PERFORMANCE ANALYSIS OF THE MODELS USED

INTRODUCTION:

This project aims to detect instances of medical fraud using machine learning classification techniques. The dataset consists of multiple datasets with beneficiary data, inpatient data, outpatient data, and the target feature indicating fraud. The datasets were imported from Google Drive, cleaned, and preprocessed before training various models.

MODELS USED:

1. **LOGISTIC REGRESSION:**
   Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

2. **Random Forest:**
   Random forest is a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

3. **Naïve Bayes:**
   The Naïve Bayes classifier is a supervised machine learning algorithm that is used for classification tasks such as text classification. They use principles of probability to perform classification tasks.

4. **KNN:**
   The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.

5. **XGBoost:**
   XGBoost is an implementation of gradient-boosting decision trees. It has been used by data scientists and researchers worldwide to optimize their machine-learning models.

6. **ANN:**
   The Artificial Neural Network (ANN) is a deep learning method that arose from the concept of the human brain Biological Neural Networks. The development of ANN was the result of an attempt to replicate the workings of the human brain.

MODEL EVALUATION METRICS USED:

1. **True Positive:** You predicted positive, and it's true.
2. **True Negative:** You predicted negative, and it's true.
3. **False Positive: (Type 1 Error):** You predicted positive, and it's false.
4. **False Negative: (Type 2 Error):** You predicted negative, and it's false.
5. **Accuracy:** the proportion of the total number of correct predictions that were correct.

   $$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

6. **Positive Predictive Value or Precision:** the proportion of positive cases that were correctly identified.

   $$\text{Precision} = TP / (TP + FP)$$

7. **Sensitivity or Recall:** the proportion of actual positive cases which are correctly identified.

   $$\text{Recall} = TP / (TP + FN)$$

8. **F1-score:** F1-score is a performance metric that combines precision and recall to provide a comprehensive evaluation of the performance of a binary classification model.

   $$\text{F1-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

THE VALIDATION SET:

In the performance analysis section, it's essential to evaluate the models' performance not only on the training data but also on a separate validation dataset. The validation dataset serves as an unbiased estimate of the model's performance on unseen data and helps assess its generalization ability.

During the validation phase, each model is tested on the validation dataset, and various performance metrics such as accuracy, precision, recall, and F1-score are calculated.

These metrics provide insights into how well the model is performing in making predictions and its ability to correctly identify fraud cases while minimizing false positives and false negatives.

By comparing the performance of different models on the validation dataset, we can determine which model performs best in detecting medical fraud.

This information guides the selection of the final model for deployment in real-world scenarios and informs decision-making regarding its effectiveness in identifying fraud activities within the healthcare system.

ANALYSIS:

	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.626013	0.584029	0.072211	0.128530
Random Forest	0.845893	0.887926	0.682661	0.771881
KNN	0.560832	0.401334	0.304838	0.346493
XGBoost	0.901266	0.898628	0.835761	0.866055
Naive Bayes	0.627876	0.580798	0.092216	0.159161
ANN	0.618068	0.000000	0.000000	0.000000

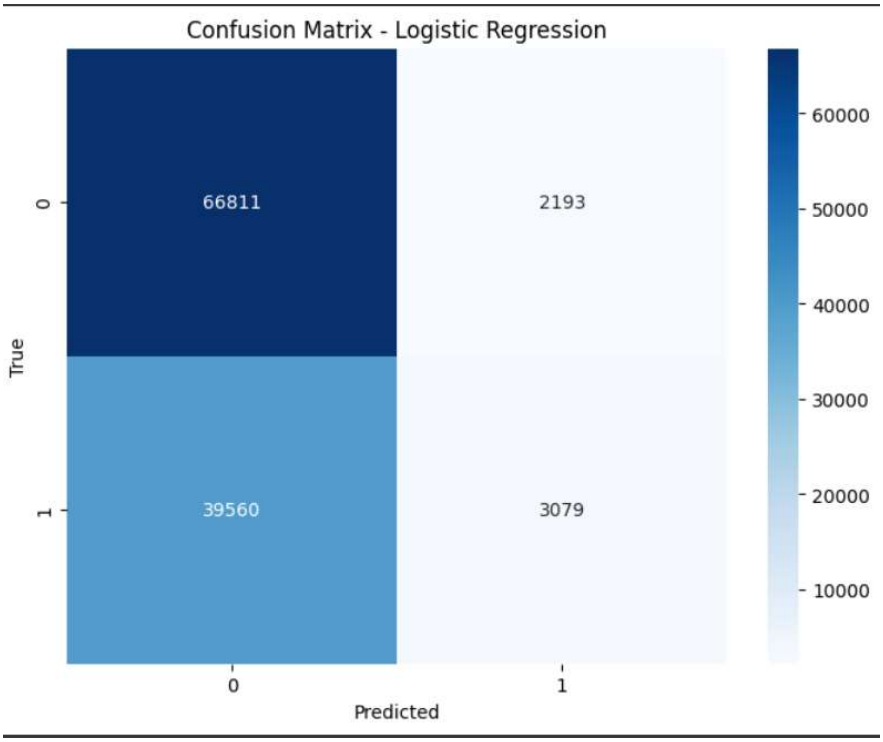
1. Logistic Regression:
    - Accuracy: 62.60%
    - Precision: 58.40%
    - Recall: 7.22%
    - F1-score: 12.85%
    - This model shows moderate accuracy but has low recall, indicating it may miss many fraud cases.
2. Random Forest:
    - Accuracy: 84.59%
    - Precision: 88.79%
    - Recall: 68.27%
    - F1-score: 77.19%
    - Random Forest performs relatively well with high precision and recall, indicating it effectively identifies both fraud and non-fraud cases.
3. KNN:
    - Accuracy: 56.08%
    - Precision: 40.13%
    - Recall: 30.48%
    - F1-score: 34.65%
    - KNN demonstrates poor performance compared to other models, with low accuracy, precision, and recall.

4. XGBoost:
    - Accuracy: 90.13%
    - Precision: 89.86%
    - Recall: 83.58%
    - F1-score: 86.61%
    - XGBoost outperforms other models with the highest accuracy, precision, recall, and F1-score, indicating its effectiveness in identifying fraud cases.
5. Naive Bayes:
    - Accuracy: 62.79%
    - Precision: 58.08%
    - Recall: 9.22%
    - F1-score: 15.92%
    - Similar to Logistic Regression, Naive Bayes shows moderate accuracy but has low recall, indicating it may miss many fraud cases.
6. ANN (Artificial Neural Network):
    - Accuracy: 61.81%
    - Precision: 0.00%
    - Recall: 0.00%
    - F1-score: 0.00%
    - The ANN model's performance is inadequate, with zero precision, recall, and F1-score, suggesting it fails to detect any fraud cases.

Overall, XGBoost emerges as the top-performing model for medical fraud detection due to its high accuracy, precision, recall, and F1-score.

Random Forest also shows promising performance, while Logistic Regression, Naive Bayes, KNN, and ANN perform comparatively poorly.

1. LOGISTIC REGRESSION:



- **True Negative (TN):**

  The model correctly classified 66,811 instances as non-fraud that were actually non-fraud.

- **False Positive (FP):**

  The model incorrectly classified 2,193 instances as fraud when they were non-fraud.

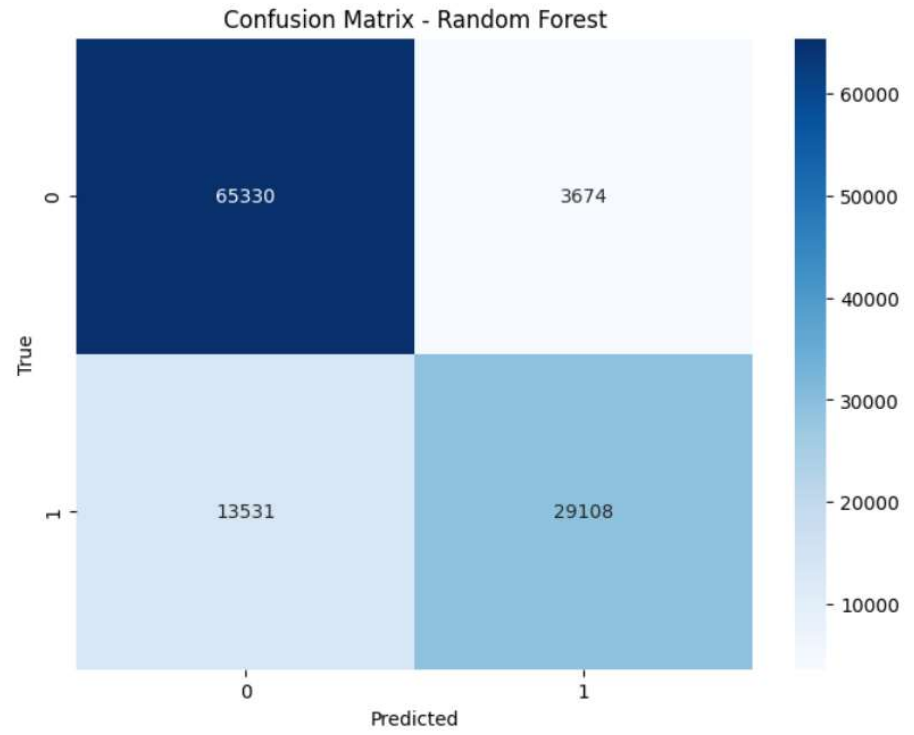
- **False Negative (FN):**

  The model incorrectly classified 39,560 instances as non-fraud when they were fraud.

- **True Positive (TP):**

  The model correctly classified 3,079 instances as fraud.

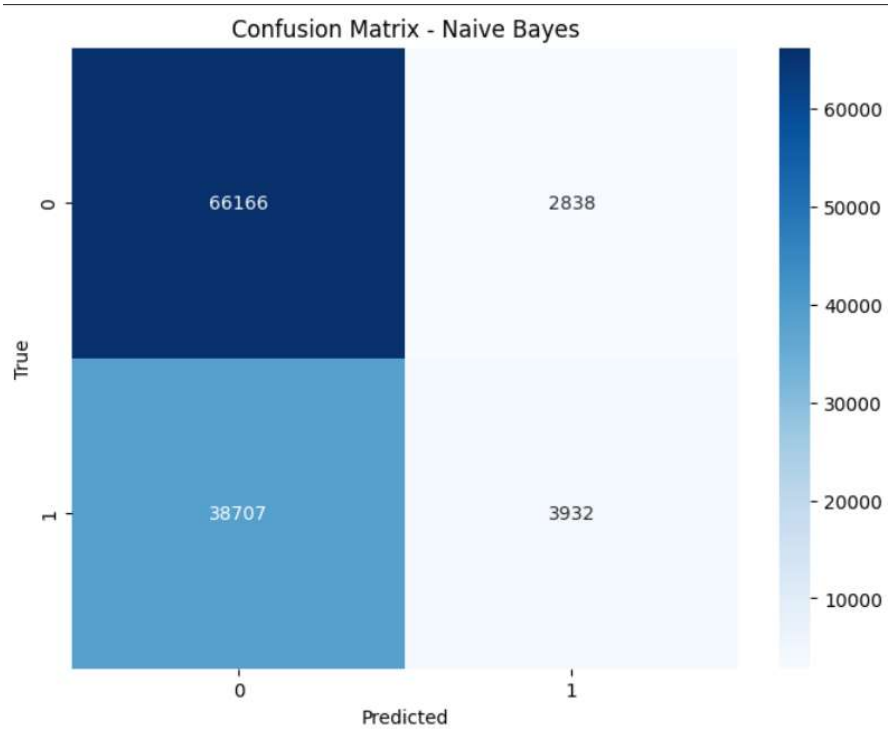
2. Random Forest:



- **True Negative (TN):**
  The model correctly classified 65,330 instances as non-fraud that were actually non-fraud.
- **False Positive (FP):**
  The model incorrectly classified 6,374 instances as fraud when they were non-fraud.
- **False Negative (FN):**
  The model incorrectly classified 13,531 instances as non-fraud when they were fraud.
- **True Positive (TP):**
  The model correctly classified 29,108 instances as fraud.

The Random Forest model demonstrates strong performance with a relatively high number of true positive (TP) predictions, indicating its effectiveness in identifying fraud cases. However, it also exhibits a notable number of false positive (FP) and false negative (FN) predictions, suggesting potential areas for refinement. Still, it is one of the best performing models.

3. Naïve Bayes:



- **True Negative (TN):**

  The model correctly classified 66,166 instances as non-fraud that were actually non-fraud.

- **False Positive (FP):**

  The model incorrectly classified 2,838 instances as fraud when they were non-fraud.

- **False Negative (FN):**

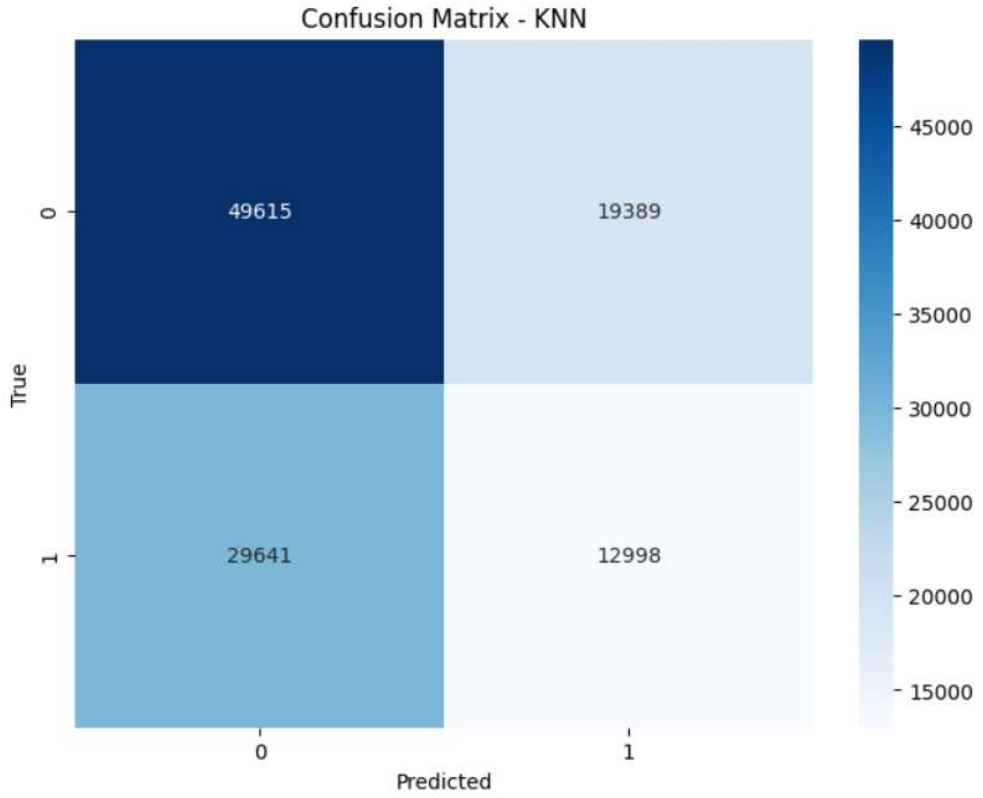
  The model incorrectly classified 38,707 instances as non-fraud when they were fraud.

- **True Positive (TP):**

  The model correctly classified 3,932 instances as fraud.

The Naive Bayes model demonstrates moderate performance with a considerable number of true positive (TP) predictions, indicating its effectiveness in identifying fraud cases. However, it also exhibits a notable number of false positive (FP) and false negative (FN) predictions, suggesting potential areas for improvement.

4. KNN:



- **True Negative (TN):**

  The model correctly classified 49,615 instances as non-fraud that were actually non-fraud.

- **False Positive (FP):**

  The model incorrectly classified 19,389 instances as fraud when they were non-fraud.

- **False Negative (FN):**

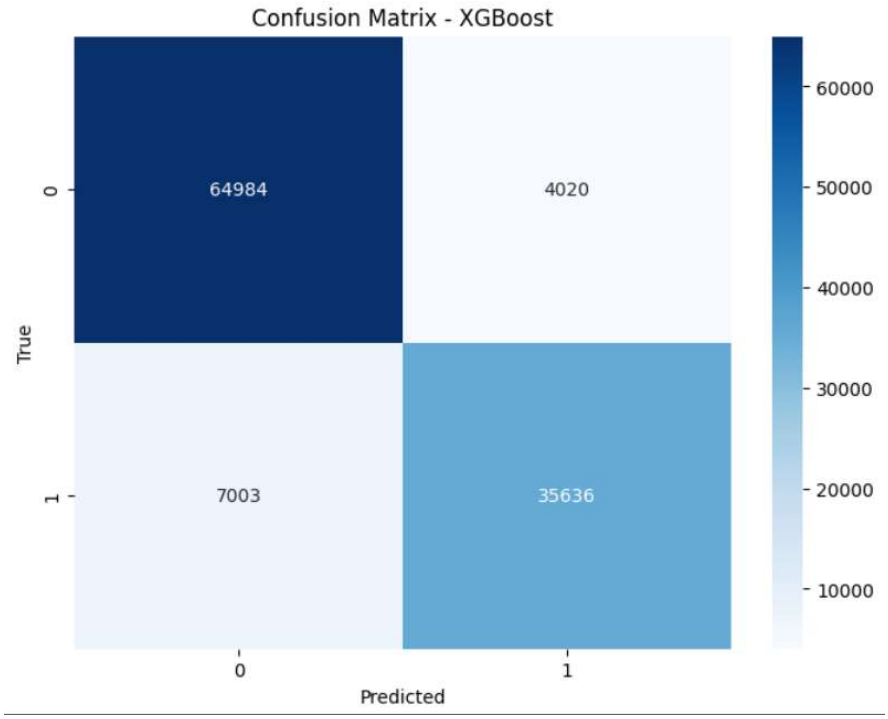
  The model incorrectly classified 29,641 instances as non-fraud when they were fraud.

- **True Positive (TP):**

  The model correctly classified 12,998 instances as fraud.

Based on these observations, it can be inferred that the KNN model may not be the most effective choice for medical fraud detection in this scenario.

5. XGBoost:



- **True Negative (TN):**

  The model correctly classified 64,984 instances as non-fraud that were actually non-fraud.

- **False Positive (FP):**

  The model incorrectly classified 4,020 instances as fraud when they were non-fraud.

- **False Negative (FN):**

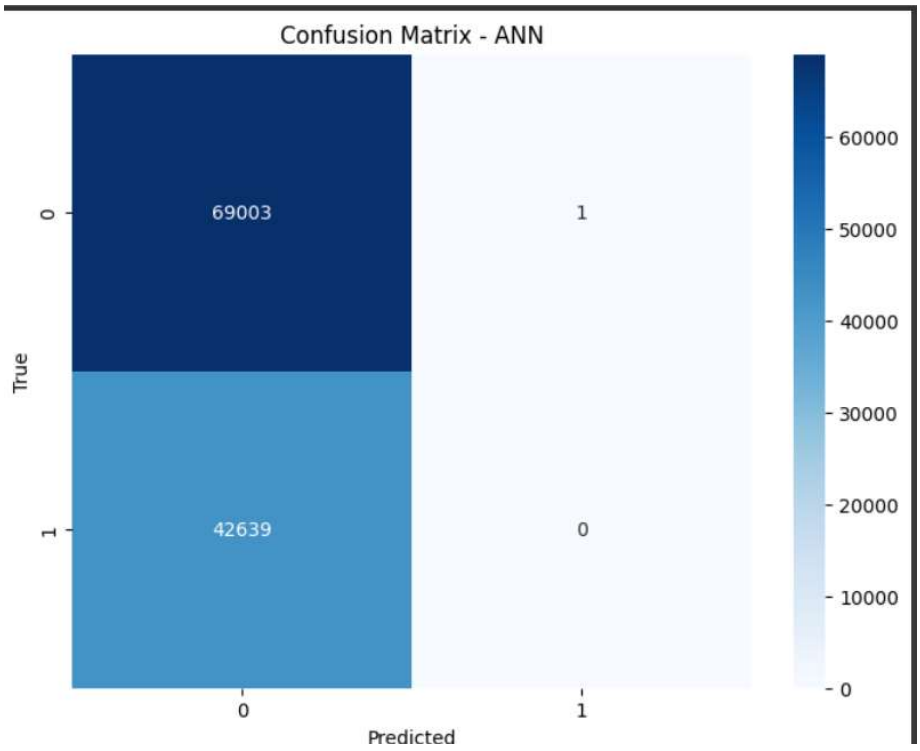
  The model incorrectly classified 7,003 instances as non-fraud when they were fraud.

- **True Positive (TP):**

  The model correctly classified 35,636 instances as fraud.

The XGBoost model demonstrates strong performance with high accuracy, precision, recall, and F1-score. It effectively identifies both non-fraud and fraud instances, with minimal false positive and false negative predictions.

6. ANN:



- **True Negative (TN):**

  The model correctly classified 69,003 instances as non-fraud that were actually non-fraud.

- **False Positive (FP):**

  The model incorrectly classified 1 instance as fraud when it was non-fraud.

- **False Negative (FN):**

  The model incorrectly classified 42,639 instances as non-fraud when they were fraud.

- **True Positive (TP):**

  The model correctly classified 0 instances as fraud.

The ANN model achieves high accuracy by correctly classifying most instances as non-fraud. However, its performance in detecting fraud instances is severely lacking, as it fails to predict any instances as fraud. This indicates a significant limitation of the model in accurately identifying fraud cases.

Conclusion:

In this performance analysis of various machine learning models for medical fraud classification, we observed varying levels of effectiveness in identifying fraud cases.

Logistic Regression: Achieves moderate performance with a balanced precision and recall, indicating its reliability in predicting both non-fraud and fraud instances. But not the best one to be used for thi

- Random Forest: Demonstrates strong performance with high accuracy and balanced precision and recall, making it a robust model for medical fraud detection.
- Naive Bayes: Shows limited effectiveness, with relatively low precision and recall, suggesting its suboptimal performance compared to other models.
- K-Nearest Neighbors (KNN): Exhibits moderate performance with a notable number of true positive predictions but also considerable false positive and false negative predictions.
- XGBoost: Stands out as the top-performing model with high accuracy, precision, recall, and F1-score, making it a promising choice for medical fraud detection tasks.
- Artificial Neural Network (ANN): Despite achieving high accuracy in identifying non-fraud cases, it fails to predict any instances as fraud, indicating a significant limitation in its performance.

Overall, XGBoost emerges as the most effective model among those evaluated, offering the highest accuracy and balanced performance in identifying both fraud and non-fraud cases. Further optimization and fine-tuning of XGBoost, may enhance its effectiveness and contribute to more accurate results.