**Report on Predicting Maternal Health Risks**

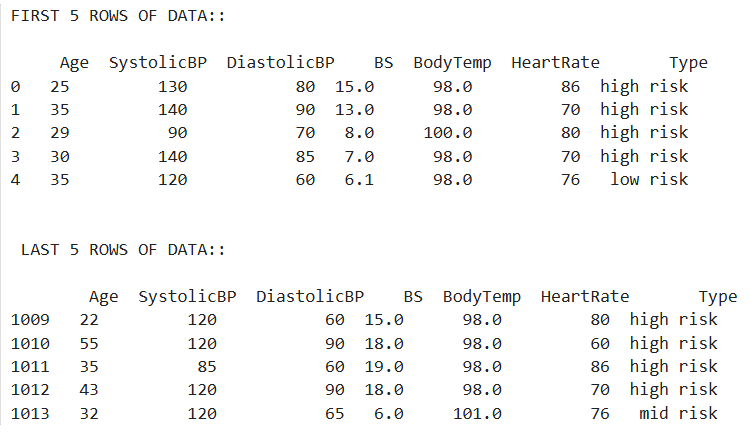
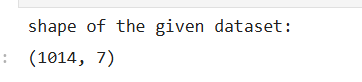
The dataset used for this analysis is the Perinatal Risk Information dataset, which comprises various health parameters such as age, Systolic blood pressure, diastolic blood pressure, Body Temperature, and Heat rate that could potentially influence maternal health risk type. The primary goal is to predict maternal health risks based on these parameters.

Classification is the suitable machine learning category for this problem because the objective is to categorize instances into distinct risk types. Each input (independent variable) will lead to a specific output class (dependent variable), making classification algorithms ideal for this task.

In the given dataset, the **Independent variables** are age, systolicBP, diastolicBP, Body Temperature, and Heat Rate, and the **Dependent Variable** is Type.

**Understand the data:**

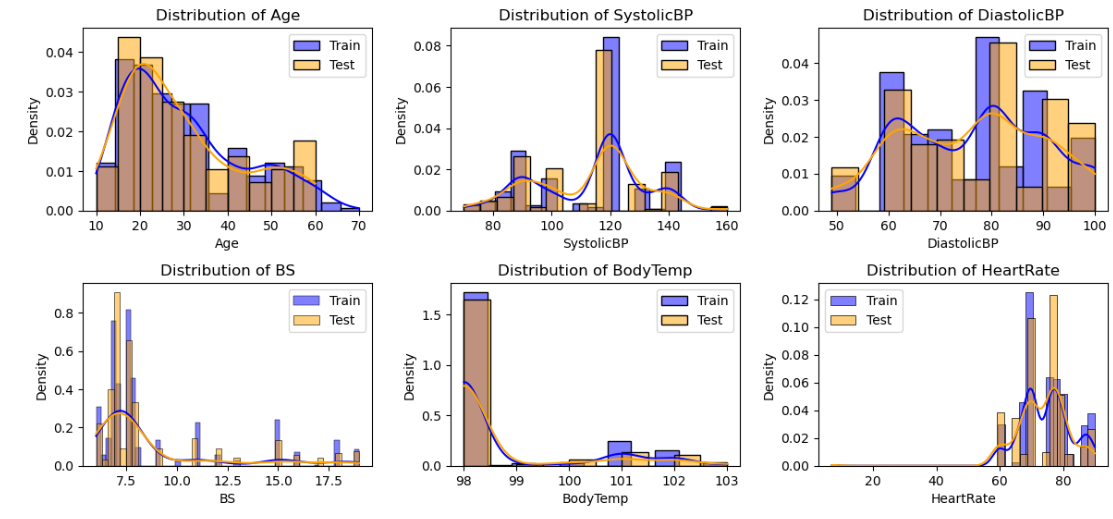
Explore the data by fetching the first and last 5 rows of the data and shape to understand the rows and columns of the data.

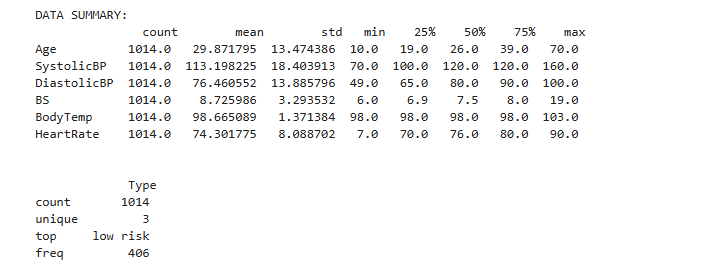
**Distribution Analysis:**

The given dataset is split into 70 % training and 30% testing sets, perform Distribution analysis to ensures that both sets adequately represent the entire dataset's characteristics.

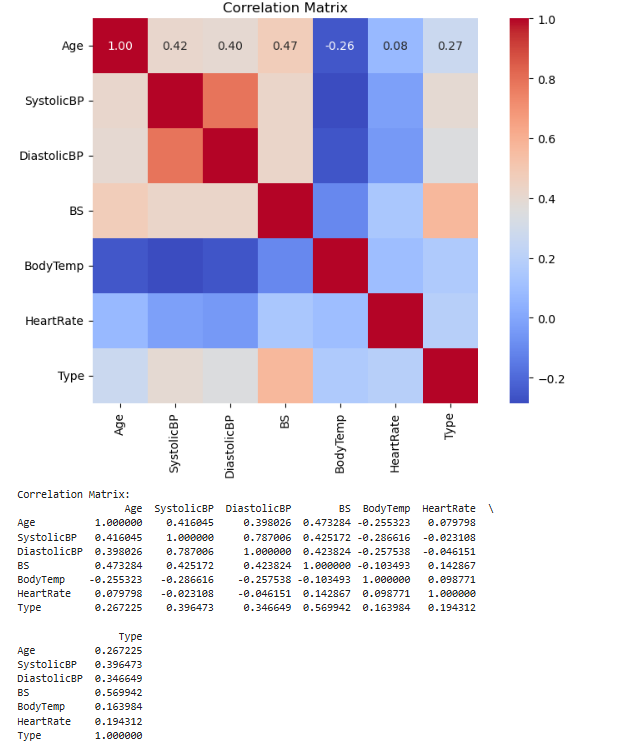
Compare the distribution of individual features for both train and test data using histograms.



View the data Statistics Summary to understand the data better

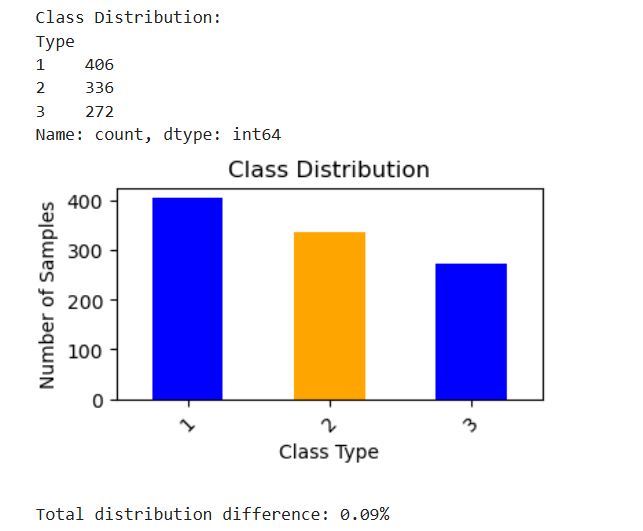


Correlation Analysis helps to understand the relationship between independent and dependent variables. Here, the body temperature parameter has a low correlation, which we can remove to get a better result.



**Imbalance dataset:**

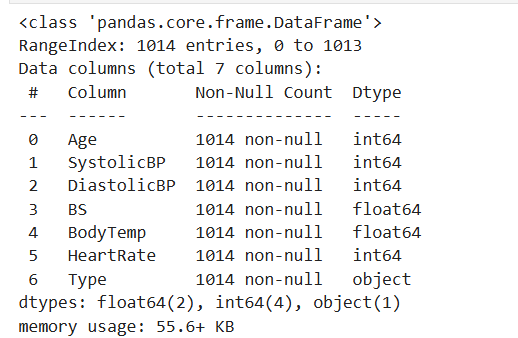
An examination of the distribution of the dependent variable reveals whether the dataset is imbalanced. A dataset is considered imbalanced if one class significantly outnumbers the other, which could lead to biased model predictions. In the given dataset, we have difference of 0.09% imbalance data.

****

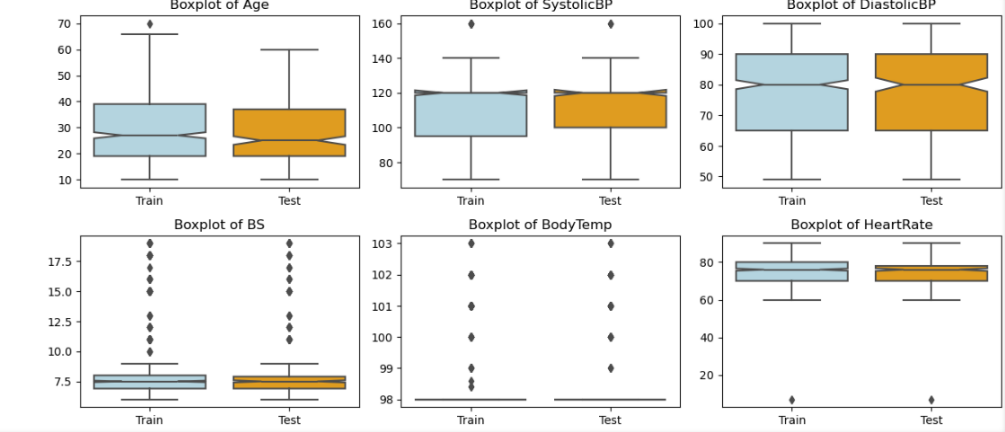
* The dataset is considered imbalanced if the output of class\_counts shows a significant disparity (e.g., 90% of samples are in one class while 10% are in another).

**Data Preprocessing:**

Check and remove Null Values or missing values: Missing values can lead to inaccurate model predictions or even model failures. No null values found in the given dataset.

****

Check and remove Outliers: Outliers can skew results and affect model performance. Identifying and handling them can lead to a more robust model. We have only a few outliers in the given dataset, which we checked using boxplot.



**Feature Scaling:**

Normalizing or standardizing features ensures that all independent variables contribute equally to the distance calculations in algorithms. Standardization is typically crucial for algorithms that rely on distance measures, like Decision Trees and Random Forests. Performed Zscore normalization in the algorithm.

ClassificationAlgorithm:

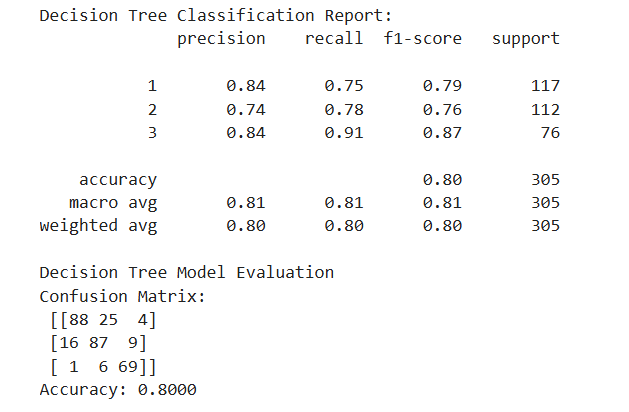
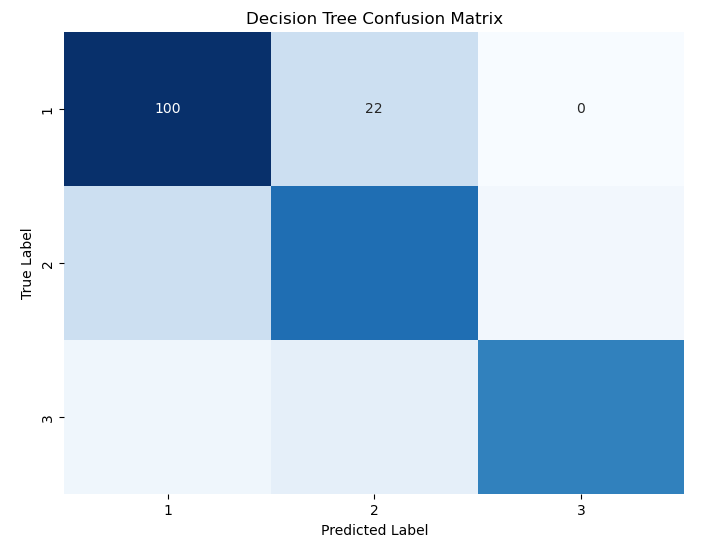
**Decision Tree Classifier:** A decision tree classifier builds a model in the form of a tree structure, where each node represents a decision based on the value of a feature. The leaf nodes represent the output classes. Decision trees are easy to interpret and can handle numerical and categorical data. Source: Mitchell, T. M. (1997). "Machine Learning."

The decision tree classifier performed well, with an overall accuracy of **80.33%** on the dataset. Below is a breakdown of its performance for each class:

Class 1: Precision: 0.79 (79% of predicted Class 1 instances were correct); Recall: 0.82 (82% of actual Class 1 instances were correctly predicted); F1-score: 0.81 (balance between precision and recall); Support: 122 instances

Class 2: Precision: 0.71 (71% of predicted Class 2 instances were correct); Recall: 0.75 (75% of actual Class 2 instances were correctly predicted); F1-score: 0.73; Support: 101 instances

Class 3: Precision: 0.96 (96% of predicted Class 3 instances were correct); Recall: 0.84 (84% of actual Class 3 instances were correctly predicted); F1-score: 0.90; Support: 82 instances

**** 

Confusion Matrix:

The confusion matrix indicates the following:

Class 1: Out of 122 instances, 100 were correctly classified as Class 1, while 22 were misclassified as Class 2.

Class 2: Out of 101 instances, 76 were correctly classified, with 22 misclassified as Class 1 and 3 as Class 3.

Class 3: Out of 82 instances, 69 were correctly classified, while 9 were misclassified as Class 2, and 4 as Class 1.

**Random Forest Classifier:** Random forests are an ensemble method that constructs multiple decision trees and merges them to produce a more accurate and stable prediction. Each tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions from all the trees, thus reducing overfitting.

Source: Breiman, L. (2001). "Random Forests."

The data was trained using the Random Forest classification algorithm, and the report of the algorithm after it was executed is below.

Class-wise Metrics:

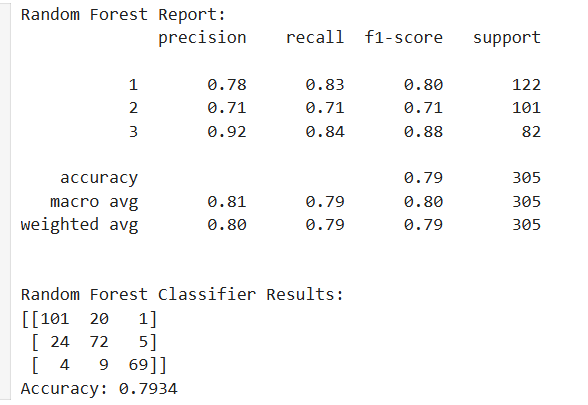
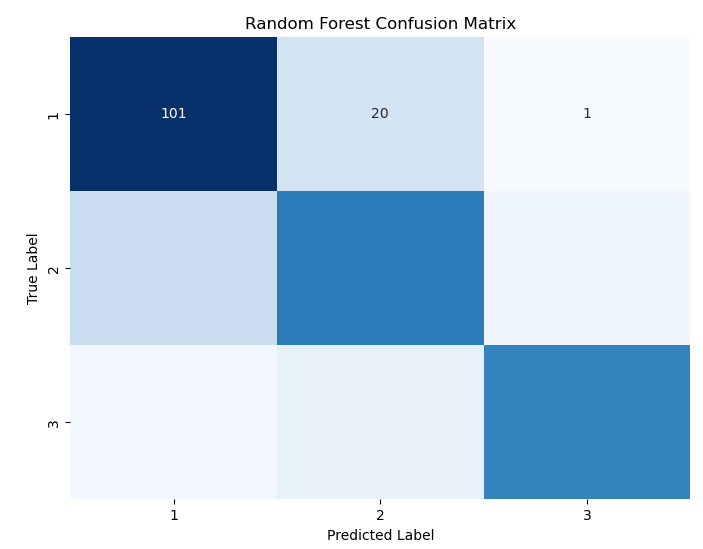
Class 1: Precision is 0.78, recall is 0.83, and the f1-score is 0.80. This means the model correctly predicted 78% of class 1 cases, and out of all actual class 1 cases, 83% were correctly identified.

Class 2: Precision is 0.71, recall is 0.71, and the f1-score is 0.71. This shows a balanced prediction rate, but with moderate performance for this class.

Class 3: Precision is high at 0.92, recall is 0.84, and the f1-score is 0.88, indicating strong performance in predicting class 3, though some cases were missed (84% recall).

Overall Performance:

Accuracy: The model achieved **79.34%** accuracy, meaning it correctly classified around 79% of the total 305 instances.

**** 

Confusion Matrix for Random Forest:

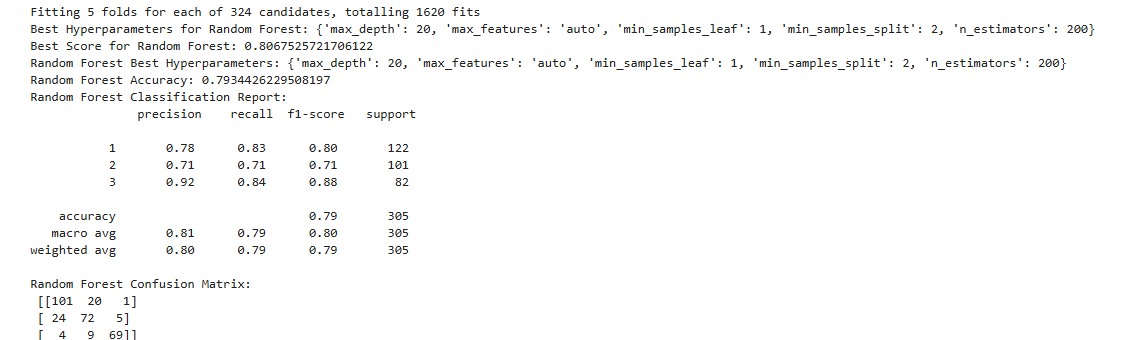
Class 1 was most often confused with class 2 (20 misclassifications).

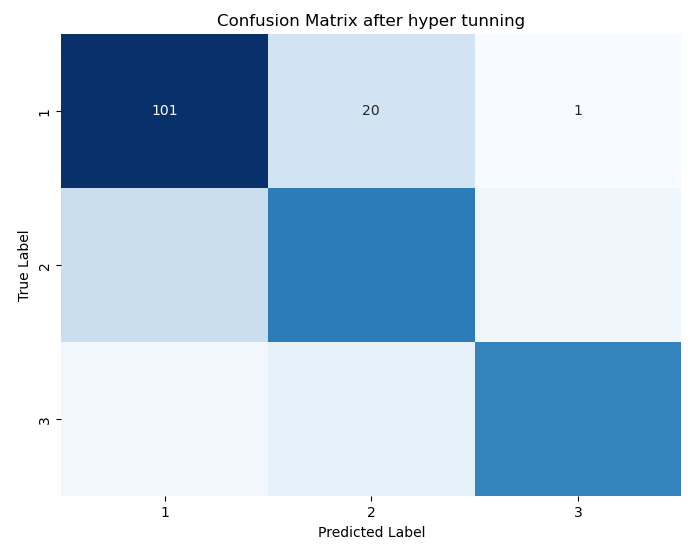
Class 2 was misclassified into class 1 (24 instances) and class 3 (5 instances).

Class 3 mainly was well-identified, with minor confusion in class 2 (9 instances).

**Hyperparameter Tunning:**

Hyperparameter tuning was performed using GridSearchCV to find the best combination of parameters for the Random Forest model.





Tuning the hyperparameters improved the model's performance from 79.34 to 80.67 by optimizing the combination of trees and their respective depths, resulting in better accuracy and reduced overfitting.

Ethical Implications in Healthcare

The application of machine learning models in healthcare, particularly in predicting maternal health risks, raises several ethical considerations:

* Bias: If the training data is biased or not representative of the entire population, the model may inadvertently reinforce existing health disparities.
* Privacy: Sensitive health information must be handled carefully, ensuring patient confidentiality.
* Accountability: Accountability becomes crucial in cases where predictions lead to critical health decisions. Misdiagnoses resulting from model errors could have severe implications.

Reflecting on these ethical concerns is vital for deploying machine learning models in real-world healthcare settings.

The involved understanding a maternal health risk dataset, conducting data exploration and pre-processing, applying classification algorithms, performing hyperparameter tuning, and evaluating the models. The ethical implications highlight the necessity of responsible AI practices in healthcare, emphasizing fairness, accountability, and transparency in model predictions.

By adopting these approaches, we can develop robust machine-learning solutions that contribute positively to maternal health outcomes.