# Evaluation of the quality of synthetic dataset for fraud detection

Ali Valiyev

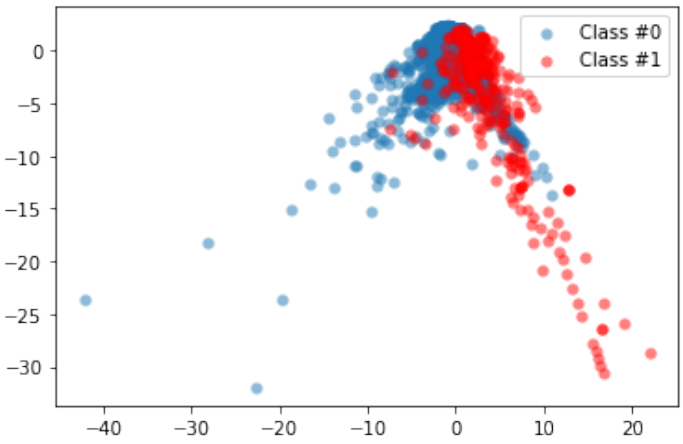
March 2022

## 1 Introduction

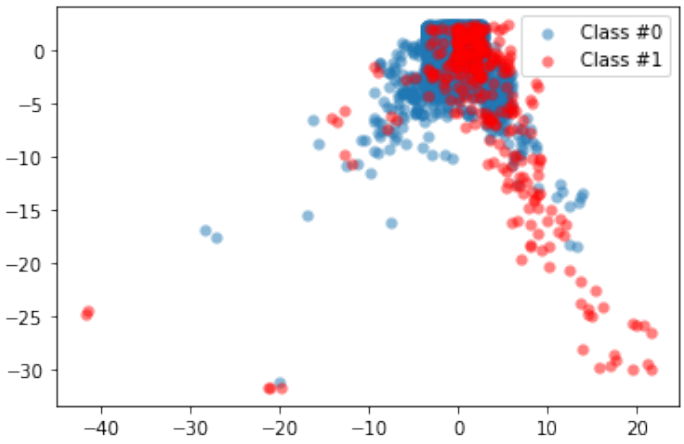
The objective of our experiment is to analyze the performance of Random Forest, Naïve Bayes, Logistic Regression, and K-nearest neighbors machine learning models for evaluation of the utility of synthesized data. We will use the accuracy, ROC curve, classification report, confusion matrix, and precision-recall curve to evaluate the model's performance.

Let us start with understanding our data with scatter plots of our datasets. From these graphs, we can visualize fraud and no fraud distributions of raw, SDV Gaussian, SDV TVAE, and DS correlated attribute datasets. As can be noticed from the graphs, the closest distribution to the raw dataset is synthetic data synthesized with DS with correlated attributes. In the scatter plot for the SDV Gaussian (Figure (d)); however, we observe that the distribution of points class 1 and class 0 is different than the raw dataset in general, and all fraud elements are distributed in one area of this graph. Hence, the worst distribution in terms of closeness to the raw data distribution of data points is in SDV Gaussian.

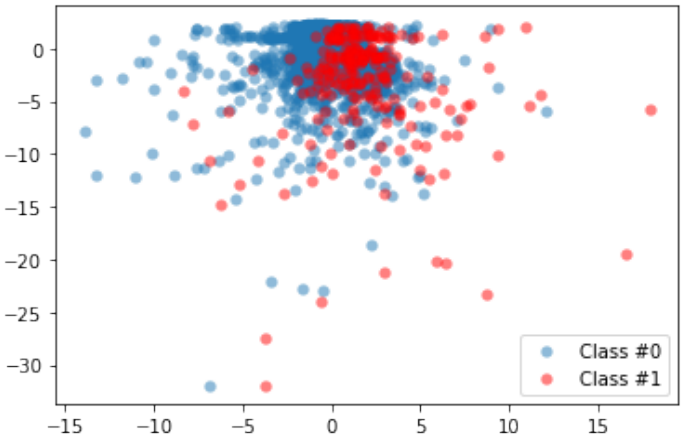
In the following chapters, we will continue to analyze the performance metrics of different machine-learning prediction algorithms on raw and synthesized data.



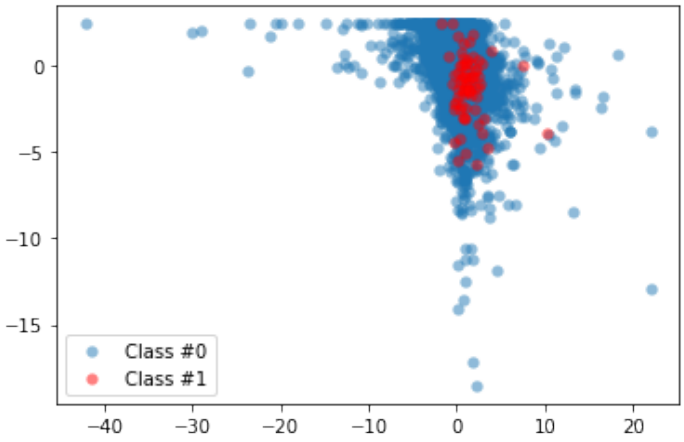

(a) Scatter plot of a raw dataset




(b) Scatter plot of a synthetic dataset(DS correlated)




(c) Scatter plot of a synthetic dataset(SDV TVAE)



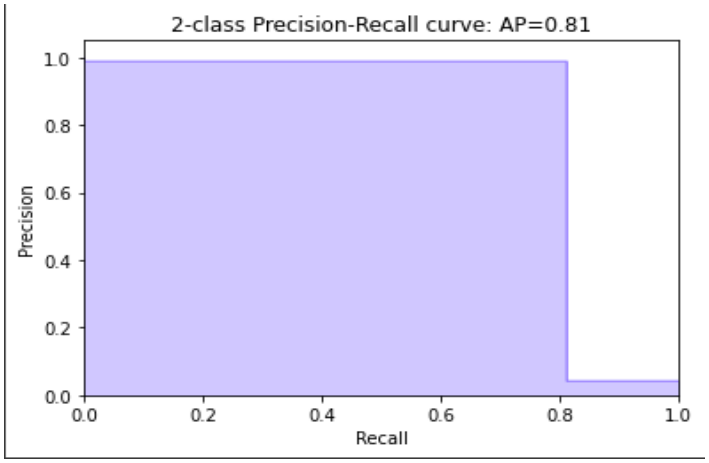

(d) Scatter plot of a synthetic dataset(SDV Gaussian)

## 2 Random forest model

We can see from Figure 3 that the performance metrics of the synthetic dataset formed with DataSythesizer with the correlated attribute are very close to the performance of the raw dataset (Figure 2). After DataSythesizer correlated attribute, the second-best performance on the Random forest model is observed on SDV TVAE data (Figure 5). For SDV Gaussian (Figure 4), although the accuracy score and ROC curve results are very close to the performance of a raw dataset, from the classification report, it can easily be seen that the precision, recall, and f1-scores are equal to 0 for the class 1 column, which is very low with the compared to a raw dataset.

```
Accuracy of Random Forest: 0.9917808219178083

The ROC curve of Random Forest: 0.961093456329283

Classification report of Random Forest:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00      2099
         1.0       0.99      0.81      0.89        91

    accuracy                           0.99      2190
   macro avg       0.99      0.91      0.94      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of Random Forest:
 [[2098    1]
 [  17   74]]
```

(a) Performance metrics on a raw dataset

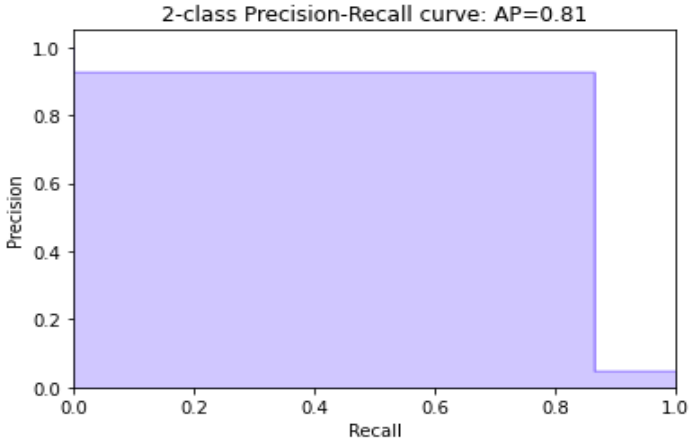



(b) Precision Recall Curve on a raw dataset

Figure 2: Performance metrics of a Random Forest model on a raw dataset

```
Accuracy of Random Forest: 0.9904109589041096

The ROC curve of Random Forest: 0.9461934877203334

Classification report of Random Forest:
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99      2086
         1.0       0.93      0.87      0.90       104

    accuracy                           0.99      2190
   macro avg       0.96      0.93      0.95      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of Random Forest:
 [[2079    7]
 [  14   90]]
```

(a) Performance metrics on a synthetic dataset(DS correlated)

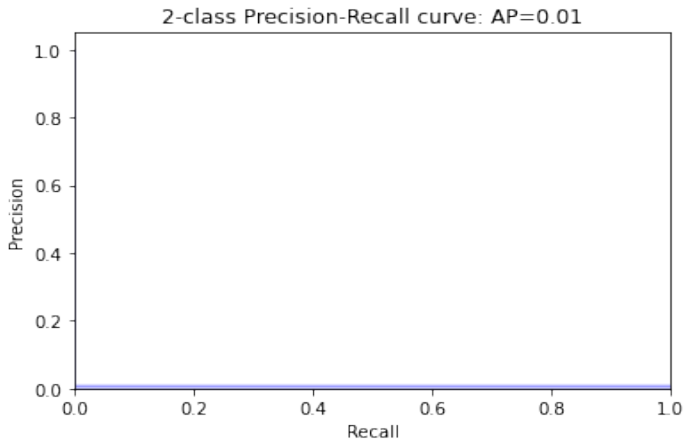



(b) Precision Recall Curve on a synthetic dataset(DS correlated)

Figure 3: Performance metrics of a Random Forest model on a synthetic dataset(DS correlated)

```
Accuracy of Random Forest: 0.9904109589041096

The ROC curve of Random Forest: 0.9615469055303079

Classification report of Random Forest:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00      2169
         1.0       0.00      0.00      0.00        21

    accuracy                           0.99      2190
   macro avg       0.50      0.50      0.50      2190
weighted avg       0.98      0.99      0.99      2190


Confusion matrix of Random Forest:
 [[2169    0]
 [  21    0]]
```

(a) Performance metrics on a synthetic dataset (SDV Gaussian)

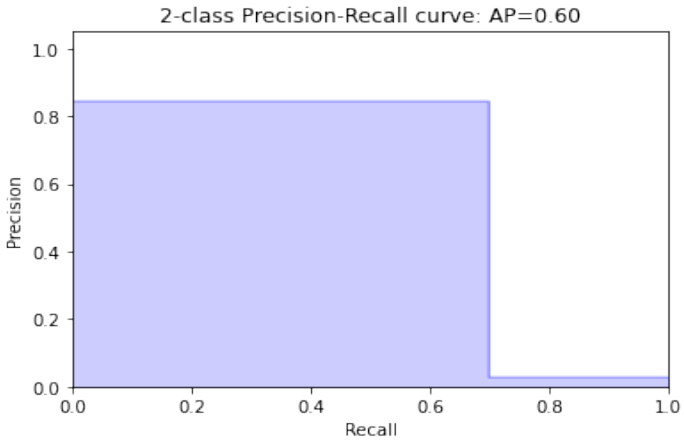



(b) Precision Recall Curve on a synthetic dataset(SDV Gaussian)

Figure 4: Performance metrics of a Random Forest model on a synthetic dataset (SDV Gaussian)

```
Accuracy of Random Forest: 0.9876712328767123

The ROC curve of Random Forest: 0.9903993253781688

Classification report of Random Forest:
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99      2127
         1.0       0.85      0.70      0.77        63

    accuracy                           0.99      2190
   macro avg       0.92      0.85      0.88      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of Random Forest:
 [[2119    8]
 [  19   44]]
```

(a) Performance metrics on a synthetic dataset(SDV TVAE)




(b) Precision Recall Curve on a synthetic dataset(SDV TVAE)

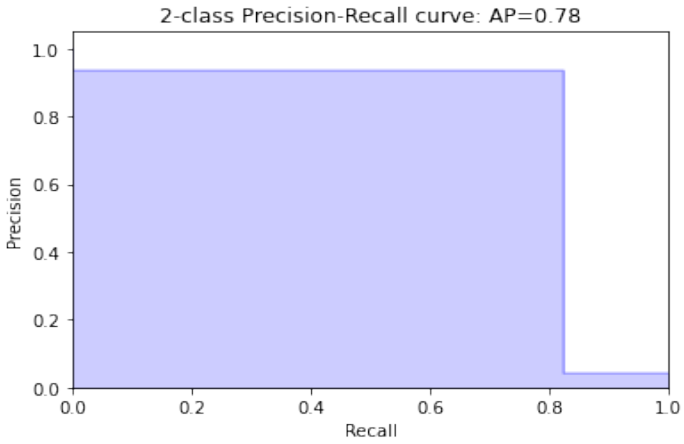
Figure 5: Performance metrics of a Random Forest model on a synthetic dataset (SDV TVAE)

## 3 Logistics Regression model

Better results for SDV Gaussian compared to other models can be observed when using the Logistic Regression model. The precision, recall, and f1 scores are higher than 30 percent, whereas these performance metrics' accuracies were equal to 0 percent in KNN and Random Forest. Furthermore, the performance metrics of DS correlated are almost higher than 80 percent and again very close to the model trained on the real dataset. The synthetic dataset formed by SDV TVAE also has a good performance on the Logistic Regression model.

```
Accuracy of Logistic Regression: 0.9904109589041096

The ROC curve of Logistic Regression: 0.968781575737269

Classification report of Logistic Regression:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00      2099
         1.0       0.94      0.82      0.88        91

    accuracy                           0.99      2190
   macro avg       0.96      0.91      0.94      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of Logistic Regression:
 [[2094    5]
 [  16   75]]
```

(a) Performance metrics on a raw dataset

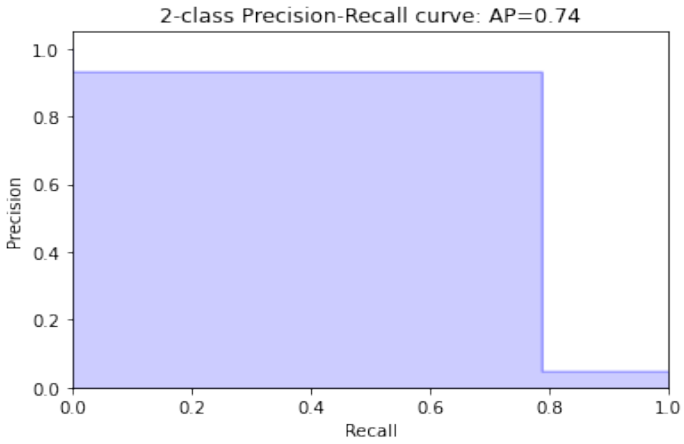



(b) Precision Recall Curve on a raw dataset

Figure 6: Performance metrics of a Logistics Regression model on a raw dataset

```
Accuracy of Logistic Regression: 0.9872146118721461

The ROC curve of Logistic Regression: 0.9591968434250313

Classification report of Logistic Regression:
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99      2086
         1.0       0.93      0.79      0.85       104

    accuracy                           0.99      2190
   macro avg       0.96      0.89      0.92      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of Logistic Regression:
 [[2080    6]
 [  22   82]]
```

(a) Performance metrics on a synthetic dataset(DS correlated)

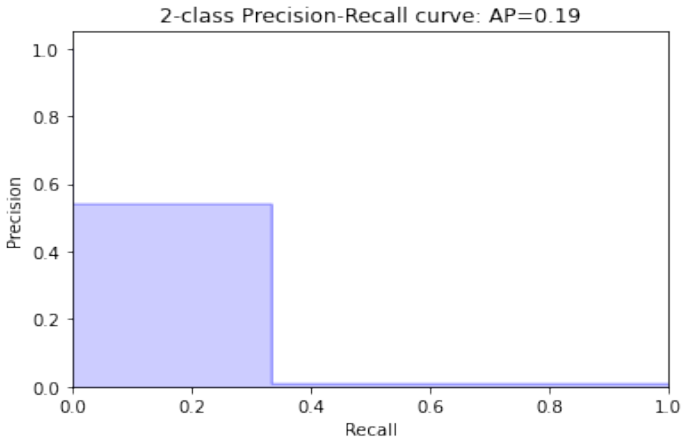



(b) Precision Recall Curve on a synthetic dataset(DS correlated)

Figure 7: Performance metrics of a Logistics Regression model on a synthetic dataset(DS correlated)

```
Accuracy of Logistic Regression: 0.9908675799086758

The ROC curve of Logistic Regression: 0.9757184570462578

Classification report of Logistic Regression:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00      2169
         1.0       0.54      0.33      0.41        21

    accuracy                           0.99      2190
   macro avg       0.77      0.67      0.70      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of Logistic Regression:
 [[2163    6]
 [  14    7]]
```

(a) Performance metrics on a synthetic dataset(SDV Gaussian)

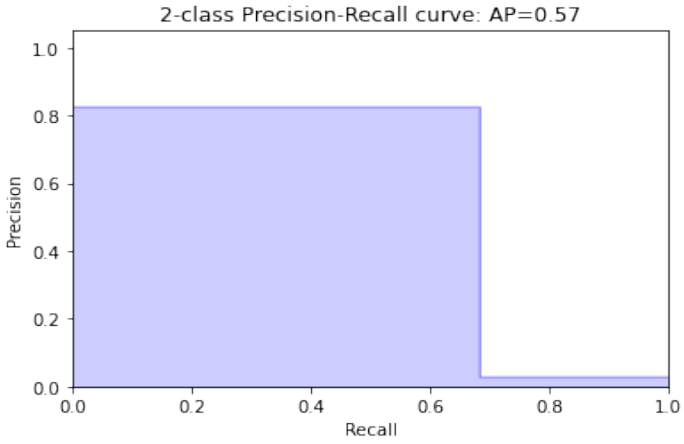



(b) Precision Recall Curve on a synthetic dataset(SDV Gaussian)

Figure 8: Performance metrics of a Logistics Regression model on a synthetic dataset(SDV Gaussian)

```
Accuracy of Logistic Regression: 0.9867579908675799

The ROC curve of Logistic Regression: 0.9859105529063216

Classification report of Logistic Regression:
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99      2127
         1.0       0.83      0.68      0.75        63

    accuracy                           0.99      2190
   macro avg       0.91      0.84      0.87      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of Logistic Regression:
 [[2118    9]
 [  20   43]]
```

(a) Performance metrics on a synthetic dataset(SDV TVAE)




(b) Precision Recall Curve on a synthetic dataset(SDV TVAE)

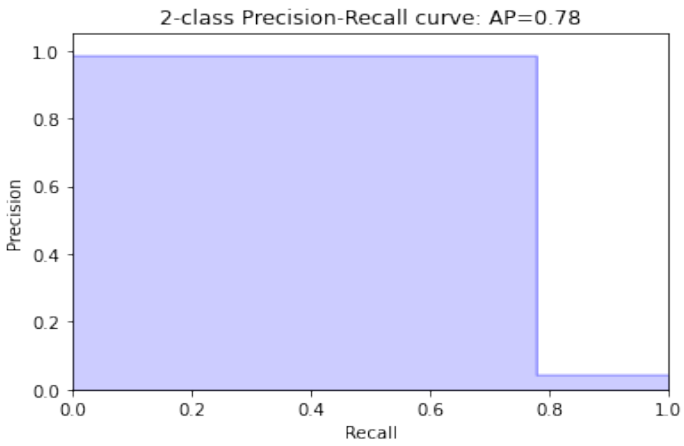
Figure 9: Performance metrics of a Logistics Regression model on a synthetic dataset(SDV TVAE)

# 4 K-nearest neighbors model

We continue the discussion with K-nearest neighbors. From Figure 12, we can observe that the true negative value is equal to 0 in the confusion matrix, whereas this value on a raw dataset is equal to 71, which is a huge difference. For SDV TVAE synthetic dataset, we can see that the results of the classification report, especially for class 1 (fraud), are lower than the Logistics Regression and Random Forest models. However, performance metrics and the Precision-Recall curve of DS correlated attribute are very close to the raw dataset.

```
Accuracy of K Nearest Neighbor: 0.9904109589041096

The ROC curve of K Nearest Neighbor: 0.9465286976006365

Classification report of K Nearest Neighbor:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00      2099
         1.0       0.99      0.78      0.87        91

    accuracy                           0.99      2190
   macro avg       0.99      0.89      0.93      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of K Nearest Neighbor:
 [[2098    1]
 [  20   71]]
```

(a) Performance metrics n a raw dataset

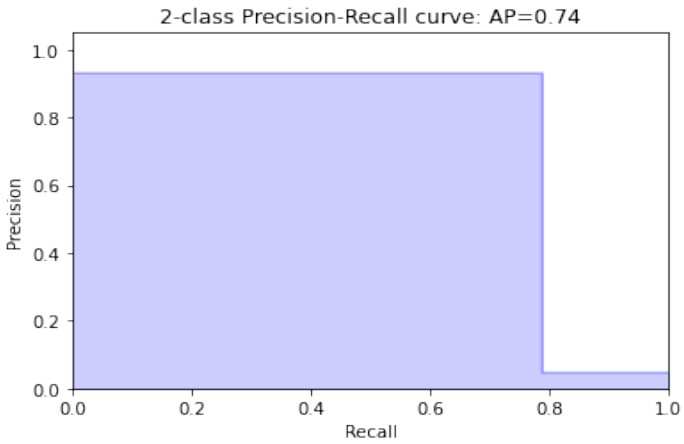



(b) Precision Recall Curve on a raw dataset

Figure 10: Performance metrics of a K-nearest neighbors model on a raw dataset

```
Accuracy of K Nearest Neighbor: 0.9872146118721461

The ROC curve of K Nearest Neighbor: 0.9324433955306438

Classification report of K Nearest Neighbor:
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99      2086
         1.0       0.93      0.79      0.85       104

    accuracy                           0.99      2190
   macro avg       0.96      0.89      0.92      2190
weighted avg       0.99      0.99      0.99      2190


Confusion matrix of K Nearest Neighbor:
 [[2080    6]
 [  22   82]]
```

(a) Performance metrics on a synthetic dataset(DS correlated)

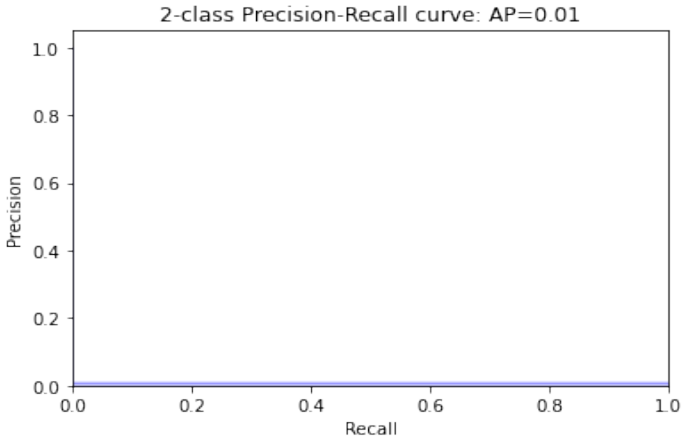



(b) Precision Recall Curve on a synthetic dataset(DS correlated)

Figure 11: Performance metrics of a K-nearest neighbors model on a synthetic dataset(DS correlated)

```
Accuracy of K Nearest Neighbor: 0.9904109589041096

The ROC curve of K Nearest Neighbor: 0.8593712265911435

Classification report of K Nearest Neighbor:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00      2169
         1.0       0.00      0.00      0.00        21

    accuracy                           0.99      2190
   macro avg       0.50      0.50      0.50      2190
weighted avg       0.98      0.99      0.99      2190


Confusion matrix of K Nearest Neighbor:
 [[2169    0]
 [  21    0]]
```

(a) Performance metrics on a synthetic dataset(SDV Gaussian)

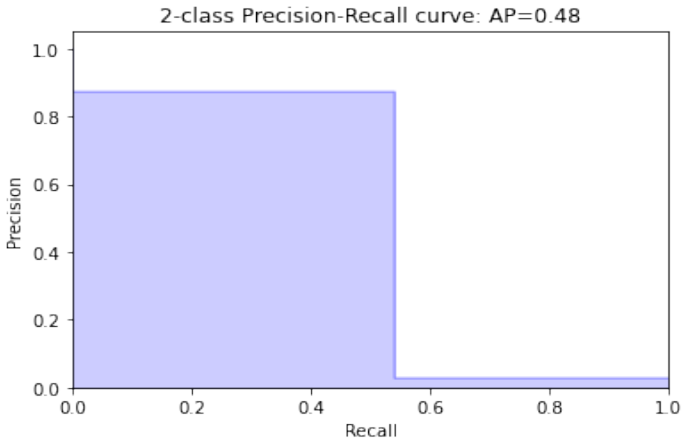



(b) Precision Recall Curve on a synthetic dataset(SDV Gaussian)

Figure 12: Performance metrics of a K-nearest neighbors model on a synthetic dataset(SDV Gaussian)

```
Accuracy of K Nearest Neighbor: 0.9844748858447488

The ROC curve of K Nearest Neighbor: 0.945888463518929

Classification report of K Nearest Neighbor:
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99      2127
         1.0       0.87      0.54      0.67        63

    accuracy                           0.98      2190
   macro avg       0.93      0.77      0.83      2190
weighted avg       0.98      0.98      0.98      2190


Confusion matrix of K Nearest Neighbor:
 [[2122    5]
 [  29   34]]
```

(a) Performance metrics on a synthetic dataset(SDV TVAE)




(b) Precision Recall Curve on a synthetic dataset(SDV TVAE)

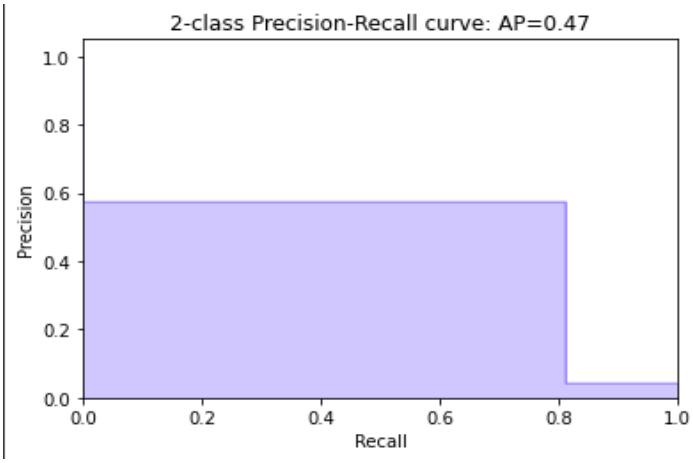
Figure 13: Performance metrics of a K-nearest neighbors model on a synthetic dataset(SDV TVAE)

# 5 Naive Bayes model

At first glance, it appears that the accuracy scores and ROC curve of raw, DS, and SDV are close to the original dataset. However, from the confusion matrix, we can see that the number of false-positive predictions is very high for both raw and synthetic datasets. In addition, the values of precision and f1-score for class 1 (frauds) are lower than 60 percent, which is insignificant compared to other models that we have analyzed so far. The Precision-Recall curve also shows us that the Naive Bayes model is not appropriate for the fraud detection task.

```
Accuracy of Naive Bayes: 0.9671232876712329

The ROC curve of Naive Bayes: 0.9529708024229226

Classification report of Naive Bayes:
              precision    recall  f1-score   support

         0.0       0.99      0.97      0.98      2099
         1.0       0.57      0.81      0.67        91

    accuracy                           0.97      2190
   macro avg       0.78      0.89      0.83      2190
weighted avg       0.97      0.97      0.97      2190


Confusion matrix of Naive Bayes:
 [[2044   55]
 [  17   74]]
```

(a) Performance metrics on a raw dataset

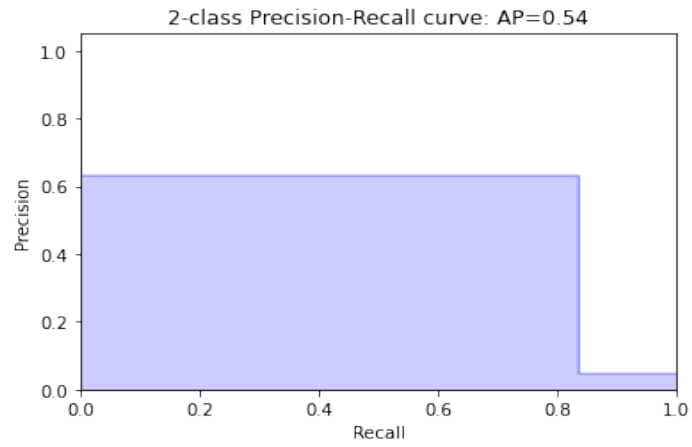



(b) Precision Recall Curve on a raw dataset

Figure 14: Performance metrics of a Naive Bayes model on a raw dataset

```
Accuracy of Naive Bayes: 0.9689497716894977

The ROC curve of Naive Bayes: 0.9426764510657127

Classification report of Naive Bayes:
              precision    recall  f1-score   support

         0.0       0.99      0.98      0.98      2086
         1.0       0.63      0.84      0.72       104

    accuracy                           0.97      2190
   macro avg       0.81      0.91      0.85      2190
weighted avg       0.97      0.97      0.97      2190


Confusion matrix of Naive Bayes:
 [[2035   51]
 [  17   87]]
```

(a) Performance metrics on a synthetic dataset(DS correlated)

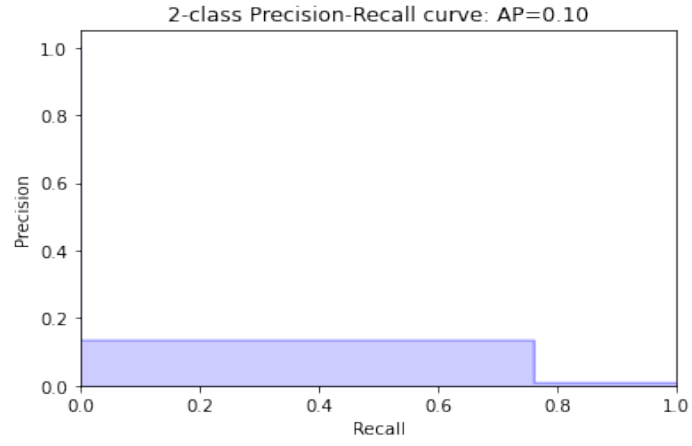



(b) Precision Recall Curve on a synthetic dataset(DS correlated)

Figure 15: Performance metrics of a Naive Bayes model on a synthetic dataset(DS correlated)

```
Accuracy of Naive Bayes: 0.9506849315068493

The ROC curve of Naive Bayes: 0.9367713890535467

Classification report of Naive Bayes:
              precision    recall  f1-score   support

         0.0       1.00      0.95      0.97      2169
         1.0       0.13      0.76      0.23        21

    accuracy                           0.95      2190
   macro avg       0.57      0.86      0.60      2190
weighted avg       0.99      0.95      0.97      2190


Confusion matrix of Naive Bayes:
 [[2066  103]
 [   5   16]]
```

(a) Performance metrics on a synthetic dataset(SDV Gaussian)

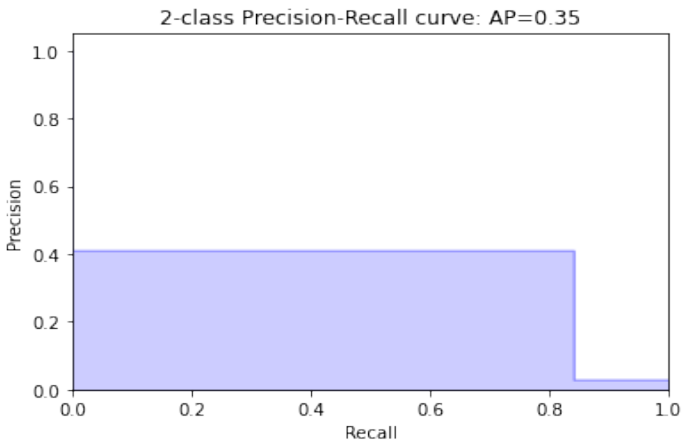



(b) Precision Recall Curve on a synthetic dataset(SDV Gaussian)

Figure 16: Performance metrics of a Naive Bayes model on a synthetic dataset(SDV Gaussian)

```
Accuracy of Naive Bayes: 0.9602739726027397

The ROC curve of Naive Bayes: 0.9736494503772359

Classification report of Naive Bayes:
              precision    recall  f1-score   support

         0.0       1.00      0.96      0.98      2127
         1.0       0.41      0.84      0.55        63

    accuracy                           0.96      2190
   macro avg       0.70      0.90      0.76      2190
weighted avg       0.98      0.96      0.97      2190


Confusion matrix of Naive Bayes:
 [[2050   77]
 [  10   53]]
```

(a) Performance metrics on a synthetic dataset(SDV TVAE)




(b) Precision Recall Curve on a synthetic dataset(SDV TVAE)

Figure 17: Performance metrics of a Naive Bayes model on a synthetic dataset(SDV TVAE)

# 6 Conclusions

Based on the analysis presented, we can conclude that the data synthesized by DS correlated attribute with disabled differential privacy is more suitable for fraud detection in supervised machine learning. It is recommended to use the Random Forest, Logistic Regression, and K-nearest neighbors models for unbalanced data because their performance metrics and Precision-Recall curves on SDV TVAE and especially DS correlated attribute synthetic datasets have higher accuracy scores and closer results to the raw datasets. However, for SDV Gaussian synthetic dataset, all models have very low precision, recall, and f1-score. Also, we need to point out that the Logistic Regression's model result for SDV Gaussian was relatively better compared to other models.