Classification Analysis

(Anticipating Consumer Satisfaction)

INDEX

Abstract

This report delves into the application of machine learning techniques, specifically Decision Trees and Naive Bayes algorithms, to predict customer satisfaction based on various features. The study rigorously evaluates the performance of these models using training and test datasets. Key metrics such as accuracy, recall, and F1-score are used to assess the effectiveness of the models. The objective of this analysis is to present a clear and professional overview of the findings, emphasizing the significance and efficiency of the implemented models in predicting customer satisfaction. This report aims to provide valuable insights into the application of these machine learning techniques in a practical setting, offering a comprehensive view of their capabilities and limitations in the context of customer satisfaction prediction.

data

In the realm of data analysis, particularly in machine learning, the quality and structure of the dataset play a crucial role in the performance of predictive models. Two common issues that often arise in datasets are **'noise'** and **'redundant features'**. **Noise refers to erroneous data that can obscure the underlying patterns the model is trying to learn. This can lead to lower accuracy and generalization capabilities of the model. On the other hand, redundant features are those that provide no additional information.**

For this project, two distinct kind of datasets were created to evaluate the impact of **noise** and **Redundant features** on model performance. The first dataset includes **noise and Redundant features,** this dataset will be used for the last part of this project where is required to **(Slightly alter the absolutely right rules and generate another set of data)** since the noise created is for altering the balance in **the absolutely right rules**, and the second dataset has only **Redundant features,** and according to the definition, these features should not significantly affect the model performance, since these features are not related to the class we are going to predict.

Settings

Inputs

- **Items: Consist in a list of 50 different items** ,in order to not affect the algorithms, I converted this category feature in a labeled encoded feature.
- **Price: Int Value given to the items.** there is a 30% or less of probability that the prices will be raised by 30% of the original price, and a 25% of probability of 30% to have lower price than the original price.
- **Time Waiting: binary feature where short=0, long=1,** This is a binary feature describes the time that the customer waits in the queue, there is a 50% or less of probability that the customer will wait for a long time then the value will be 1 otherwise will be 0.
- **Customer service: binary feature where bad=0, good=1,**This feature explain how well the customer was served while he is purchasing items,**(in the dataset with noise this feature is affected by cashier mood, which means that if the cashier is in a bad mood, it will give a bad service to the customer thus the satisfaction of the client will be affected indirectly)** ,in the dataset without noise , there is a 50% or less that the customer get a bad service.
- **Mood: binary feature where bad=0, good=1,**This feature explains the mood of the customer when he/her purchase an item **,(in the dataset with noise this feature is affected by the weather which means that when there is a bad weather , the mood of the customer will be affected ,thus the satisfaction will be affected indirectly),** in the dataset without noise there is a 30% or less of probability that the costumer has a bad mood.
- **Good Environment: binary feature where bad=0, good=1,**this feature explains how the environment in the store may affect the satisfaction of the customer, for this the probability was adjusted to 40% or less that one store has a bad environment.

Output

- **Satisfaction: Binary class where Satisfied=1,Unsatisfied=0,** this is the feature we want to predict.

Noise

- **Weather: binary feature where bad=0, good=1,** this feature only appears in the data with noise, and this explains how a bad weather influences in a customer mood, in other to make this feature really noise that can affect the Satisfaction feature indirectly, I adjust its probability up to 90% .
- **Cashier mood: binary feature where bad=0, good=1,** this explains how the mood of the cashier could affect how he/her treats customer, the probability that the cashier has a bad mood was adjusted up to 90%, in order to introduce noise in the dataset.

Redundant features

- **Payment Method: costumer has 70% or less of probability to pay using credit card, this a binary feature where credit card=1, cash=0** since this is not related to the satisfaction of the costumer , this is a redundant feature.
- **Cashier Gender: Cashier has 50% or less to be male, binary feature where Female=0, Male=1,** since this is not related to the satisfaction of the costumer , this is a redundant feature.
- **Hungry? : Costume has 20% or less to be hungry, this is binary feature where No hungry=0, Hungry=1,** since this is not related to the satisfaction of the costumer , this is a redundant feature.

Absolutely Right Rules

- If the item price is higher than the original price, the possibility of unsatisfaction will increase, in the other hand if the price is less or equal to the original price , the chance of satisfaction increase.
- If the customer does not need to wait for a long time in the queue, this may increase the chance of satisfaction if is related with other

positive Absolutely Rules like the first one , otherwise the satisfaction may be (0)Unsatisfied.

- If the costumer service is good and also this rule is combined with other positive Absolutely Rules, the satisfaction will be positive (1), otherwise will be (0)Unsatisfied.
- If the costumer mood is good , and this rule is combined with other positive Absolutely Rules, the satisfaction will be positive (1), otherwise will be (0)Unsatisfied.
- If the environment of the store is good, and this rule is combined with other positive Absolutely Rules the satisfaction will be positive (1), otherwise will be (0)Unsatisfied.

Data Appearance

- This is the dataset with 2 noise features and 3 redundant features. **12 features in total and 994 items.**

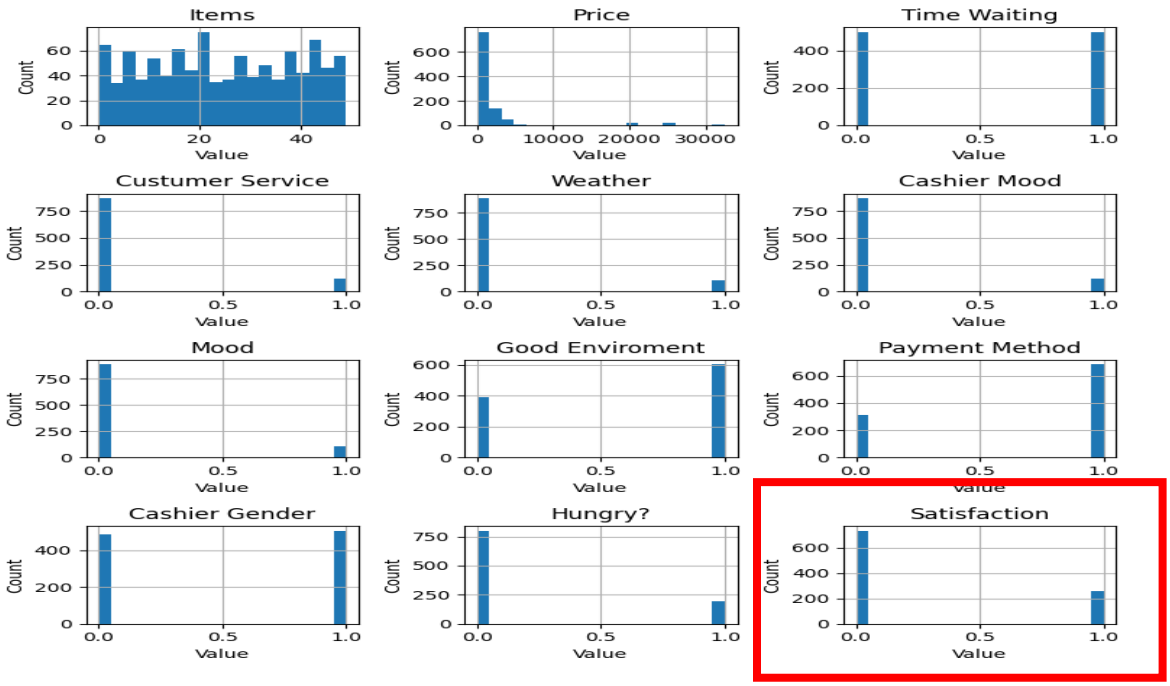
Items	Price	Time Waiting	Custumer Service	Weather	Cashier Mood	Mood	Good Enviroment	Payment Method	Cashier Gender	Hungry?	Satisfaction
18.0	150.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
41.0	130.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0
18.0	195.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
2.0	120.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0
21.0	130.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0
39.0	2340.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
16.0	90.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0
38.0	20000.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
29.0	230.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
14.0	1950.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0
7.0	910.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0
8.0	2000.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
43.0	3400.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
49.0	150.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0
22.0	150.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0

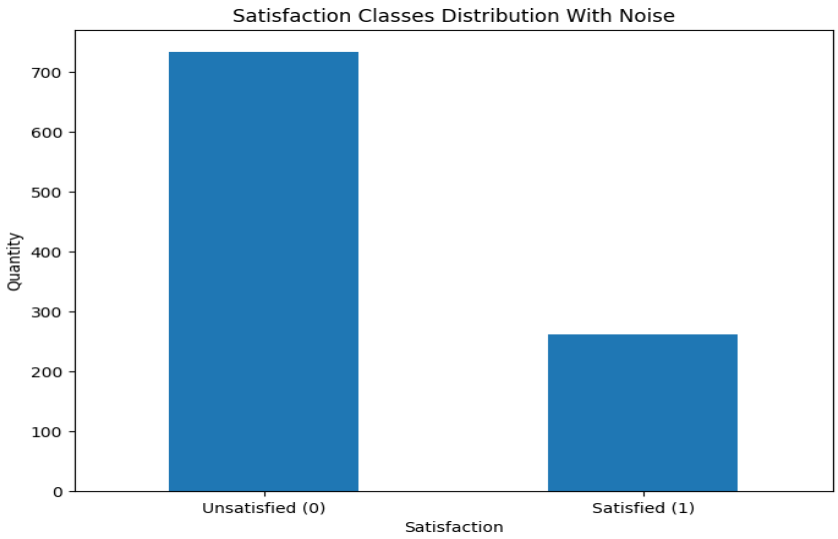
. This is the dataset without noise but with 3 redundant features. **10 features in total and 994 items.**

Items	Price	Time Waiting	Custumer Service	Mood	Good Enviroment	Payment Method	Cashier Gender	Hungry?	Satisfaction
6.0	2730.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0
7.0	700.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0
49.0	195.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
22.0	195.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0
42.0	130.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0
0.0	30.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
34.0	221.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
33.0	260.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0
15.0	150.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0
8.0	2000.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0
12.0	208.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0
36.0	110.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0
24.0	25000.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0
4.0	200.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0
1.0	140.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0

Dataset with Noise

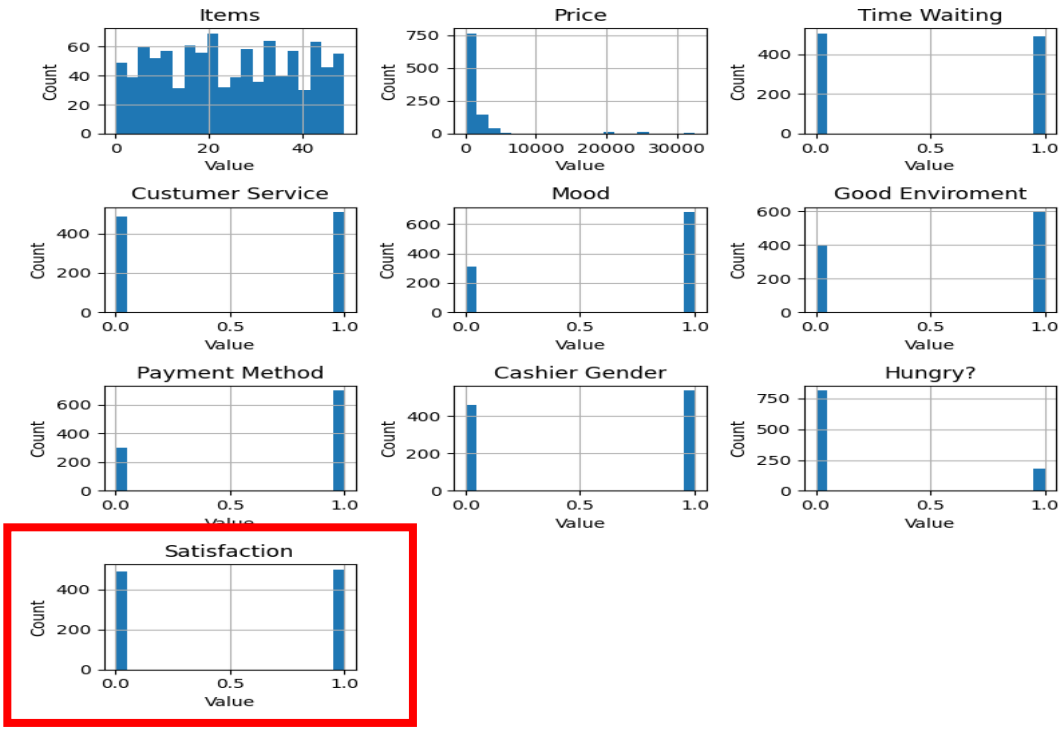
Incorporating the **'Weather'** feature, which is **linked to mood**, and the **'Cashier Mood'** feature, **connected to customer service**, **introduces noise** into the dataset as they indirectly affect **'Consumer Satisfaction'**. Poor weather conditions negatively impact the customer's mood, and a negative cashier mood adversely influences the quality of service provided. Consequently, **these factors lead to a significant rise in 'unsatisfied' cases within the dataset.** This results in a **skewed dataset**, dominated by negative classes in the key feature we aim to predict: Consumer Satisfaction. Such an addition underscores the complexity of indirect influences in predictive modeling and the importance of careful feature selection.

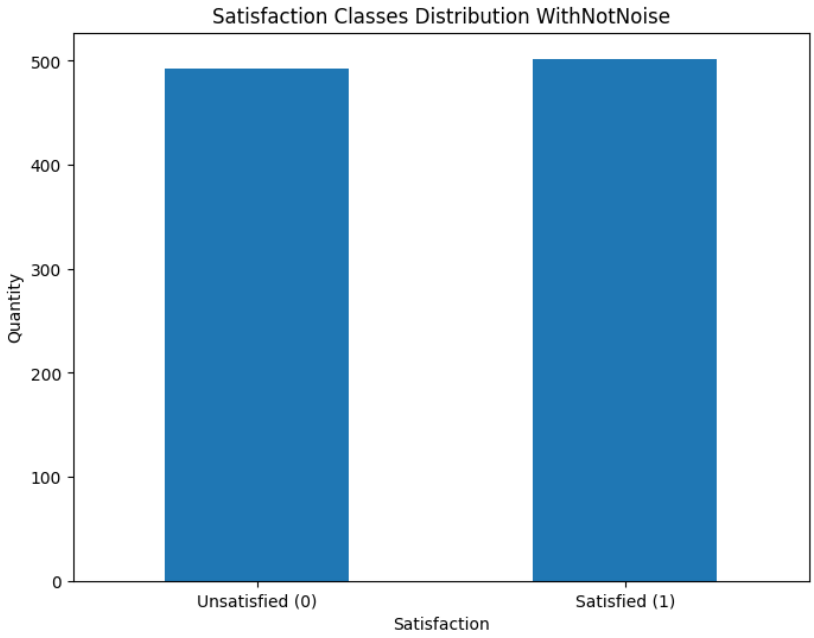




Dataset Without Noise

When the **'weather' and "cashier mood"** features, which influences the costumer **"Mood" and the "costumer service", subsequently the 'Satisfaction' feature**, are removed from the dataset, there is a notable shift in the data balance. Without this external impact, the positive class within the 'Satisfaction' feature increases, **leading to a more evenly distributed dataset between positive and negative classes**. This highlights the significance of feature selection in dataset composition, demonstrating how removing a disruptive **Noise** features can enhance the balance and potentially the predictive accuracy of the models applied to customer satisfaction analysis.


Items
Price
Time Waiting
Custumer Service
Mood
Good Enviroment
Payment Method
Cashier Gender
Hungry?
Satisfaction
Count
Value

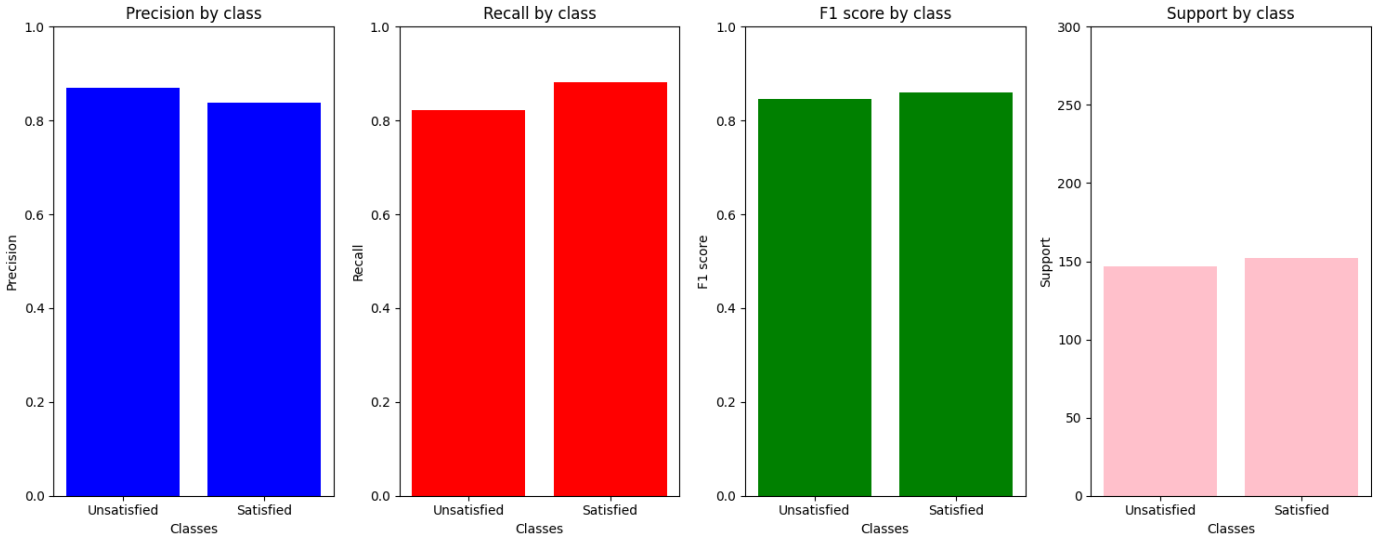

Satisfaction Classes Distribution WithNotNoise
Quantity
Unsatisfied (0)
Satisfied (1)
Satisfaction

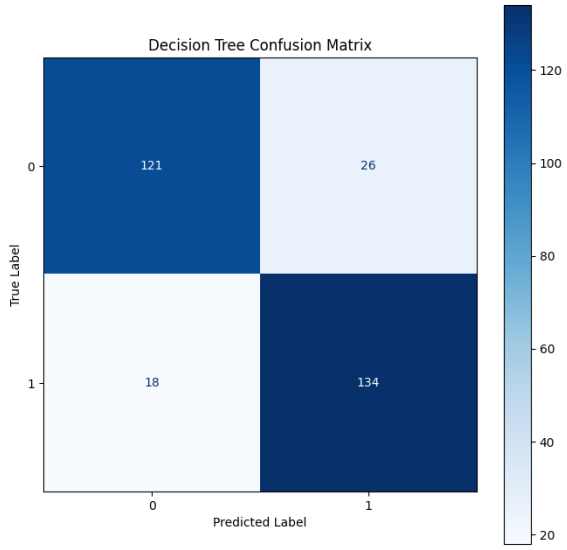
Model and Evaluation

Decision Tree:

. Dataset without Noise but with the Redundant features.

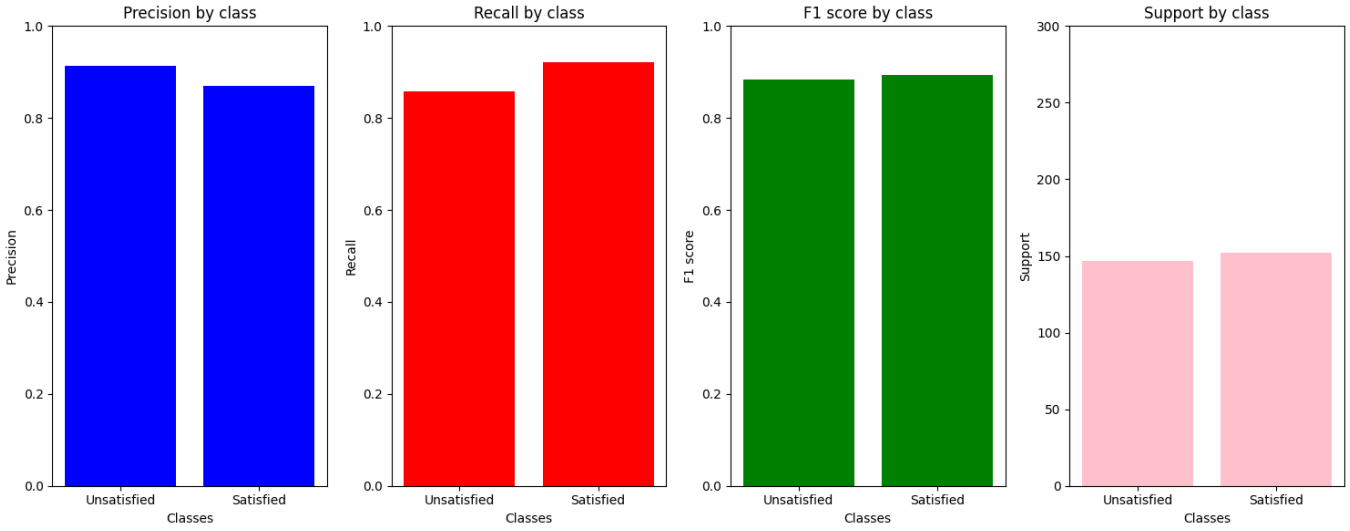
Accuracy: 0.862876254180602

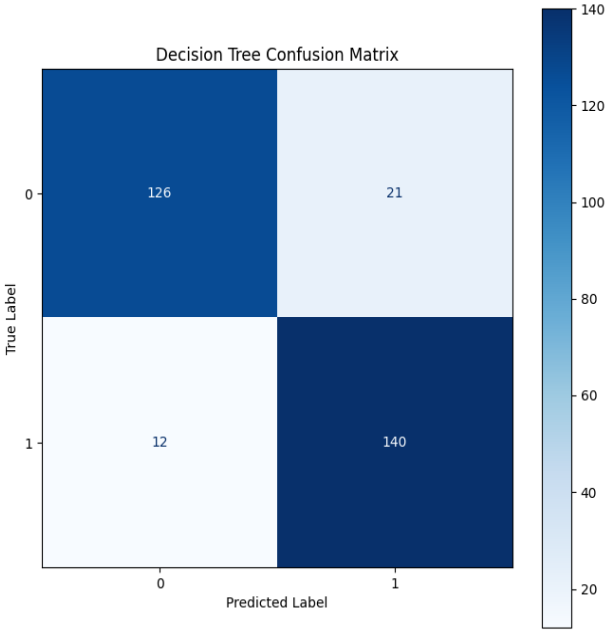




. Dataset without Noise and without the Redundant features.

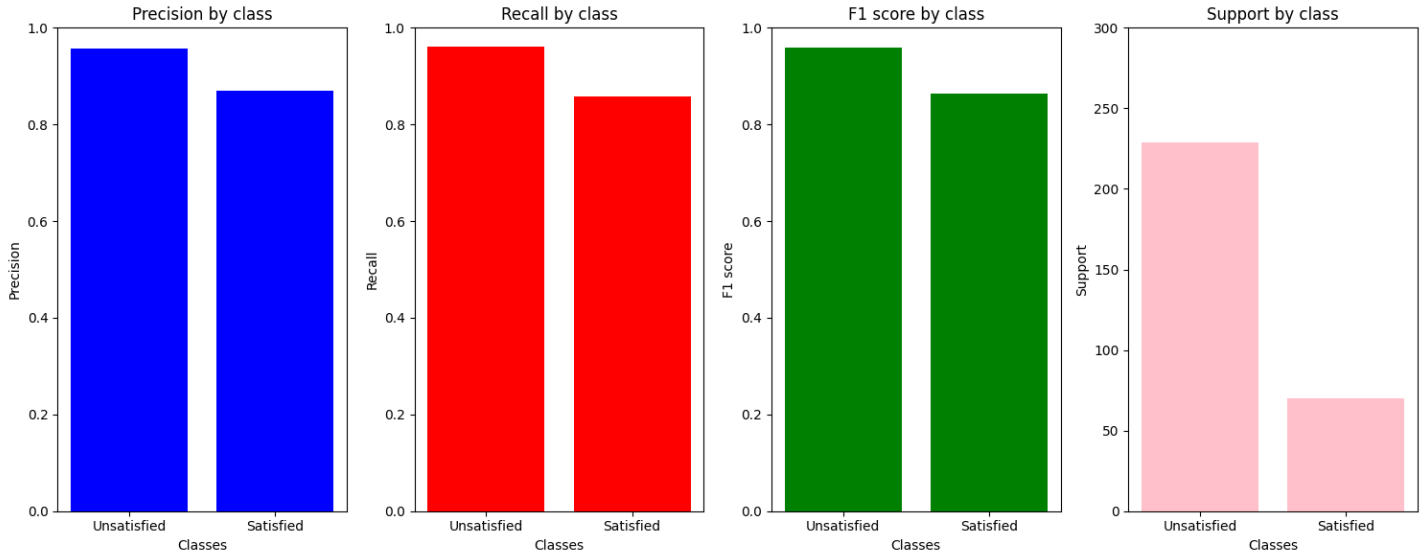
Accuracy: 0.8963210702341137

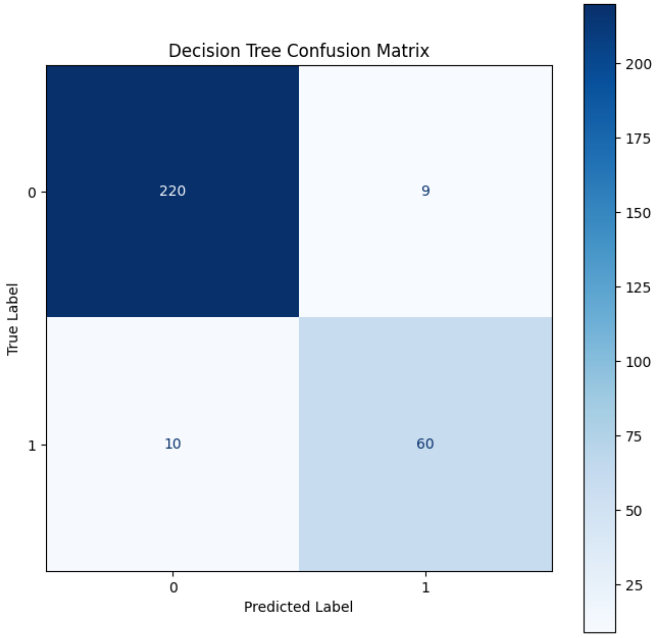


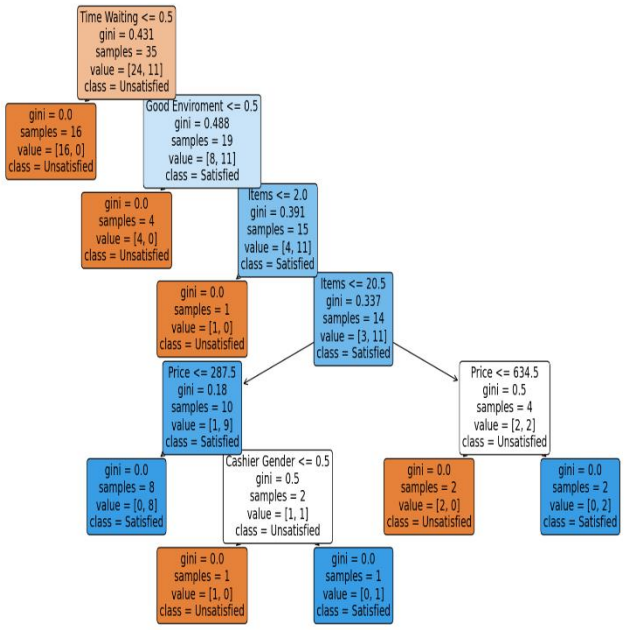


. Dataset with Noise and with Redundant features.

Accuracy: 0.9364548494983278



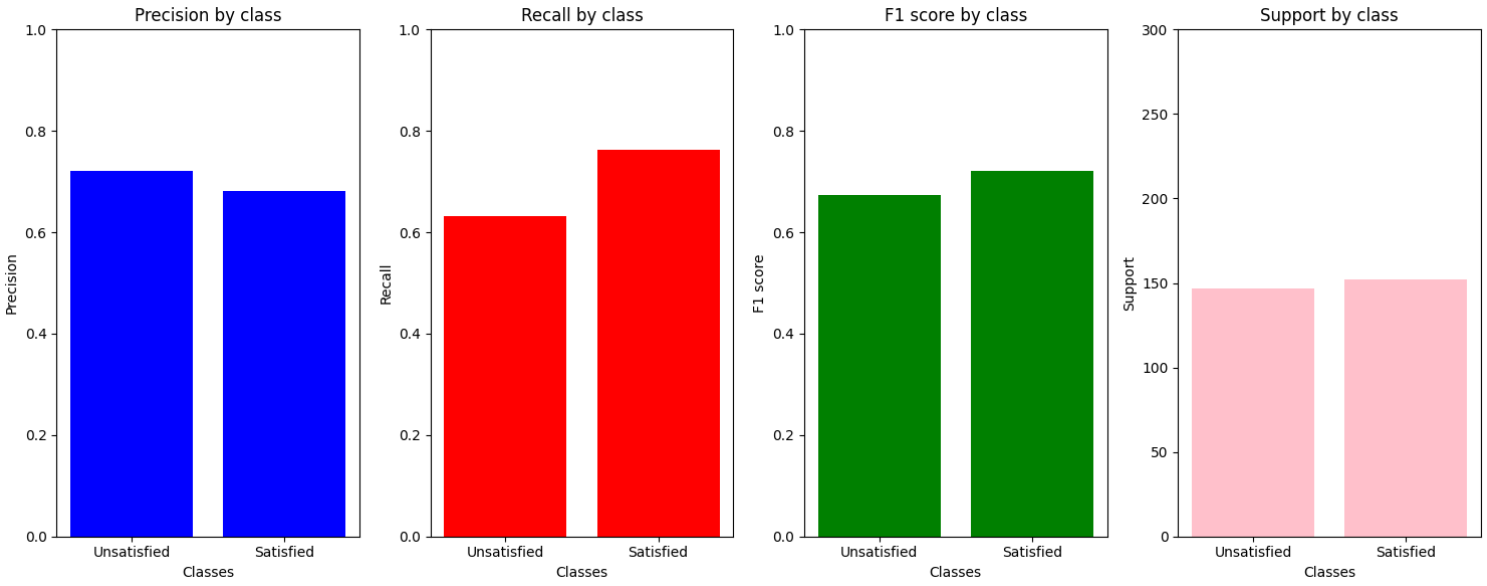


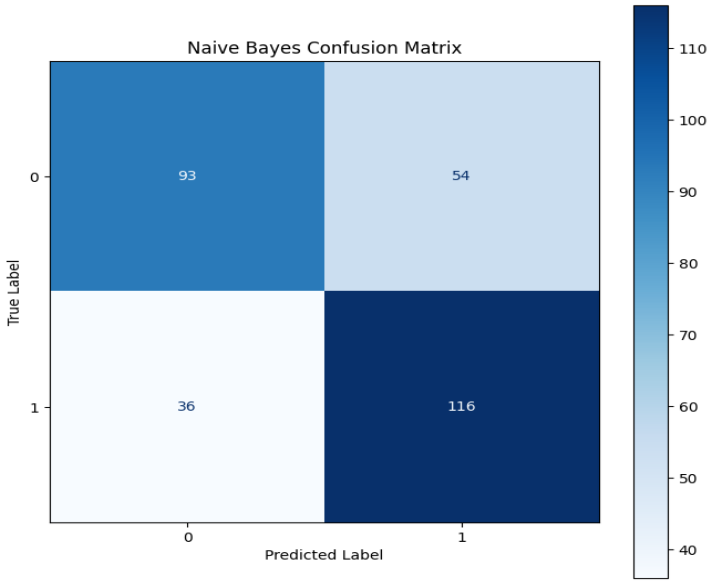


Gaussian Naive Bayes:

. Dataset without Noise but with the Redundant features.

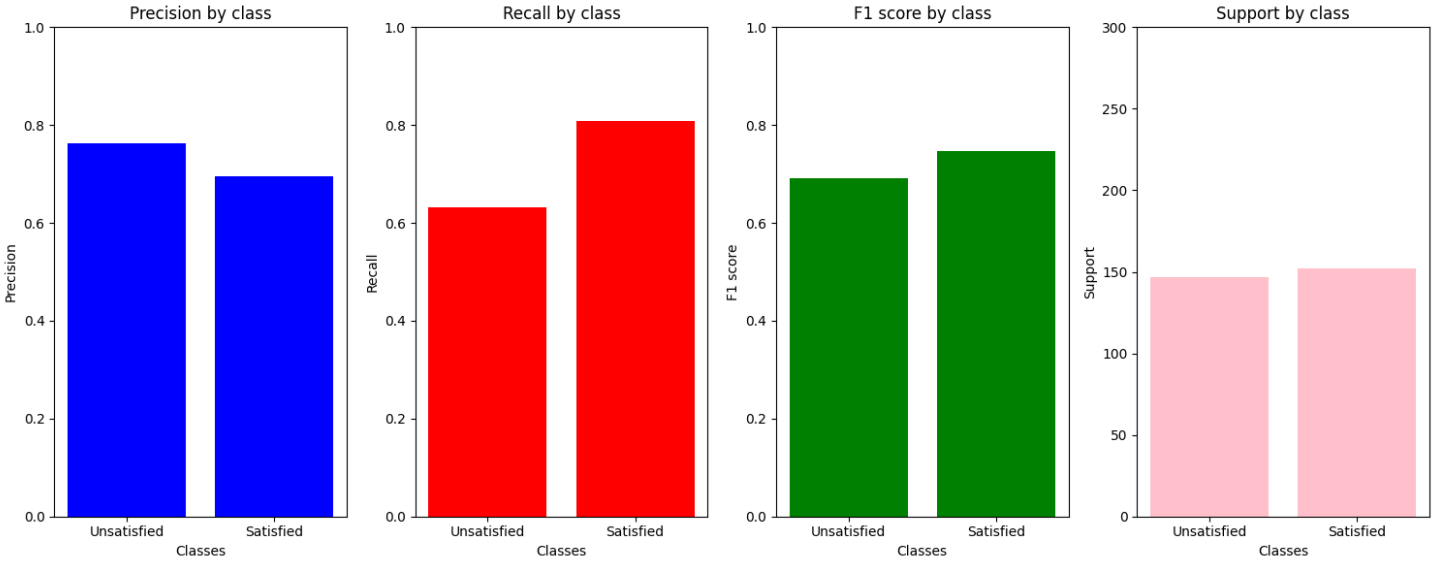
Accuracy: 0.6989966555183946

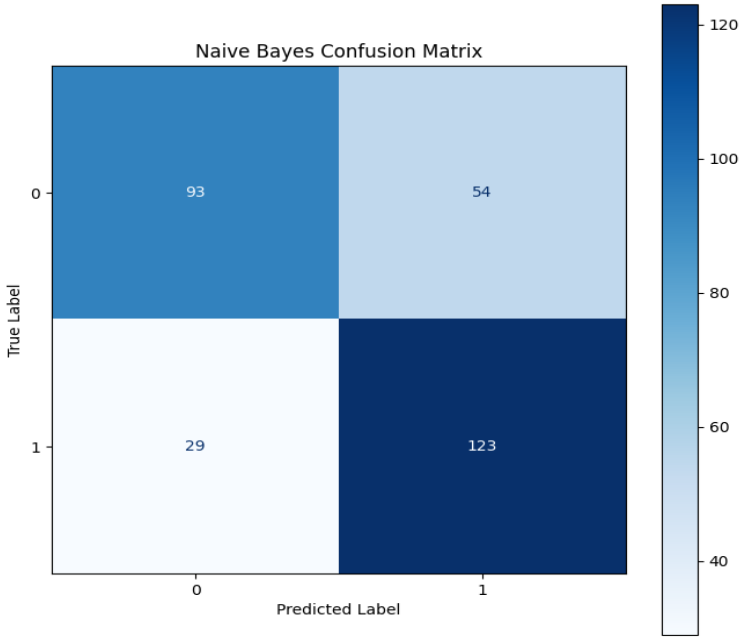




. Dataset without Noise and without the Redundant features.

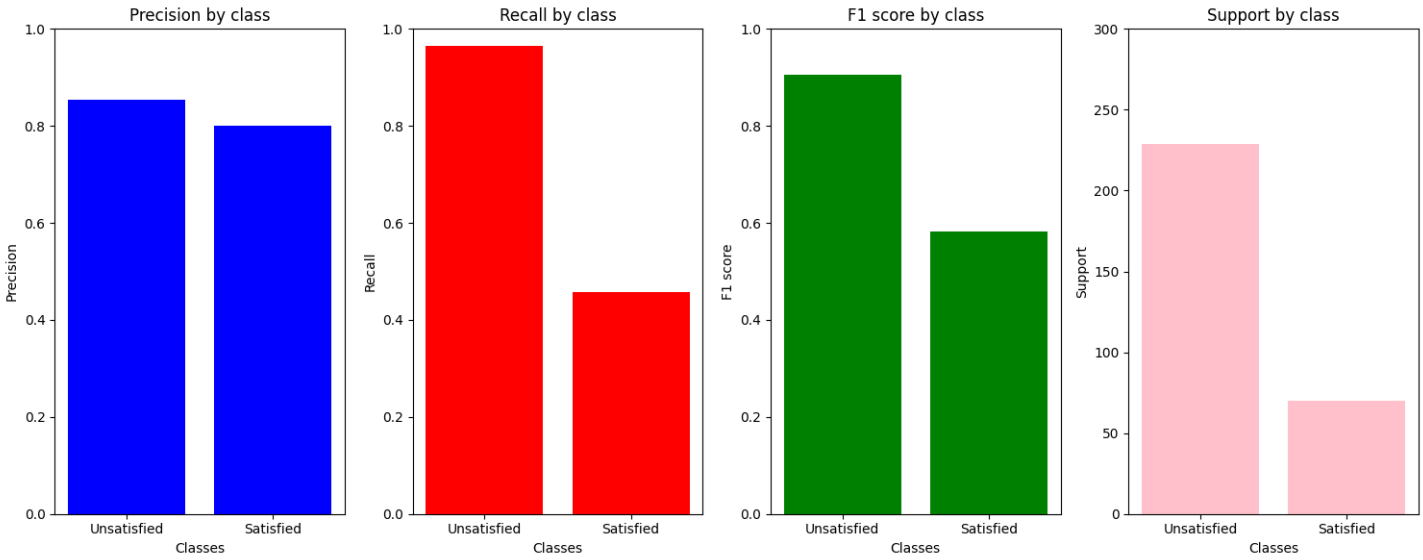
Accuracy: 0.7224080267558528

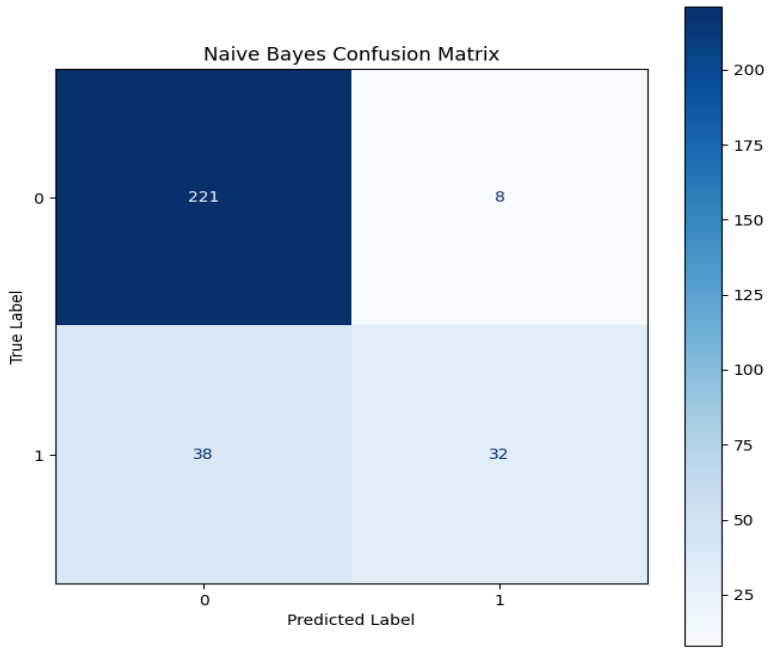




. Dataset with Noise and with Redundant features.

Accuracy: 0.8461538461538461





Conclusion

After conducting various tests and observing the results of both models under different conditions, I have arrived at the following conclusions:

. **The Decision Tree** was subjected to testing under **three different scenarios.** In the initial scenario, it was trained and tested on a dataset that, while free from noise, included redundant features. The model achieved an acceptable precision of 86.28%. According to **the confusion matrix, predictions were stable in spite of these redundant features.** In the second scenario, the model was provided with a clean dataset, absent both noise and redundant features, which led to an increased precision of 89.63%. **Based on the accuracy, recall, and F1 score metrics, the Decision Tree demonstrated consistent outcomes**, a stability that was also echoed in the confusion matrix with very consistent predictions. In the final test, the Decision Tree was challenged with a noisy and imbalanced dataset, where the negative class was dominant. Remarkably, not only did the Decision Tree enhance its precision to 93.64%, but it also retained its stability. Although the accuracy, recall, and F1 score metrics showed a slight inclination towards the negative class, the confusion matrix displayed stable predictions from the Decision Tree. **This suggests that the Decision Tree is an effective algorithm capable of performing well in scenarios marked by data noise and imbalance, potentially due to its hierarchical method of making predictions.**

. **The Gaussian Naive Bayes algorithm** underwent the same tests as the Decision Tree, yet its performance varied significantly. In the first test, with a dataset free from noise but laden with redundant features, it achieved a modest precision of 69.89%. Furthermore, **the confusion matrix indicated irregular behavior,** with an **exceptionally high number of false negatives**, casting doubt on its efficacy. In the second test, **involving a clean dataset without noise or redundant features**, precision saw a marginal improvement to 72.24%. Despite this cleaner and more balanced dataset, the algorithm still underperformed, as evidenced by a **considerable number of false negatives in the confusion** matrix. Finally, when trained with a **noisy and feature-redundant**

dataset, precision unexpectedly rose to 84.61%. However, a closer examination of other metrics like recall and F1 score revealed biased predictions. The recall and F1 score disproportionately increased for the dominant negative class, indicating an imbalance in the model's performance. The confusion matrix also suggested an inability to generalize, due to a significant number of false positives, skewing predictions towards the less observed positive class in the training. In conclusion, these tests have led me to understand that in situations with data noise and imbalance, the Decision Tree significantly outperforms the Naive Bayes algorithm.