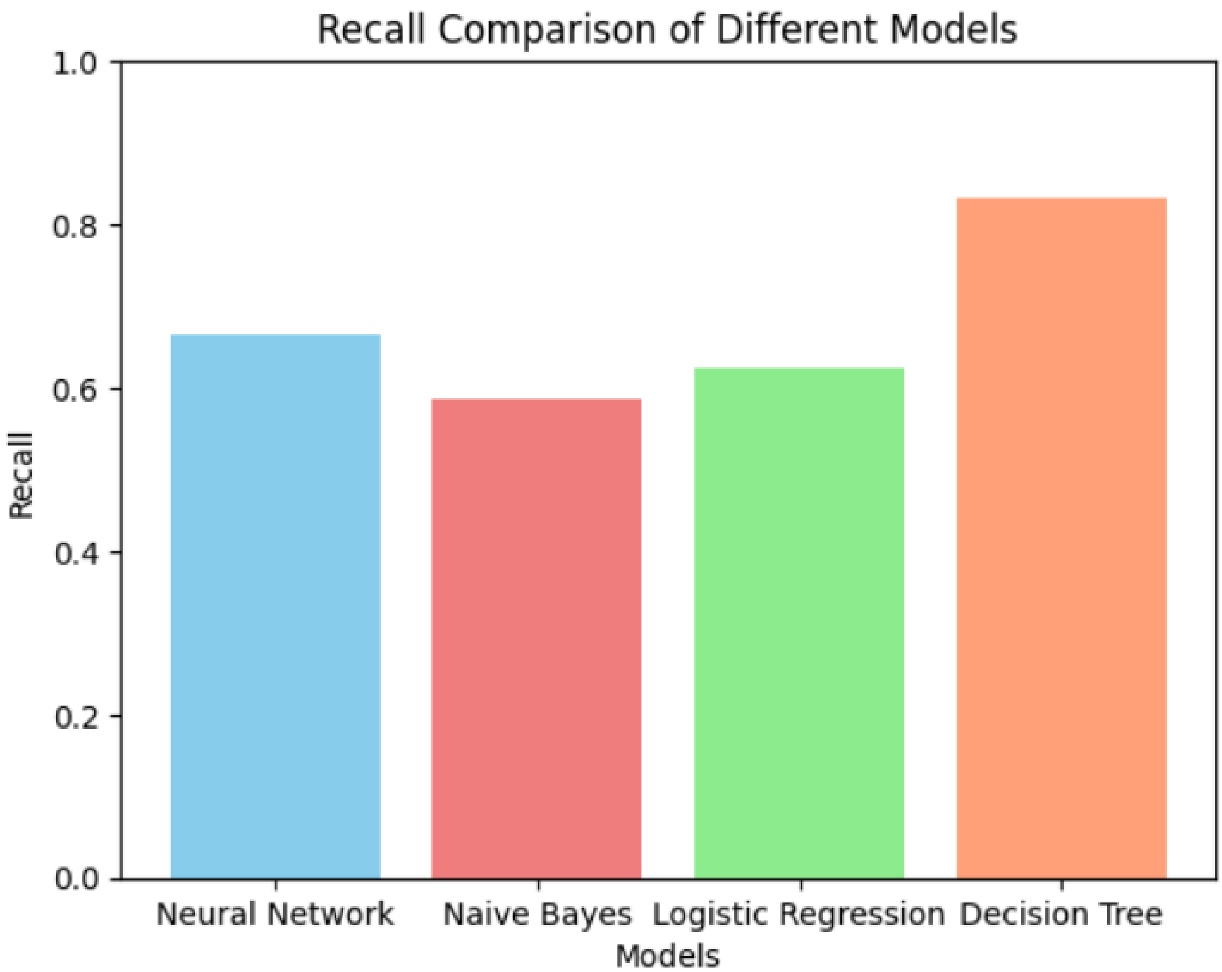
```
[[99 19 4]
 [15 80 6]
 [5  4  73]]
```



Model Comparison:




Precision Comparison of Different Models
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Neural Network
Naive Bayes
Logistic Regression
Decision Tree
Models



8. Conclusion:

In this project, we aimed to predict maternal health risks using four machine learning models: Neural Network, Naive Bayes, Logistic Regression, and Decision Tree. The dataset was pre-processed, the numerical features were scaled and the models' performance was assessed using accuracy, precision, and recall. The results show that the Decision Tree model achieved the highest accuracy, precision, and recall, indicating its superior ability to learn complex patterns in data. Other models like Neural Network and Logistic Regression also performed reasonably well, but Decision Tree''s performance was significantly better. The results were visually represented through confusion matrices and bar charts, providing insights into the models' strengths and weaknesses in classifying different risk levels. Future work could involve exploring advanced algorithms, hyperparameter tuning, and feature engineering to identify more predictive features. A larger dataset could also improve the models' generalizability and evaluation effectiveness. Overall, this project successfully demonstrated the viability of using machine learning models for health risk level classification. With further refinements and advancements, such models can play a crucial role in maternal health risk prediction and decision making.