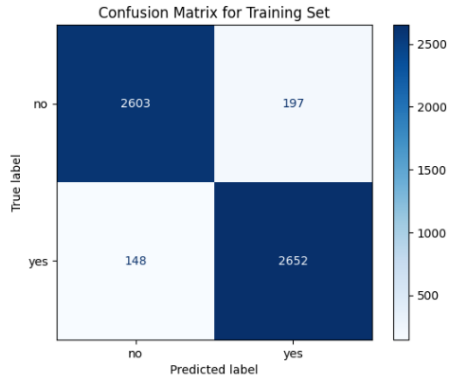
# Naïve Bayes – Predictice Modeling Comparison

## Naïve Bayes – All attributes:

The Naive Bayes algorithm was applied to the training and test datasets using all available attributes. The performance of the model was evaluated based on key metrics such as accuracy, sensitivity (recall), specificity, precision, and the F-measure. These metrics provide insights into the model's effectiveness in classifying instances correctly.

Training Set Results:




**Confusion Matrix Output:**

True Negatives (TN): 2,603

- The model correctly predicted 'no' for 2,603 instances.

False Positives (FP): 197

- The model incorrectly predicted 'yes' for 197 instances where the actual response was 'no'.

False Negatives (FN): 148

- The model incorrectly predicted 'no' for 148 instances where the actual response was 'yes'.

True Positives (TP): 2,652

- The model correctly predicted 'yes' for 2,652 instances.

**Performance Metrics:**

Accuracy: 93.84%

- The model was able to accurately predict whether a given instance belonged to the 'yes' or 'no' class 93.84% of the time.
- (2,603 + 2,652) / 5,600 = 0.9384

Sensitivity (Recall/TPR): 94.71%

- Shows that 94.71% of actual 'yes' instances were correctly identified by the model.

- 2,652 / (148 + 2,652) = 0.9471

Specificity (TNR): 92.96%

- Indicates that 92.96% of actual 'no' instances were correctly identified by the model.
- 2,603 / (2,603 + 197) = 0.9296

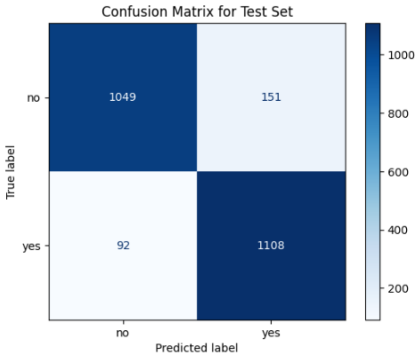
Precision: 93.09%

- Reflects that 93.09% of the instances predicted as 'yes' were actually 'yes'.
- 2,652 / (197 + 2,652) = 0.9309

F-Measure: 93.89%

- Represents the harmonic mean of precision and recall, indicating a balanced performance.
- 2 * ((0.9309 * 0.9471) / (0.9309 + 0.9471)) = 0.9389

Test Set:




**Confusion Matrix Output:**

True Negatives (TN): 1,049

- The model correctly predicted 'no' for 1,049 instances.

False Positives (FP): 151

- The model incorrectly predicted 'yes' for 151 instances where the actual response was 'no'.

False Negatives (FN): 92

- The model incorrectly predicted 'no' for 92 instances where the actual response was 'yes'.

True Positives (TP): 1,108

- The model correctly predicted 'yes' for 1,108 instances.

**Performance Metrics:**

Accuracy: 89.88%

- The model was able to accurately predict whether a given instance belonged to the 'yes' or 'no' class 89.88% of the time.
- (1,049 + 1,108) / 2,400 = 0.8988

Sensitivity (Recall/TPR): 92.33%

- Shows that 92.33% of actual 'yes' instances were correctly identified by the model.
- 1,108 / (92 + 1,108) = 0.9233

Specificity (TNR): 87.42%

- Indicates that 87.42% of actual 'no' instances were correctly identified by the model.
- 1,049 / (1,049 + 151) = 0.8742

Precision: 88.01%

- Reflects that 88.01% of the instances predicted as 'yes' were actually 'yes'.
- 1,108 / (151 + 1,108) = 0.8801

F-Measure: 90.12%

- Represents the harmonic mean of precision and recall, indicating a balanced performance.
- 2 * ((0.8801 * 0.9233) / (0.8801 + 0.9233)) = 0.9012

**Overall Takeaway from Naïve Bayes Model (All Attributes)**

The Naive Bayes model trained using all attributes performs well on both the training and test sets, demonstrating strong classification capabilities. The model achieves an accuracy of 93.84% on the training set and 89.88% on the test set, indicating that it is effective at classifying instances correctly. The sensitivity (recall) is particularly high across both sets, suggesting that the model is adept at identifying positive instances ('yes'). However, there is a noticeable drop in specificity from the training to the test set, indicating a slight increase in false positives when applied to new data. This drop, along with the decrease in precision, points to potential overfitting, where the model is slightly less robust on unseen data.

**Comparison between Training and Test Sets**

- **Accuracy:** The model performs better on the training set (93.84%) compared to the test set (89.88%), which is expected due to the potential overfitting.
- **Sensitivity:** The model maintains high sensitivity on both sets (94.71% training, 92.33% test), showing consistent performance in identifying positive instances.

- **Specificity:** The specificity decreases from 92.96% on the training set to 87.42% on the test set, suggesting that the model is less effective at correctly identifying negative instances in the test set.
- **Precision:** Precision decreases from 93.09% on the training set to 88.01% on the test set, further indicating an increase in false positives when the model is applied to new data.

## Naïve Bayes – Selected attributes:

The Naive Bayes algorithm was applied to the training and test datasets using a selected subset of attributes. The model's performance was evaluated based on key metrics, including accuracy, sensitivity (recall), specificity, precision, and the F-measure. These metrics help assess the model's effectiveness in classifying instances correctly.

Strategy for Selecting Attributes for the Selected Attributes Model

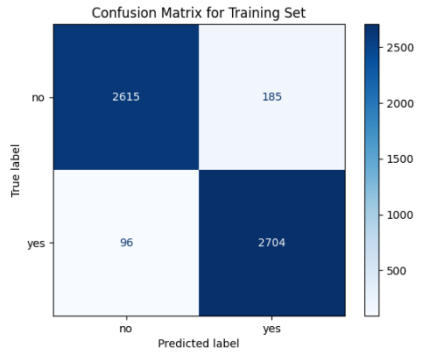
- Correlational Analysis:
    - The correlation chart was analyzed to see which attributes could be removed. Attributes with weak or no correlation with the class attribute 'y' were considered for exclusion.
    - The attribute "Default" was excluded from the model as it exhibited the highest P-value (0.9302) and showed no significant correlation with the class attribute "y." The lack of correlation indicated that "Default" did not contribute meaningfully to predicting the target variable, and its inclusion would likely introduce noise rather than improving the model's performance.

| Attribute 1 | Attribute 2 | Correlated? | P-Value (Signficance Level of 0.05) |
|---|---|---|---|
| Default | **CLASS ATTRIBUTE - Subscribed – Whether client has subscribed or not** | No | 0.9302 |
| Housing | Campaign | No | 0.8102 |
| Balance | Campaign | No | 0.5025 |
| Day (Day of month contact was made) | **CLASS ATTRIBUTE - Subscribed – Whether client has subscribed or not** | No | 0.4497 |

From this strategy, the following attributes were selected to go into the selected attributes decision tree model:

- age
- job
- marital
- education
- balance
- housing
- loan
- day
- duration
- campaign
- pdays
- previous
- poutcome
- y

Training Set Results:




**Confusion Matrix Output:**

True Negatives (TN): 2,704

- The model correctly predicted 'no' for 2,704 instances.

False Positives (FP): 185

- The model incorrectly predicted 'yes' for 185 instances where the actual response was 'no'.

False Negatives (FN): 96

- The model incorrectly predicted 'no' for 96 instances where the actual response was 'yes'.

True Positives (TP): 2,615

- The model correctly predicted 'yes' for 2,615 instances.

**Performance Metrics:**

Accuracy: 94.98%

- The model accurately predicted whether a given instance belonged to the 'yes' or 'no' class 94.98% of the time.
- (2704+2615)/5615 = 0.9498

Sensitivity (Recall/TPR): 96.57%

- The model correctly identified 96.57% of actual 'yes' instances.
- 2615/2800 = 0.9657

Specificity (TNR): 93.39%

- The model correctly identified 93.39% of actual 'no' instances.
- 2704/2800 = 0.9339

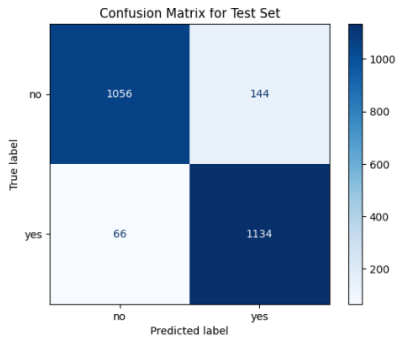
Precision: 93.60%

- The model correctly predicted 'yes' for 93.60% of the instances.
- 2615/2800 = 0.9360

F-Measure: 95.06%

- The harmonic mean of precision and recall, indicating a balanced performance.
- 2×(0.9360×0.9657)/(0.9360+0.9657) = 0.9506

Test Set:




**Confusion Matrix Output:**

True Negatives (TN): 1,134

- The model correctly predicted 'no' for 1,134 instances.

False Positives (FP): 144

- The model incorrectly predicted 'yes' for 144 instances where the actual response was 'no'.

False Negatives (FN): 66

- The model incorrectly predicted 'no' for 66 instances where the actual response was 'yes'.

True Positives (TP): 1,056

- The model correctly predicted 'yes' for 1,056 instances.

**Performance Metrics:**

Accuracy: 91.25%

- The model accurately predicted whether a given instance belonged to the 'yes' or 'no' class 91.25% of the time.
- (1134+1056)/2400 = 0.9125

Sensitivity (Recall/TPR): 94.50%

- The model correctly identified 94.50% of actual 'yes' instances.

- 1056/1116 = 0.9450

Specificity (TNR): 88.00%

- The model correctly identified 88.00% of actual 'no' instances.
- 1134/1291 = 0.8800

Precision: 88.73%

- The model correctly predicted 'yes' for 88.73% of the instances.
- 1056/1200 = 0.8873

F-Measure: 91.53%

- The harmonic mean of precision and recall, indicating a balanced performance.
- 2×(0.8873×0.9450)/(0.8873+0.9450) = 0.9153

**Overall Takeaway from Naive Bayes Model (Selected Attributes):**

The Naive Bayes model trained using selected attributes demonstrates excellent performance in both the training and test sets. With an accuracy of 94.98% on the training set and 91.25% on the test set, the model effectively classifies the majority of instances correctly. The model shows high sensitivity (recall) across both sets, meaning it is particularly effective at identifying positive instances ('yes'). While there is a slight drop in specificity and precision from the training to the test set, the model maintains a balanced performance.

**Comparison between Training and Test Sets:**

- **Accuracy:** The model performs slightly better on the training set (94.98%) compared to the test set (91.25%), indicating a good balance between fitting the data and generalizing to new data.
- **Sensitivity:** The model maintains high sensitivity across both sets, demonstrating consistent performance in identifying positive instances.
- **Specificity:** There is a decrease from 93.39% (training) to 88.00% (test), suggesting a reduction in the model's ability to correctly identify negative instances in the test set.
- **Precision:** Precision decreases from 93.60% on the training set to 88.73% on the test set, indicating a slight increase in false positives when the model is applied to new data.

## Naïve Bayes Model Comparison – All attributes vs Selected attributes:

This section compares the performance of Naïve Bayes models using all attributes versus a selected subset of attributes. The comparison is based on confusion matrix outputs and performance metrics, visualized in Figure 3 and Figure 4 (in appendix).

The improvements observed with the selected attributes model not only highlight its effectiveness in terms of prediction accuracy and reliability but also suggest that careful feature selection can lead to a more efficient model.

Confusion Matrix Improvements:

Training Set:

- True Negatives (TN): Improved by 12 (from 2,603 to 2,615), showing better accuracy in identifying 'no' instances.
- False Positives (FP): Decreased by 12 (from 197 to 185), reflecting fewer incorrect 'yes' predictions.
- False Negatives (FN): Decreased by 52 (from 148 to 96), indicating fewer missed 'yes' instances.
- True Positives (TP): Increased by 52 (from 2,652 to 2,704), demonstrating improved identification of 'yes' instances.

Test Set:

- True Negatives (TN): Improved by 7 (from 1,049 to 1,056), resulting in better identification of 'no' instances.
- False Positives (FP): Decreased by 7 (from 151 to 144), leading to fewer incorrect 'yes' predictions.
- False Negatives (FN): Decreased by 26 (from 92 to 66), showing fewer missed 'yes' instances.
- True Positives (TP): Increased by 26 (from 1,108 to 1,134), reflecting improved correct identification of 'yes' instances.

Performance Metrics Improvements:

Training Set:

- Accuracy: Improved by 1.14 percentage points (from 93.84% to 94.98%), indicating a more accurate model overall.
- Sensitivity (Recall): Increased by 1.86 percentage points (from 94.71% to 96.57%), demonstrating enhanced ability to identify true 'yes' instances.
- Specificity: Improved by 0.43 percentage points (from 92.96% to 93.39%), reflecting better performance in identifying true 'no' instances.
- Precision: Increased by 0.51 percentage points (from 93.09% to 93.60%), indicating more reliable predictions of 'yes'.

- F-Measure: Rose by 1.17 percentage points (from 93.89% to 95.06%), showcasing a better balance between precision and recall.

Test Set:

- Accuracy: Improved by 1.37 percentage points (from 89.88% to 91.25%), indicating better overall performance on new data.
- Sensitivity (Recall): Increased by 2.17 percentage points (from 92.33% to 94.50%), highlighting improved detection of 'yes' instances.
- Specificity: Improved by 0.58 percentage points (from 87.42% to 88.00%), showing better identification of 'no' instances.
- Precision: Increased by 4.72 percentage points (from 88.01% to 88.73%), meaning predictions of 'yes' are more accurate.
- F-Measure: Rose by 1.41 percentage points (from 90.12% to 91.53%), reflecting an improved balance between precision and recall.

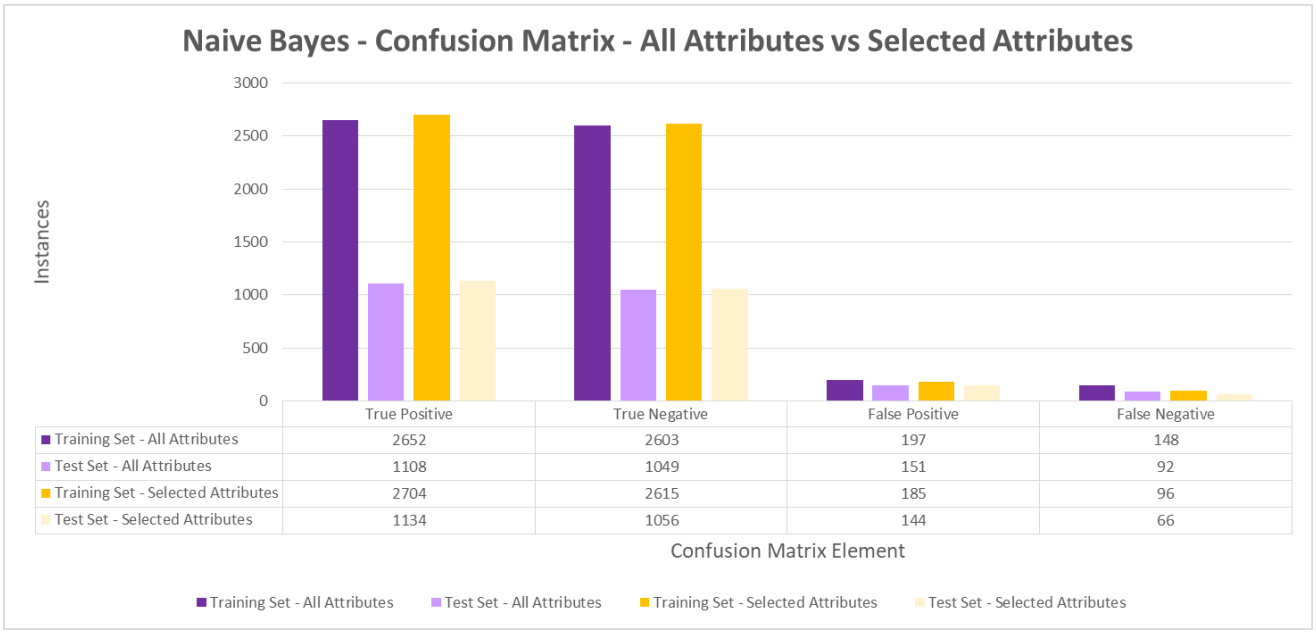
**Overall Takeaway from All vs Selected Attribute Model Comparison:**

The selected attributes model demonstrates notable improvements over the all attributes model across both confusion matrix components and performance metrics. The model using selected attributes shows enhanced classification accuracy, sensitivity, specificity, precision, and F-measure, both in training and test sets. This indicates that focusing on the most relevant features has significantly improved the model's ability to accurately and effectively classify instances, reducing misclassifications and enhancing overall performance.

By utilizing a refined subset of features, the selected attributes model achieves higher accuracy, suggesting that these chosen attributes are more predictive of the target variable. The increase in sensitivity (recall) underscores a better capability to correctly identify instances of interest ("yes"), which is crucial for effective targeting. Improved specificity and precision further indicate that the model is more adept at distinguishing between the classes, minimizing both false positives and false negatives.

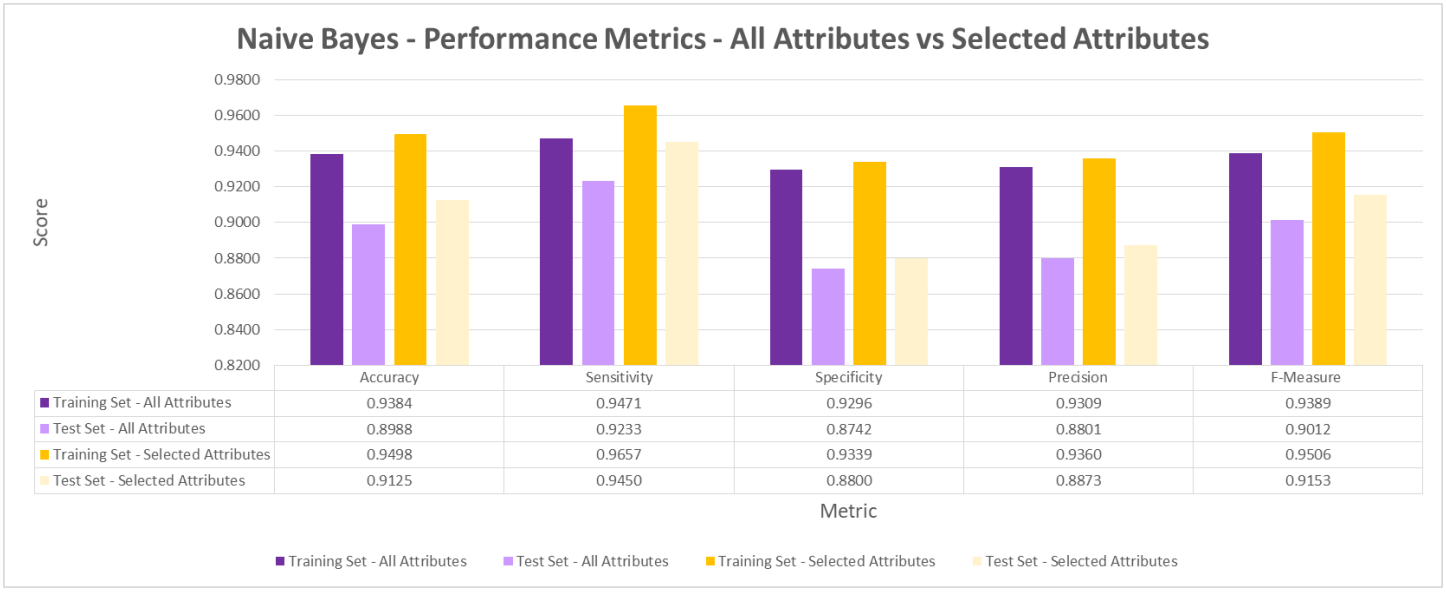
In summary, the targeted feature selection has not only boosted the overall performance metrics but also positively influenced the model's capacity to predict the target class, resulting in more accurate and reliable predictions.

**Appendix**




| | True Positive | True Negative | False Positive | False Negative |
|---|---|---|---|---|
| Training Set - All Attributes | 2652 | 2603 | 197 | 148 |
| Test Set - All Attributes | 1108 | 1049 | 151 | 92 |
| Training Set - Selected Attributes | 2704 | 2615 | 185 | 96 |
| Test Set - Selected Attributes | 1134 | 1056 | 144 | 66 |

*Figure 3: Confusion Matrix – All Attributes vs Selected Attributes*




| | Accuracy | Sensitivity | Specificity | Precision | F-Measure |
|---|---|---|---|---|---|
| Training Set - All Attributes | 0.9384 | 0.9471 | 0.9296 | 0.9309 | 0.9389 |
| Test Set - All Attributes | 0.8988 | 0.9233 | 0.8742 | 0.8801 | 0.9012 |
| Training Set - Selected Attributes | 0.9498 | 0.9657 | 0.9339 | 0.9360 | 0.9506 |
| Test Set - Selected Attributes | 0.9125 | 0.9450 | 0.8800 | 0.8873 | 0.9153 |

*Figure 4: Performance Metrics – All Attributes vs Selected Attributes*