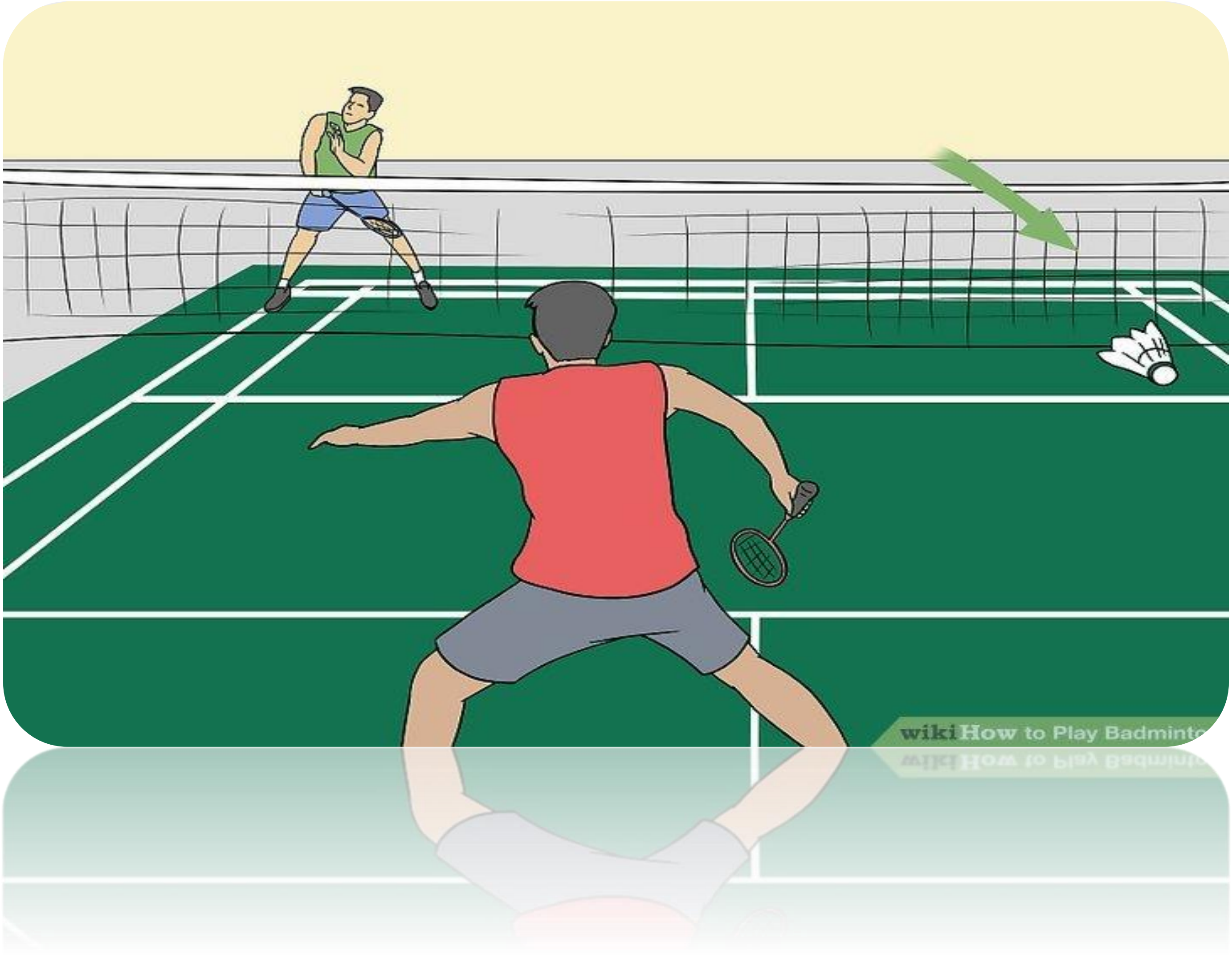
Fahad Ur Rehman ~ 18228



Badminton Sport Analysis

In the past few years, people have loved to play badminton and badminton was the passion of the majority of people in the community, but somehow players of badminton have shown low interest in the game and are not playing badminton as that was their favorite sport for the players. Therefore, It is very necessary that we analyze the reasons that are the obstacles to players’ passion which are making them not play badminton sport, and to prevent badminton sport from vanishing as it is slowly dying in the community.

Fahad Ur Rehman ~ 18228

Assumptions

1. Badminton players avoid playing due to some weather conditions.
2. When it's raining players avoid playing badminton.
3. The most significant factor influencing players to play badminton is the clear weather and normal temperature.

Research Question

1. What are the columns in the dataset, and what types of data do they contain (e.g., numerical, categorical)?
2. Provide a summary of the dataset, including the number of rows, columns, and basic statistics (mean, median, standard deviation, etc.) for numerical columns.
3. Check for missing values in the dataset. Which columns have missing values, and what is the proportion of missing data in each column?
4. Describe the methods you used to handle missing data. Did you choose to remove, impute, or use another method for dealing with missing values? Justify your choice.-
5. Identify any outliers in the numerical columns. How did you detect them, and what method did you use to handle them?
6. Explore relationships between variables using scatter plots, or heatmaps. Discuss any interesting patterns or correlations you observe.
7. Describe the type of Naive Bayes model you chose (e.g., Gaussian, Multinomial, Bernoulli) and why it is appropriate for your dataset.
8. What are your Interpretations of the metrics you have used for model evaluation?

9. Use scaling and normalization techniques if necessary and observe their impact on the model’s performance.

Answers to the questions

1. What are the columns in the dataset, and what types of data do they contain (e.g., numerical, categorical)?

   Founded Answer: The columns in the dataset are five, which are 'Outlook', 'Temperature', 'Humidity', 'Wind', and 'Play_Badminton' and the types of data that they contain are object or categorical.

2. Provide a summary of the dataset, including the number of rows, columns, and basic statistics (mean, median, standard deviation, etc.) for numerical columns.

   Founded Answer: The data has info about the weather and if people can play badminton. The dataset comprises 36 rows and 5 columns. The dataset includes data or records like outlook (like sunny or rainy), temperature (like hot or cold), humidity (how damp it is), and wind strength. Most days are cloudy, the temperature is usually cool, and it's often humid. Weak winds are common. People usually don't play badminton, with 24 "No" and 12 "Yes" answers.

   Further checking the distribution of data of the dataset, reveals that the dataset includes categorical variables such as Outlook, Temperature, Humidity, Wind, and Play_Badminton, each with balanced distributions. Outlook has 12 instances each of Overcast, Sunny, and Rainy. Temperature is split evenly among Cool, Mild, and Hot, each with 12 occurrences. Humidity and Wind both have equal counts for their categories: 18 instances each of High and Normal humidity, and 18 instances each of Weak and Strong wind. The Play_Badminton outcome shows 24 'No' and 12 'Yes'.

Notably, badminton is rarely played, with 24 occurrences of "No" compared to 12 "Yes" instances.

3. Check for missing values in the dataset. Which columns have missing values, and what is the proportion of missing data in each column?

   Founded Answer: The dataset has no missing values which means that the dataset is complete, without any gaps in the recorded information.

4. Describe the methods you used to handle missing data. Did you choose to remove, impute, or use another method for dealing with missing values? Justify your choice.

   Founded Answer: Upon utilizing the "isnull().sum()" method to check for missing data in the data frame. Since the output indicated zero missing values for all columns, therefore, did not need to employ any methods for handling missing data. There was no need for removal, imputation, or any other method, as the dataset was complete without any gaps in the recorded information.

5. Identify any outliers in the numerical columns. How did you detect them, and what method did you use to handle them?

   Founded Answer: All columns have 'object' type of data, indicating that they contain categorical data rather than numerical data. Therefore, traditional methods for detecting outliers in numerical columns, such as calculating quartiles and identifying values outside a certain range, are not applicable in this case.

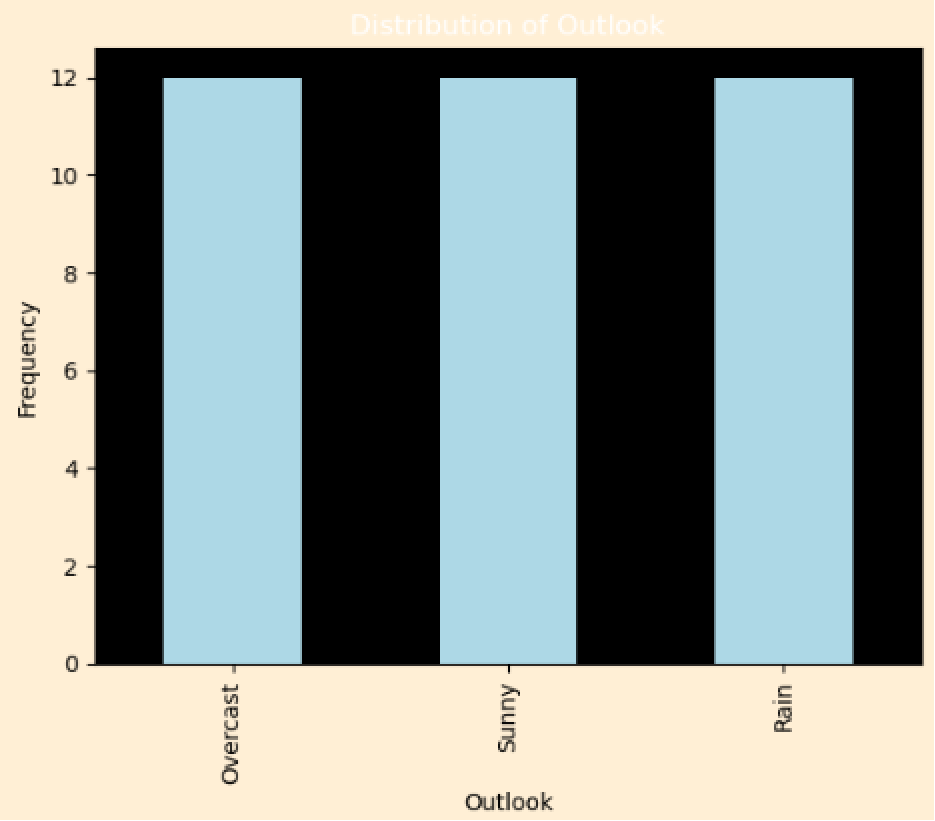
   The dataset is so small that we can find and detect outliers by analysis or looking through the dataset and there are no outliers.

6. Explore relationships between variables using scatter plots, or heatmaps. Discuss any interesting patterns or correlations you observe.

Fahad Ur Rehman ~ 18228

Founded Answer:

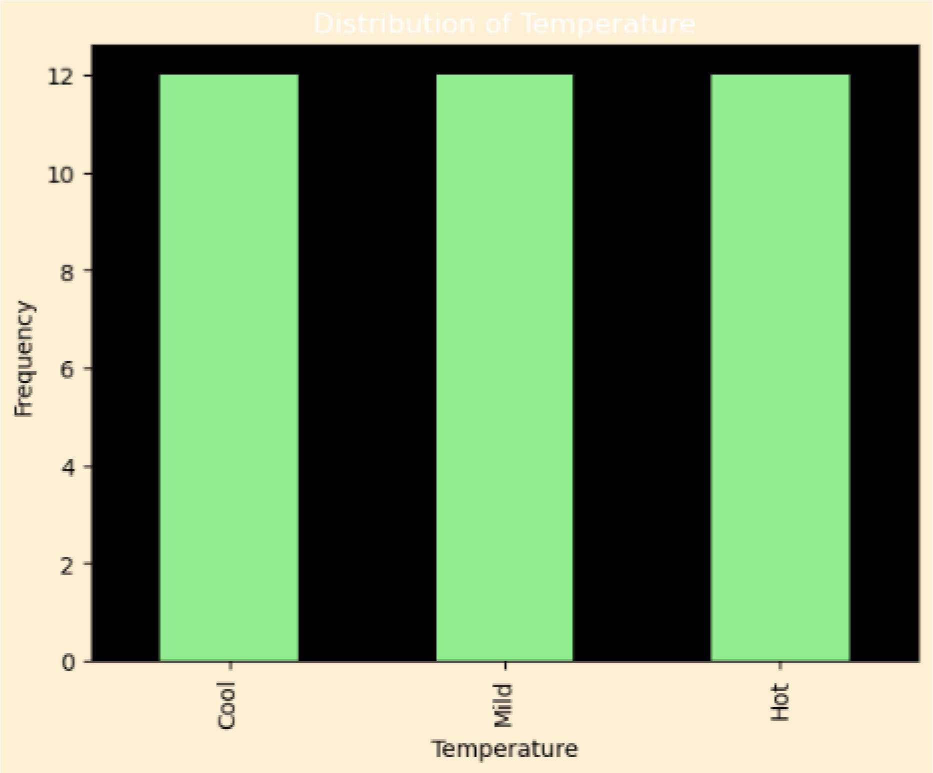
Let’s explore each feature and then the relationships between them



The insights from the plot are:

The column `Outlook` represents three groups of data which are `Overcast`, `Sunny`, and `Rain`. The column group's frequencies are equally distributed. The highest frequency is `12`, which is the same for every group of the column.

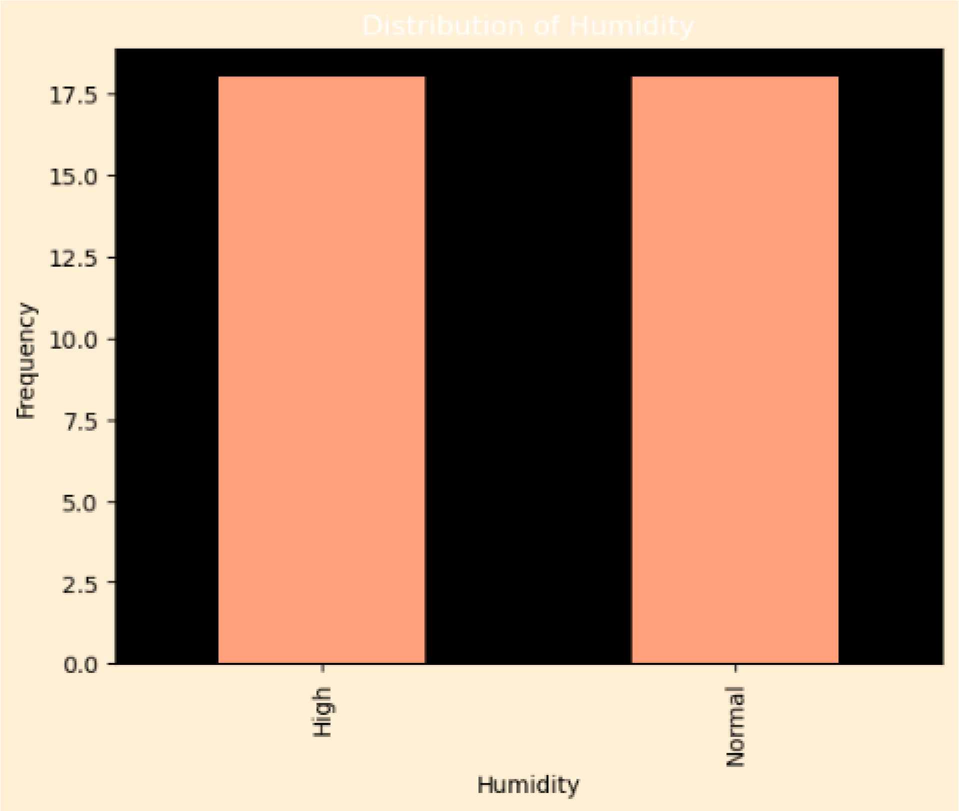
Fahad Ur Rehman ~ 18228



The insights from the plot are:

The column `Temperature` represents three groups of data which are `COOL`, `MILD`, and `HOT`. The column group's frequencies are equally distributed. The highest frequency is `12`, which is the same for every group of the column.

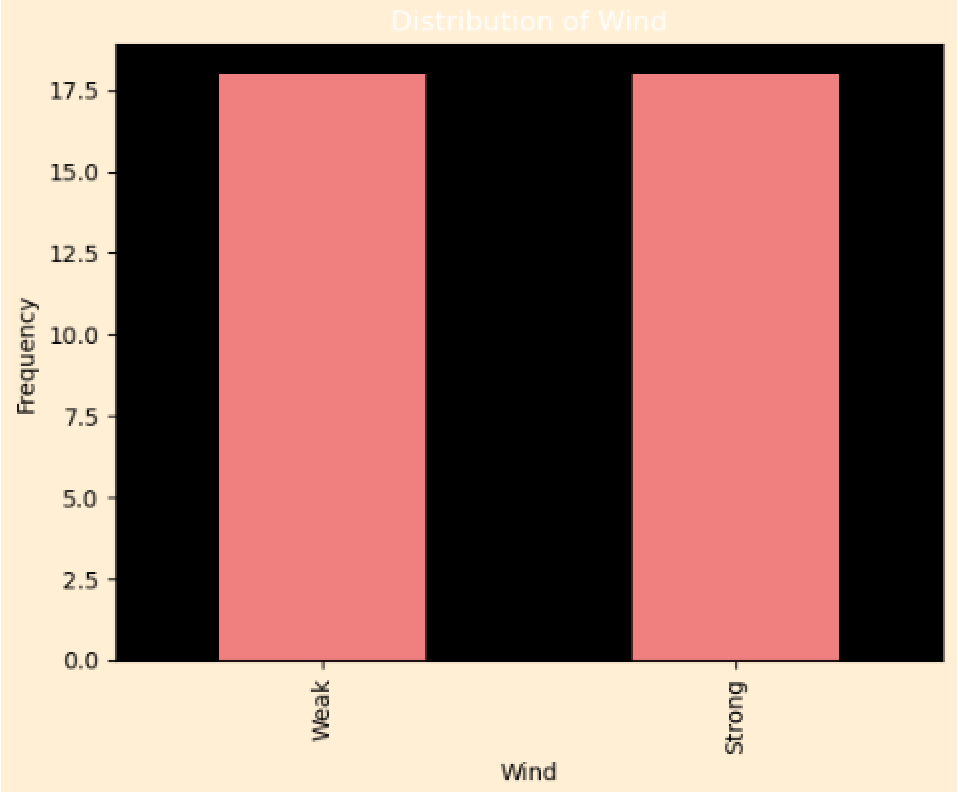
Fahad Ur Rehman ~ 18228



The insights from the plot are:

The column `Humidity` represents two groups of data which are `High` and `Normal`. The column group's frequencies are equally distributed. The highest frequency is `18`, which is the same for both groups of the column.

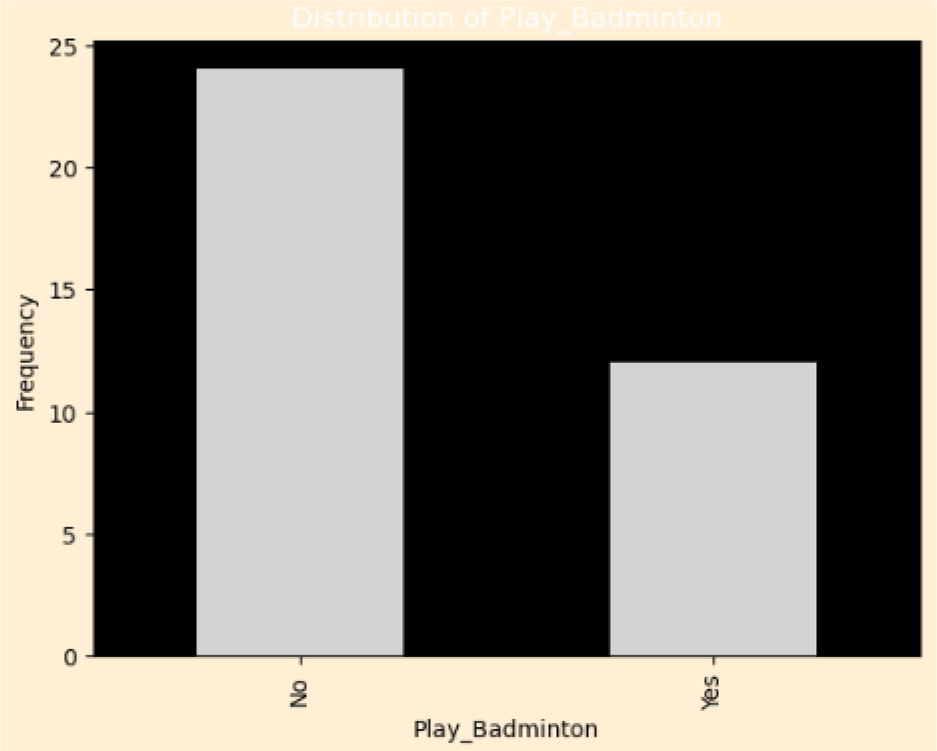
Fahad Ur Rehman ~ 18228



The insights from the plot are:

The column `WIND` represents two groups of data which are `Weak` and `Strong`. The column group's frequencies are equally distributed. the highest frequency is `18`, which is the same for both groups of the column.

Fahad Ur Rehman ~ 18228



The insights from the plot are:

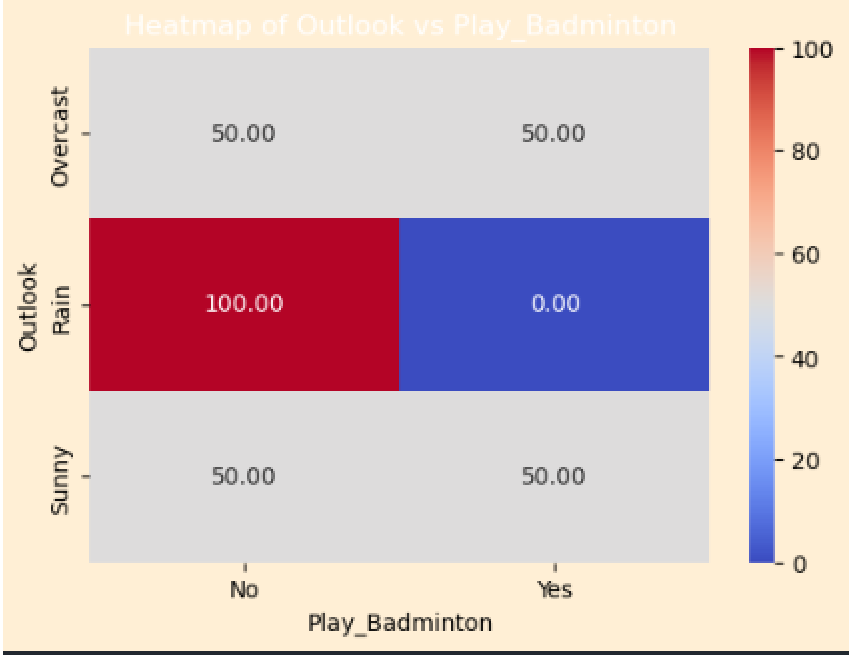
The column `Play Badminton` represents two groups of data which are `No` and `Yes`. The column groups are not equally distributed. the no group of the column has the highest frequency which is `24`, and the `yes` group of data frequency is `13`. Which reveals a shocking insight.

Shocking Insight:

The visualization reveals that nowadays due to `weather inconsistency` people `don't play badminton` and we have to `analyze the weather condition` on which conditions the people `avoid playing badminton`.

Fahad Ur Rehman ~ 18228

Relation of Outlook and Play Badminton columns

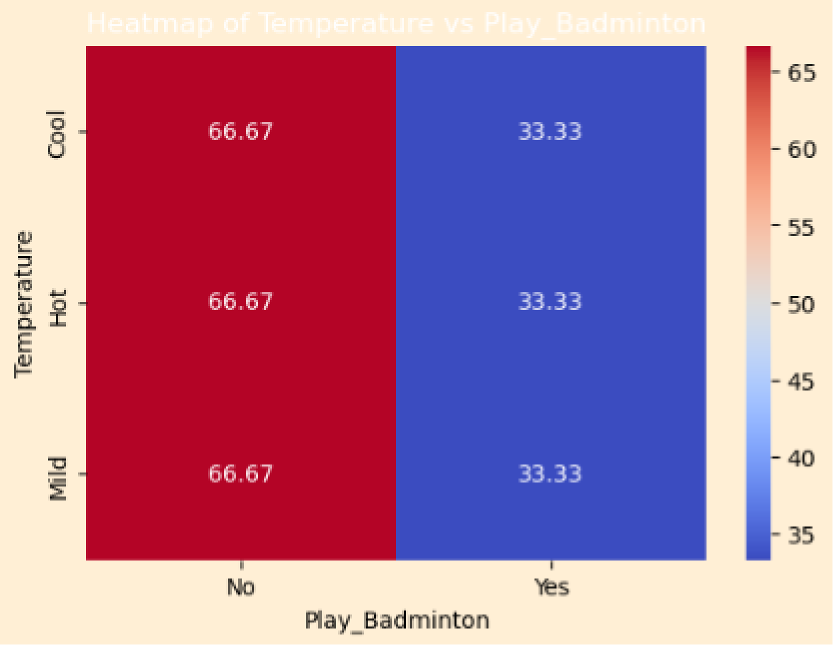


The insights from the plot are:

The column Outlook seems to have a relationship with the badminton playing, the plot shows that:

1. When it's `overcast` then there are `50%` chance of playing badminton and a `50%` chance of not playing badminton.
2. When it's `Raining` then there is a '0%` chance of playing badminton and a `100%` chance of not playing badminton.
3. When it's `Sunny` then there is a `50%` chance of `both` playing and not playing badminton.

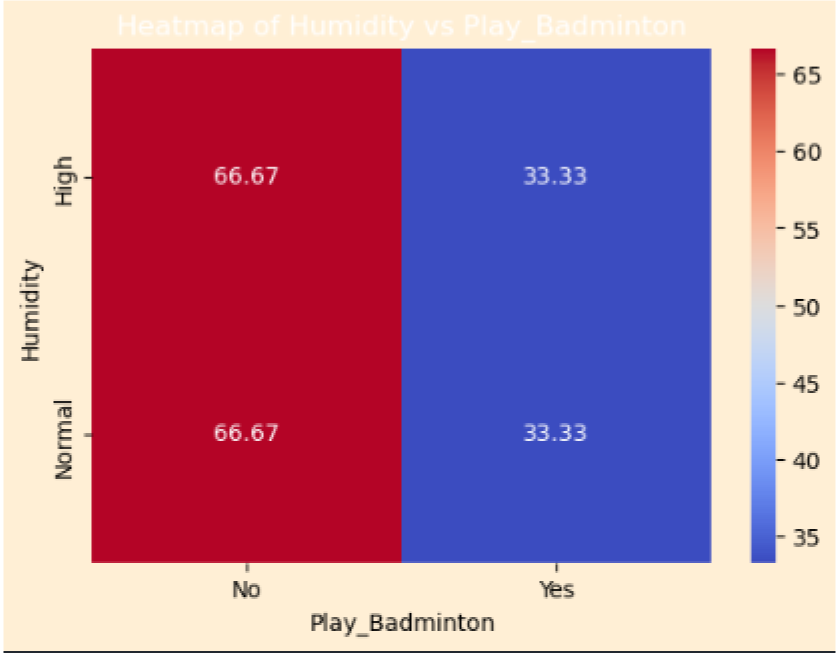
Fahad Ur Rehman ~ 18228



The insights from the plot are:

1. The column `Temperature` seems to have a relationship with the badminton-playing column, the plot shows that:
2. When it's `Cool` then there are `66.67%` chances of not playing badminton and `33.33%` chances of playing badminton.
3. When it's `Hot` then there are `66.67%` chances of not playing badminton and `33.33%` chances of playing badminton.
4. When it's `Mild` then there is a `66.67%` chance of not playing badminton and a`33.33%` chance of playing badminton.

Fahad Ur Rehman ~ 18228



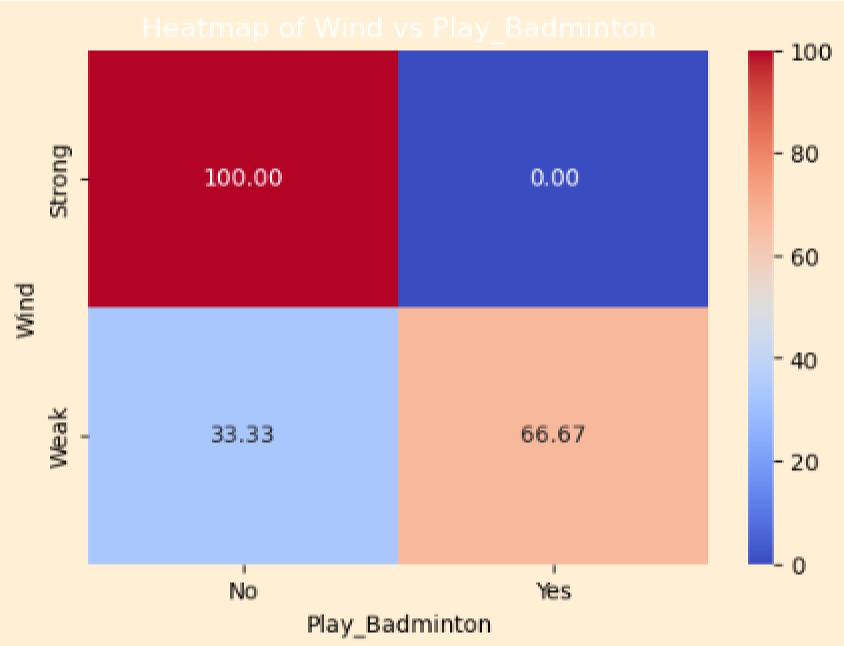
The insights from the plot are:

The column `Humidity` seems to have a relationship with the badminton-playing column, the plot shows that:

1. When it's `HIGH` then there are `66.67%` chances of not playing badminton and `33.33%` chances of playing badminton.

2. When it's `NORMAL` then there are `66.67%` chances of not playing badminton and `33.33%` chances of playing badminton.

Fahad Ur Rehman ~ 18228



The insights from the plot are:

The column `WIND` seems to have a relationship with the badminton-playing column, the plot shows that:

1. When it's `Strong` then there is a `100%` chance of not playing badminton and a `0%` chance of playing badminton.

2. When It's `Weak` then there are `33.33%` chances of not playing badminton and `66.67%` chances of playing badminton.

Fahad Ur Rehman ~ 18228

7. Describe the type of Naive Bayes model you chose (e.g., Gaussian, Multinomial, Bernoulli) and why it is appropriate for your dataset.

   Founded Answer: The choice of which Naive Bayes algorithm is best for a dataset depends on several factors, including the nature of the dataset and the assumptions we're willing to make about the data.

   Since our dataset contains categorical features related to weather conditions and a binary target variable ('Play_Badminton'), we may consider using either Multinomial Naive Bayes or Bernoulli Naive Bayes.

   However, since our features represent categories rather than word counts or frequencies (as in text classification), `Bernoulli Naive Bayes` may be more suitable for this dataset. It treats each feature as a binary variable indicating the presence or absence of a category (e.g., 'Sunny', 'Overcast', 'Rain' for 'Outlook'). Therefore, Bernoulli Naive Bayes may be the best choice for your dataset.

8. What are your Interpretations of the metrics you have used for model evaluation?

   ### Interpretations of the Metrics:

   The evaluation metrics for the model are exceptional, with all metrics (accuracy, precision, recall, and F1 score) achieving perfect scores of 1.0. This indicates that the model has successfully classified all instances in the test set without any errors.

   The confusion matrix further confirms this outstanding performance, showing that there are no misclassifications. Specifically, there are 5 instances correctly classified as negative (True Negatives) and 3 instances correctly classified as positive (True Positives), with no false positives or false negatives.

In summary, the model demonstrates flawless performance on the test data, achieving perfect accuracy, precision, recall, and F1 score. This suggests that the model has learned the underlying patterns in the data extremely well and can make highly accurate predictions across both positive and negative classes.

9. Use scaling and normalization techniques if necessary and observe their impact on the model's performance.

   Founded Answer:

   Normalization and Scaling weren't implemented due to two major reasons which are as follows:

   1. The performance accuracy of the model was 100% according to the metrics, so there was no need for normalization and scaling to enhance the prediction or the performance accuracy.

   2. BernoulliNB typically works with binary data, so scaling might not be necessary.