### **Naive Bayes Algorithms**

We are evaluating the performance of several Naive Bayes algorithms, and here's a detailed overview of each model's suitability, performance, and hypotheses for the results:

#### **Multinomial Naive Bayes**

* **Suitability:** Best for categorical data or text data where features represent counts or frequencies (e.g., word counts in documents).
* **Performance:** Moderate performance with an **accuracy of 66.1%** and a **ROC AUC of 0.857**, with relatively high **specificity** (88.3%) and lower **sensitivity** (66.9%), indicating the model is better at identifying negatives (non-events).
* **Hypothesis for Performance:** The moderate performance suggests that the dataset might not be purely categorical or contain non-textual data, making it less well-suited for this algorithm. The data may not align with the assumptions of frequency-based features or text-based data.

#### **Gaussian Naive Bayes**

* **Suitability:** Suitable for continuous data, assuming features are normally distributed.
* **Performance:** Achieved **perfect performance** across all metrics, including **accuracy** (1.0), **sensitivity** (1.0), and **specificity** (1.0), which could indicate overfitting on the dataset.
* **Hypothesis for Performance:** The dataset might contain features that follow a **normal distribution** or exhibit clear patterns that align with Gaussian assumptions, leading to perfect predictions. However, the perfect scores suggest potential overfitting, where the model may not generalize well to unseen data.

#### **Complement Naive Bayes**

* **Suitability:** Designed to improve on **Multinomial Naive Bayes**, particularly for **imbalanced datasets**.
* **Performance:** Lower performance with an **accuracy of 54.9%** and a **ROC AUC of 0.676**, with low **sensitivity** (38.7%), indicating difficulty in identifying positives.
* **Hypothesis for Performance:** **Complement Naive Bayes** works well for imbalanced datasets, but the dataset might not be as imbalanced as expected, which could explain its poorer performance compared to other models. The **low sensitivity** also suggests that it struggles to identify the positive class effectively.

#### **Bernoulli Naive Bayes**

* **Suitability:** Best for binary/Boolean features.
* **Performance:** Achieved **perfect scores** across all metrics, including **accuracy** (1.0), **sensitivity** (1.0), and **specificity** (1.0).
* **Hypothesis for Performance:** The dataset likely consists of **binary** or **near-binary** features, making **Bernoulli Naive Bayes** a good fit. The model’s perfect performance suggests that binary features or simple decision boundaries align well with the data.

#### **Categorical Naive Bayes**

* **Suitability:** Effective for categorical data where features are discrete.
* **Performance:** Nearly **perfect performance** with an **accuracy of 99.6%** and a **ROC AUC of 0.99998**.
* **Hypothesis for Performance:** The dataset is likely composed primarily of **categorical features**, making **Categorical Naive Bayes** the best choice. Its high accuracy, sensitivity, specificity, and ROC AUC indicate that it aligns perfectly with the dataset’s structure

#### **Complex Bayes Net (Optional)**

* **Suitability:** More complex models that can capture dependencies between features.
* **Performance:** Could perform well if the dataset contains complex interactions or correlations between features.
* **Hypothesis for Performance:** **Complex Bayes nets** could outperform simpler models if the dataset contains intricate dependencies between features. However, their performance is highly dependent on the dataset's structure and might not always outperform simpler models, especially if the relationships between features are not complex.

### **Evaluation Metrics:**

* **Accuracy:** Measures the proportion of correct predictions. However, it can be misleading in **imbalanced datasets**.
* **Sensitivity (Recall):** Measures the proportion of true positives correctly identified. High sensitivity is important when minimizing **false negatives**.
* **Specificity:** Measures the proportion of true negatives correctly identified. Useful for ensuring negatives are classified well.
* **Precision:** Measures the proportion of true positives among all positive predictions. High precision is crucial when the cost of **false positives** is high.
* **ROC AUC:** Measures the area under the **receiver operating characteristic curve**, balancing sensitivity and specificity. A higher ROC AUC indicates a better overall model.

### **Naive Bayes Algorithms**

#### **Multinomial Naive Bayes**

* **Suitability:** Best for categorical data or text data where features represent counts or frequencies (e.g., word counts in documents).
* **Performance:** Moderate performance with an **accuracy of 66.1%** and a **ROC AUC of 0.857**, with relatively high **specificity** (88.3%) and lower **sensitivity** (66.9%), indicating the model is better at identifying negatives (non-events).
* **Hypothesis for Performance:** The moderate performance suggests that the dataset might not be purely categorical or contain non-textual data, making it less well-suited for this algorithm. The data may not align with the assumptions of frequency-based features or text-based data.

#### **Gaussian Naive Bayes**

* **Suitability:** Suitable for continuous data, assuming features are normally distributed.
* **Performance:** Achieved **perfect performance** across all metrics, including **accuracy** (1.0), **sensitivity** (1.0), and **specificity** (1.0), which could indicate overfitting on the dataset.
* **Hypothesis for Performance:** The dataset might contain features that follow a **normal distribution** or exhibit clear patterns that align with Gaussian assumptions, leading to perfect predictions. However, the perfect scores suggest potential overfitting, where the model may not generalize well to unseen data.

#### **Complement Naive Bayes**

* **Suitability:** Designed to improve on **Multinomial Naive Bayes**, particularly for **imbalanced datasets**.
* **Performance:** Lower performance with an **accuracy of 54.9%** and a **ROC AUC of 0.676**, with low **sensitivity** (38.7%), indicating difficulty in identifying positives.
* **Hypothesis for Performance:** **Complement Naive Bayes** works well for imbalanced datasets, but the dataset might not be as imbalanced as expected, which could explain its poorer performance compared to other models. The **low sensitivity** also suggests that it struggles to identify the positive class effectively.

#### **Bernoulli Naive Bayes**

* **Suitability:** Best for binary/Boolean features.
* **Performance:** Achieved **perfect scores** across all metrics, including **accuracy** (1.0), **sensitivity** (1.0), and **specificity** (1.0).
* **Hypothesis for Performance:** The dataset likely consists of **binary** or **near-binary** features, making **Bernoulli Naive Bayes** a good fit. The model’s perfect performance suggests that binary features or simple decision boundaries align well with the data.

#### **Categorical Naive Bayes**

* **Suitability:** Effective for categorical data where features are discrete.
* **Performance:** Nearly **perfect performance** with an **accuracy of 99.6%** and a **ROC AUC of 0.99998**.
* **Hypothesis for Performance:** The dataset is likely composed primarily of **categorical features**, making **Categorical Naive Bayes** the best choice. Its high accuracy, sensitivity, specificity, and ROC AUC indicate that it aligns perfectly with the dataset’s structure.

#### **Complex Bayes Net (Optional)**

* **Suitability:** More complex models that can capture dependencies between features.
* **Performance:** Could perform well if the dataset contains complex interactions or correlations between features.
* **Hypothesis for Performance:** **Complex Bayes nets** could outperform simpler models if the dataset contains intricate dependencies between features. However, their performance is highly dependent on the dataset's structure and might not always outperform simpler models, especially if the relationships between features are not complex.

### **Evaluation Metrics:**

* **Accuracy:** Measures the proportion of correct predictions. However, it can be misleading in **imbalanced datasets**.
* **Sensitivity (Recall):** Measures the proportion of true positives correctly identified. High sensitivity is important when minimizing **false negatives**.
* **Specificity:** Measures the proportion of true negatives correctly identified. Useful for ensuring negatives are classified well.
* **Precision:** Measures the proportion of true positives among all positive predictions. High precision is crucial when the cost of **false positives** is high.
* **ROC AUC:** Measures the area under the **receiver operating characteristic curve**, balancing sensitivity and specificity. A higher ROC AUC indicates a better overall model.

### **Proposed Solutions:**

#### **Best Performing Models in Each Metric:**

* **Accuracy:** **Categorical Naive Bayes** (99.1%)
* **Sensitivity:** **Gaussian Naive Bayes** (perfect sensitivity across all classes)
* **Specificity:** **Bernoulli Naive Bayes** (perfect specificity across all classes)
* **Precision:** **Categorical Naive Bayes** (high precision)
* **Recall:** **Categorical Naive Bayes** (high recall)
* **ROC AUC:** **Categorical Naive Bayes** (0.999)

### **Naive Bayes Algorithms**

#### **Multinomial Naive Bayes**

* **Suitability:** Best for categorical data or text data where features represent counts or frequencies (e.g., word counts in documents).
* **Performance:** Moderate performance with an **accuracy of 66.1%** and a **ROC AUC of 0.857**, with relatively high **specificity** (88.3%) and lower **sensitivity** (66.9%), indicating the model is better at identifying negatives (non-events).
* **Hypothesis for Performance:** The moderate performance suggests that the dataset might not be purely categorical or contain non-textual data, making it less well-suited for this algorithm. The data may not align with the assumptions of frequency-based features or text-based data.

#### **Gaussian Naive Bayes**

* **Suitability:** Suitable for continuous data, assuming features are normally distributed.
* **Performance:** Achieved **perfect performance** across all metrics, including **accuracy** (1.0), **sensitivity** (1.0), and **specificity** (1.0), which could indicate overfitting on the dataset.
* **Hypothesis for Performance:** The dataset might contain features that follow a **normal distribution** or exhibit clear patterns that align with Gaussian assumptions, leading to perfect predictions. However, the perfect scores suggest potential overfitting, where the model may not generalize well to unseen data.

#### **Complement Naive Bayes**

* **Suitability:** Designed to improve on **Multinomial Naive Bayes**, particularly for **imbalanced datasets**.
* **Performance:** Lower performance with an **accuracy of 54.9%** and a **ROC AUC of 0.676**, with low **sensitivity** (38.7%), indicating difficulty in identifying positives.
* **Hypothesis for Performance:** **Complement Naive Bayes** works well for imbalanced datasets, but the dataset might not be as imbalanced as expected, which could explain its poorer performance compared to other models. The **low sensitivity** also suggests that it struggles to identify the positive class effectively.

#### **Bernoulli Naive Bayes**

* **Suitability:** Best for binary/Boolean features.
* **Performance:** Achieved **perfect scores** across all metrics, including **accuracy** (1.0), **sensitivity** (1.0), and **specificity** (1.0).
* **Hypothesis for Performance:** The dataset likely consists of **binary** or **near-binary** features, making **Bernoulli Naive Bayes** a good fit. The model’s perfect performance suggests that binary features or simple decision boundaries align well with the data.

#### **Categorical Naive Bayes**

* **Suitability:** Effective for categorical data where features are discrete.
* **Performance:** Nearly **perfect performance** with an **accuracy of 99.6%** and a **ROC AUC of 0.99998**.
* **Hypothesis for Performance:** The dataset is likely composed primarily of **categorical features**, making **Categorical Naive Bayes** the best choice. Its high accuracy, sensitivity, specificity, and ROC AUC indicate that it aligns perfectly with the dataset’s structure.

#### **Complex Bayes Net (Optional)**

* **Suitability:** More complex models that can capture dependencies between features.
* **Performance:** Could perform well if the dataset contains complex interactions or correlations between features.
* **Hypothesis for Performance:** **Complex Bayes nets** could outperform simpler models if the dataset contains intricate dependencies between features. However, their performance is highly dependent on the dataset's structure and might not always outperform simpler models, especially if the relationships between features are not complex.

### **Evaluation Metrics:**

* **Accuracy:** Measures the proportion of correct predictions. However, it can be misleading in **imbalanced datasets**.
* **Sensitivity (Recall):** Measures the proportion of true positives correctly identified. High sensitivity is important when minimizing **false negatives**.
* **Specificity:** Measures the proportion of true negatives correctly identified. Useful for ensuring negatives are classified well.
* **Precision:** Measures the proportion of true positives among all positive predictions. High precision is crucial when the cost of **false positives** is high.
* **ROC AUC:** Measures the area under the **receiver operating characteristic curve**, balancing sensitivity and specificity. A higher ROC AUC indicates a better overall model.

### **Proposed Solutions:**

#### **Best Performing Models in Each Metric:**

* **Accuracy:** **Categorical Naive Bayes** (99.1%)
* **Sensitivity:** **Gaussian Naive Bayes** (perfect sensitivity across all classes)
* **Specificity:** **Bernoulli Naive Bayes** (perfect specificity across all classes)
* **Precision:** **Categorical Naive Bayes** (high precision)
* **Recall:** **Categorical Naive Bayes** (high recall)
* **ROC AUC:** **Categorical Naive Bayes** (0.999)

### **Hypotheses on Why Some Models Performed Better:**

* **Gaussian Naive Bayes** performed well because it assumes **normal distribution**, which might be aligned with the underlying data characteristics.
* **Categorical Naive Bayes** performed best across nearly all metrics, likely because the dataset contains **discrete categorical data**, making it the best fit.
* **Bernoulli Naive Bayes** achieved perfect specificity, likely because the dataset consists of **binary** features, and the model fits well for such cases.
* **Multinomial Naive Bayes** struggled because the dataset did not align well with the typical assumptions of this model, which are based on **frequency-based** or **text-based features**.

| Naive Bayes Algorithm | Accuracy (%) | Sensitivity (%) | Specificity (%) | Precision (%) | Recall (%) | ROC AUC |
| --- | --- | --- | --- | --- | --- | --- |
| Multinomial Naive Bayes | 55.7% | 66.9% | 66.9% | 75.9% | 66.9% | 85.7% |
| Gaussian Naive Bayes | 100% | 100% | 100% | 100% | 100% | 100% |
| Complement Naive Bayes | 54.9% | 38.7% | 38.7% | 53.5% | 54.9% | 67.6% |
| Bernoulli Naive Bayes | 100% | 100% | 100% | 100% | 100% | 100% |
| Categorical Naive Bayes | 99.6% | 99.5% | 99.8% | 99.6% | 99.5% | 99.9% |

**Gaussian Naive Bayes** and **Bernoulli Naive Bayes** are the best performing models across **Accuracy**, **Sensitivity**, **Specificity**, **Precision**, **Recall**, and **ROC AUC** with perfect scores.

**Categorical Naive Bayes** performs excellently across **Accuracy**, **Sensitivity**, **Specificity**, and **ROC AUC**, though its precision is slightly lower than the perfect scores of the other two