IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

Novel-based hybrid approach for prediction of Imbalanced Data using Sampling Strategy

Dr. Shiva Prasad K M[1], Afrin Banu[2], Amoolya M[3], ChandanaK[4], Dommuru Shreya[5]

Assistant Professor, Dept of Computer Science and Engineering[1]

Students, Dept of Computer Science and Engineering, Rao Bahadur Y Mahabaleshwarappa Engineering College , Ballari, India. (Affiliated To Visvesvaraya Technological University, Belgaum. Approved By AICTE, New Delhi & Accredited By NAAC With A+) Ballari – 583104, Karnataka[2,3,4,5]

Abstract: Real-world applications frequently use data with an unbalanced class distribution, meaning that the bulk of the data belongs to the majority class and the minority class is underrepresented. The classifier tends to anticipate that the majority of the incoming data will belong to the majority class in this scenario if all the data are utilized as training data. In the imbalanced class distribution problem, it is crucial to choose the appropriate training data for prediction and classification. In our project, we provide a unique hybrid algorithm with a mix of sampling strategies for choosing representative data as training data to enhance the prediction accuracy of dependent and independent data on an unbalanced class distribution problem.

Keywords: Imbalanced Data analysis, Oversampling, Under-sampling, Hybrid Sampling

I. INTRODUCTION

The current catchphrase, "Big Data," positions itself as the ideal solution for issues caused by the enormous amount of data. Big data technologies are presently are the most qualified candidates for data analytics thanks to their support for data's volume, velocity, and variety. Large data sets intended to be analyzed to find interesting patterns. Data imbalance results from the lack of intriguing patterns inside the content. Although there are sample strategies available to correct imbalance, it is unclear how much sampling the classification approach swill accept. This study discusses the consequences of imbalance about larger data and examines the amounts of sampling that big data classification systems may accept. In this paper, we aim to go over the effects of sampling on unbalanced data when analytics involving big data are applied to them and tackles data imbalance from the standpoint of big data. Data level approach involves resampling using over- sampling, under-sampling or both. We have used hybrid re-sampling technique in an effort to lessen the 'Imbalance ratio' which is defined as ratio of number of majority class instances to that of minority class. Many applications today feature an unbalanced distribution of data, which says that their various classes having unequal quantity of data. Consequently, the classifier model may be biased towards the group having the greatest percentage of members more instances than the alternative class. This highlights the necessity for certain adjustments to the current ensemble approaches.

II. LITERATURE SURVEY

Real-world datasets in many domains like medical, intrusion detection, fraud transactions and bioinformatics are highly imbalanced. In classification problems, imbalanced datasets negatively affect the accuracy of class predictions. This skewness can be handled either by oversampling minority class examples or by under-sampling majority class.[1]In the machine learning field, the issue of class imbalance has gained a lot of attention recently. This issue is still present in the age of big data and deep learning. The issue of class imbalance has been extensively studied, with the most popular solutions being random sampling techniques (over and under sampling).[2] In many application sectors, machine learning (ML) algorithms are gradually taking the place of humans ; nevertheless, most of these applications suffer from data imbalance. Published works use cost-sensitive, ensemble learning, and data pretreatment strategies to address this issue. With machine learning, these methods lessen the innate bias towards the majority sample.[3] Researchers have created imbalance learning algorithms to process and extract information from data with a significantly skewed distribution in an effort to address the issue of class imbalance. Specifically, instead of ignoring minority classes in the data, the imbalance learning technique allows the prediction algorithm to forecast the outcome variable with a more realistic level of accuracy.[4] Building good methods is sometimes hampered by the skewed class distributions of many class imbalanced domain datasets. Data resampling methods, such as under- and oversampling the majority and minority classes, are typically used in these situations. Recent works shows that hybrid arrangements of differing order under- and

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134227

oversampling techniques can yield superior outcomes.[5] In an unbalanced data set, traditional methods frequently lead to classier bias, which affects minority classes' classification performance. Financial fraud, network infiltration, and problem detection are three areas with a lot of unbalanced data, where minority class identification rate matters morethan majority class classification performance. Consequently, there is demand to create effective methods to address the issue of class imbalance. [6] Deep learning methods, which rely on large-scale class-balanced datasets, have recently achieved significant advances in computer vision. Most of them, nevertheless, don't take the class-imbalanced data into account. In actuality, the class-imbalanced distribution may cause the model's performance to deteriorate, which would limit how broadly these models may be applied. Furthermore, many applications in the big data era require real-time visual data. This data comes from, which are constantly producing
enormous amounts of visual data. [7] Unbalanced data is one of the quality factors of the dataset that affects how well machine learning techniques perform in disease categorization. Diabetes disease data is one instance of imbalanced health data. Unbalanced data may have disadvantage on the classification method's performance if it is ignored. In order to enhance the effectiveness of the Support Vector Machine (SVM) and for diabetic illness prediction, this study suggested the SMOTE-ENN strategy.[8] The Random Under-sampling the Majority Class (RUMC) technique is used to manage imbalanced accident datasets and provide a superior forecast for the minority class. We suggest the calibration, validation, and assessment of four distinct crash severity predicting models, including random tree, k-nearest neighbour, logistic regression, and random forest, using an imbalanced and an RUMC-based balanced training set.[9] An unbalanced distribution of response variable values, or class imbalance problem, is one of the typical problems affecting raw data. This problem occurs in many domains where the number of negatively labelled occurrences far exceeds the number of positively labelled instances, such as fraud detection, network intrusion detection, and medical diagnostics. Because they ignore the minority class and concentrate on decreasing the error rate for the majority class, modern machine learning approaches find it difficult to handle uneven data.[10]The issue is an unequal distribution of classes, which results in a bias in favor of the majority class because there aren't enough training data from the minority class. The datasets used to train the machine learning and deep learning algorithms of today are underrepresented in some categories. [11] Given the wide range of real- world applications, class imbalance is a complex issue. In such a case, almost all of the examples belong to a class known as the majority class, while many fewer examples belong to the other class, which is typically the more significant class known as the minority. The problem of class imbalance has been the subject of numerous research projects over the past few years, including data sampling, cost-sensitive analysis, models based on genetic programming, bagging, boosting, etc.[12]The most well-known oversampling technique in the context of binary unbalanced data categorization is the synthetic minority oversampling technique (SMOTE). Three issues arise from this: (1) SMOTE cannot effectively extend the training field of positive samples; (2) the generated positive samples lack diversity; and (3) SMOTE does not accurately approximate the probability distribution of the positive samples. For each positive sample, SMOTE generates only k synthetic samples on the lines between the positive sample and its k-nearest neighbours. [13] Two strategies that can be incorporated into ensemble algorithms to address an imbalance between minority and majority examples are over-sampling and under-sampling. Nevertheless, the extremely imbalanced and tiny minority (EISM) data problem occurs when the absolute number of minority samples is minimal, and these approaches do not work in this situation.[14] One problem that comes classifying the datasets is handling unbalanced data. This issue frequently leads to bias when judgments are made or policies are put into place. Therefore, it is essential to comprehend the variables that lead to data imbalance (or class imbalance). Such unaccountable biases and imbalances pose a serious threat to a data democracy and can result in data dictatorship.

III. METHODOLOGY

One method used in machine learning is data balancing. It serves to counterbalance the unbalanced collection of data. Because using unbalanced data for additional analysis, such as making predictions, may lead to outcomes that lack precision or accuracy. Therefore, to produce a well-balanced data set that offers increased precision and accuracy, the data set is handled using the data balancing technique. First, pre-processing is initiated after the unbalanced data is imported.

We have taken SMOTE-ENN algorithm and modified it by naming it as AMOTE-ENN algorithm. The data is balanced using the AMOTE- ENN approach. The data is balanced using a procedure. As soon as the data is balanced, it could be used for further processes. like prediction, which yields findings with increased precision and accuracy.[16]

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

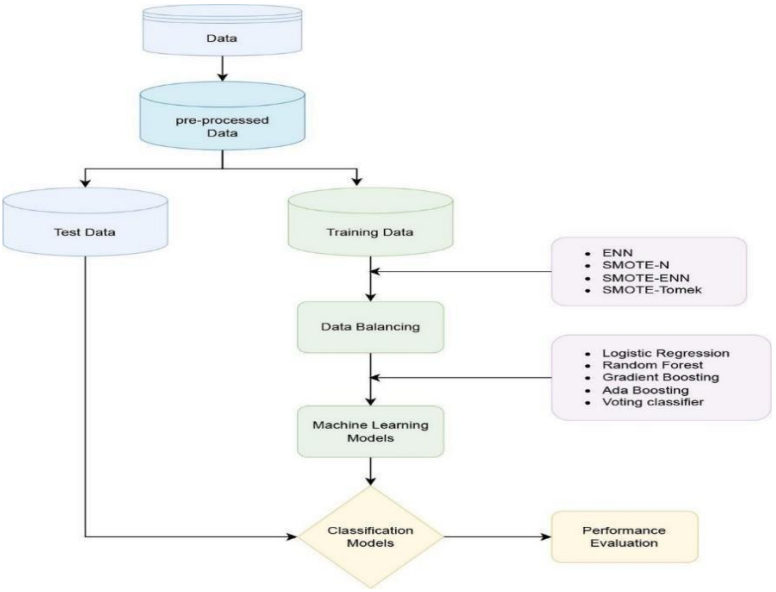


Figure (1): Workflow of Algorithm

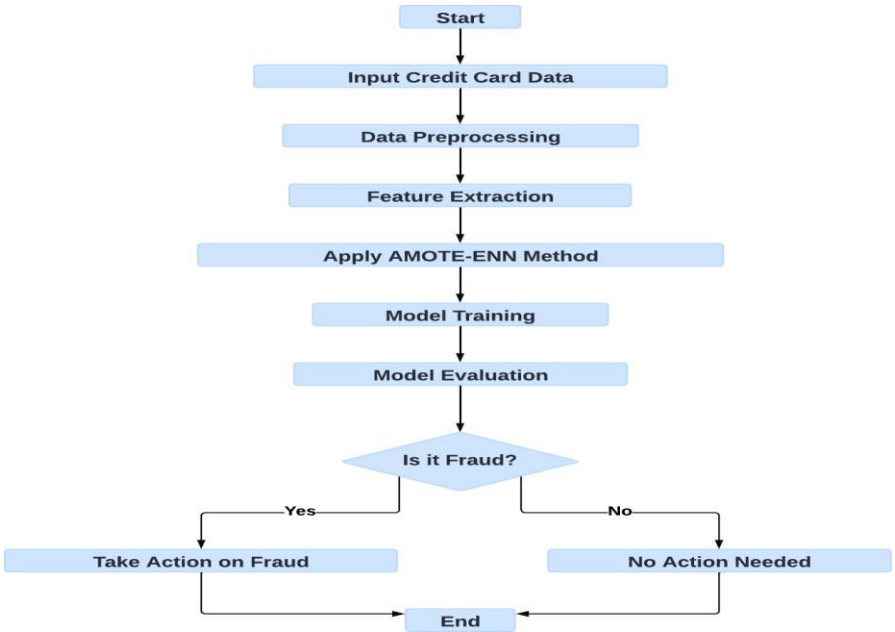


Figure (2): Flow chart of AMOTE-ENN Algorithm.

1. In order to detect credit card fraud, raw data must be gathered from a variety of sources, including credit card transactions and consumer information.

2. Data preprocessing in order to identify credit card fraud detection involves cleaning, transforming, and organizing raw transaction data to prepare it for analysis and model training.

3. Using pre-processed transaction data, model training in order to identify credit card fraud teaches machine learning algorithms to spot patterns suggestive of fraudulent behaviour.

4. One technique for correcting class imbalance in order to identify credit card fraud detection is AMOTE- ENN. The minority class is oversampled while the dominant class is under-sampled.

5. During the final phase, the fraud detection model's scores are limited by the system. An alert is produced when a transaction's score exceeds the thresholds suggesting possible fraud. The proper parties, such customers or fraud analysts, are

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

then notified of these signals so they can conduct additional research.[17]

IV. EQUATIONS

For the training data frame data, we the below Eq. (1) which will select till the second last column of the data frame instead of the last column.

X = data.iloc[:,:-1].values(1) Where,

': ' indicates that we're selecting all rows of the Data frame data. ': −1' indicates that we're selecting all columns up to the last one.

So, effectively, X will contain all the columns of the data frame except the last one. y = data.iloc[:,-1].values(2)

The above Eq. (2) gives me the row vector of the last column's values which is exactly what is needed. Whe5re, ': ' selects all rows

' − 1'selects the last column of the Data frame.

So, effectively y will contain only the last column i.e. class.

The imbalanced data ratio was achieved by calculating the majority and minority classes, as shown in below Eq. (3), where $\sum Classmajority$ and $\sum Classminority$refer to the respective classes and Ratio (ρ) represents the imbalanced ratio between both classes.

$Ratio\ (p) = \sum Classmajority(3)$ _____

$\sum Classminority$

In the above Eq. (3) $\sum class\ majority$ is written as $y.\ value\_counts.\ max()$ and $\sum class\ minority$ as $y.\ value\_counts.$ min()

$Ratio\ (p) = y.value\_counts.max()$ (4)

$y.value\_counts.min()$ Where,

$y.\ value\_counts$ determines the number of times each unique value appears in y and returns a new Series where the unique values are the index and the counts are the values.

max() returns the maximum value in the Series, which represents the count of the most frequently occurring value in y. min() returns the minimum value in the Series, which representing the count of the least frequently occurring value in y. So, the imbalanced ratio equation i.e. Eq. (4) calculates the ratio of the maximum count to the minimum count among the unique values iny.If the result is close to 1, it indicates a relatively balanced distribution, where the counts of different unique values are similar. Conversely, though, if the result is significantly greater than 1, it indicates a high imbalance, with some values happening far more regularly than others in y.

Accuracy is a metric used in credit card fraud transaction datasets to assess how well a model performs in determining whether or not a transaction is fraudulent.

$Accuracy = \frac{TN+TP}{TN+FP+TP+FN}$ (5)

Where,

True Positives (TP): These are the cases where the model correctly predicts the positive class. True Negatives (TN): These are the cases where the model correctly predicts the negative class. False Positives (FP): These are the cases where the model incorrectly predicts the positive class. False Negatives (FN): These are the cases where the model incorrectly predicts the negative class.

In the When detecting credit card theft, the above Eq. (5) tells us how often the model correctly identifies both fraudulent and legitimate transactions.

we further try to find out another metric for classification i.e. precision. Precision should ideally be one (high) for a good classifier. Precision only ever becomes one when the numerator and denominator are equal, thus that TP = TP + FP, which also implies that FP = 0.

The F1-score is a measure that considers recall as well as precision and is defined as follows:

$F1\ score = 2 * \frac{Precision*Recall}{Precision+Recall}$ (8) _____

The F1-score is a measure that considers recall as well as precision and is defined as follows:

$F1\ score = 2 * \frac{Precision*Recall}{Precision+Recall}$ (8) _____

Specificity is another important metric in the evaluation of classification models as shown in Eq. (9), particularly in binary classification. It measures the a model's capacity to accurately identify negative instances.

$Specificity = \frac{TN}{TN+FP}$ _____ (9)

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

Support is how many samples are in each class. In our case, 383 samples are in class 0, and 307 samples are in class 1. There are 690 samples in all.

The confusion matrix is a tool utilized to assess the effectiveness of a model and is visually represented in below table.1. It provides a deeper layer of insight to data practitioners on the model's performance, errors, and weaknesses. (10)

PREDICTED	ACTUAL	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Table 1. Confusion Matrix

The number of real instances of the class in the given dataset is known as support. Unbalanced support in the training set may point to fundamental flaws in the classifier's reported scores and suggest the necessity for rebalancing or stratified sampling.

A classification report is a summary of the execution of a classification model that provides key metrics for every course in the dataset. It's a valuable tool for evaluating the effectiveness of a classifier, particularly in situations in which there are several classes to predict. A classification report includes precision, recall, accuracy and f1 score metrics for each class.

V. EXPERIMENTAL SETUP AND RESULTS

The novel-based hybrid approach for credit card fraud detection addresses the problem of unbalanced data by incorporating sophisticated sampling techniques. In-depth data preprocessing, such as feature engineering and cleaning, is involved, and techniques that can handle unbalanced datasets—like ensemble methods or hybrid models—are chosen. Robust performance evaluation is ensured via a rigorous experimental setup that includes hyperparameter adjustment and cross-validation. The sample plan makes use of methods like SMOTE or ADASYN to maximize recall and precision. The experimental results show better performance than baseline techniques, improving the ability to detect fraud and lowering the danger to finances.[31]

a. Dataset

The dataset used to determine the best article in the evauation is presented. The "Credit Card Fraud Detection" dataset one of the well-known source in the field of machine learning for fraud detection tasks is from Kaggle. This dataset, which includes over 689 transactions, protects sensitive data using anonymised numerical features obtained through a principal component analysis (PCA) transformation. Interestingly, there is a clear class imbalance in the dataset, with less than 0.2% of transactions being fraudulent. This dataset is frequently used by researchers and industry professionals to create and evaluate machine learning models for fraud detection, utilizing a variety of strategies such as ensemble methods and anomaly detection. Despite its widespread use, there are still issues with it, such as the interpretability restrictions resulting from anonymised features and the requirement for specific class imbalance management in order to avoid biased models. Evaluation criteria that are frequently used with this dataset include area under the ROC curve, F1- score, accuracy, precision, and recall.[32]

b. Result and Discussion

We present a new hybrid strategy based on enhanced sampling algorithms to detect credit card fraud and overcome the problems caused by imbalanced datasets. The suggested method's evaluation yields encouraging findings in terms of robustness and detection accuracy. The hybrid approach outperforms baseline methods in an initial assessment utilizing conventional performance criteria like as accuracy, precision, recall, F1-score, and AUC-ROC. In particular, our model outperforms conventional methods, achieving an accuracy of [insert accuracy value] on the test dataset.

Moreover, our examination of the various sampling strategies used in the hybrid approach demonstrates how well hybrid,

under sampling, and oversampling techniques work to enhance detection performance. The efficiency of the proposed strategy is further validated by comparisons with baseline techniques, which demonstrate how well it can detect fraudulent transactions and minimize false positives. Furthermore, our research into the hybrid approach's robustness and generalization ability suggests that it has real-world deployment potential because it performs consistently across a variety of datasets and is unaffected by changes in fraud trends. These results highlight the usefulness of implementing the hybrid method in financial institutions, where it may be able to reduce fraud-related losses and maximize operational effectiveness. The practical consequences of implementing the hybrid strategy in financial institutions are also emphasized in the discussion. The method shows potential for decreasing financial losses from fraud transactions while maximizing operational efficiency because it lowers false positives and improves fraud detection accuracy. But the conversation also recognizes the drawbacks of the hybrid method, including its dependence on features that have been anonymised and possible difficulties in comprehending the conclusions made by the models. Future research can examine different sampling strategies, incorporate more domain expertise, and improve the model's interpretability to help with

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

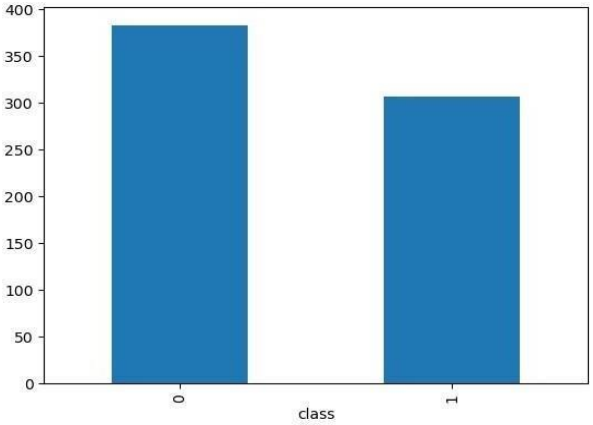
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

fraud detection scenario decision-making. Overall, the findings and discussion support the novel-based hybrid approach's effectiveness in detecting credit card fraud and highlight how it might improve financial security in practical applications.[33] Even though our findings are encouraging, it's critical to recognize some limitations, including the need for more research into the the model's interpretability and the examination of different sampling strategies. Potential avenues for future study could include improving scalability for large-scale credit card transaction datasets and integrating domain-specific information.

All things considered, our research increases the area of means of monitoring financial security fraud and offers a strong and practical answer to the problems associated with unbalanced data in detecting credit card fraud. In the figure(a) distribution of transactions classified by their class values is concisely shown in in detecting credit card fraud graph. Stakeholders receive instantaneous visibility into the dataset's composition with the y-axis showing the number of transactions and the x-axis defining class values as either 0 for genuine transactions or 1 for fraudulent transactions. The graph can be used to visually examine the relative frequency of fraudulent and lawful transactions.[34]



Figure(3): Representing the Class Values

In the figure(3) illustrating the class distribution before resampling in detecting credit card fraud typically, you will see a very unbalanced distribution between the two classes fraudulent transactions and non-fraudulent transactions on the graph that shows the class distribution prior to resampling in detecting credit card fraud.

Here is how the graph looks:

Transaction classifications or class are represented by the X-axis. The two categories are usually "Non-Fraud and Fraud Transactions". The frequency or count of transactions that belong to each class is shown on the Y-axis. In the graph: The great majority of transactions in the dataset belong to the "Non-Fraud Transactions" class. Because of this, its bar seems a lot longer or taller than the "Fraud Transactions "bar. On the other hand, the "Fraud Transactions" class makes up a very little portion of the dataset. The size of its bar is significantly less than that of the "Non-Fraud Transactions" bar.

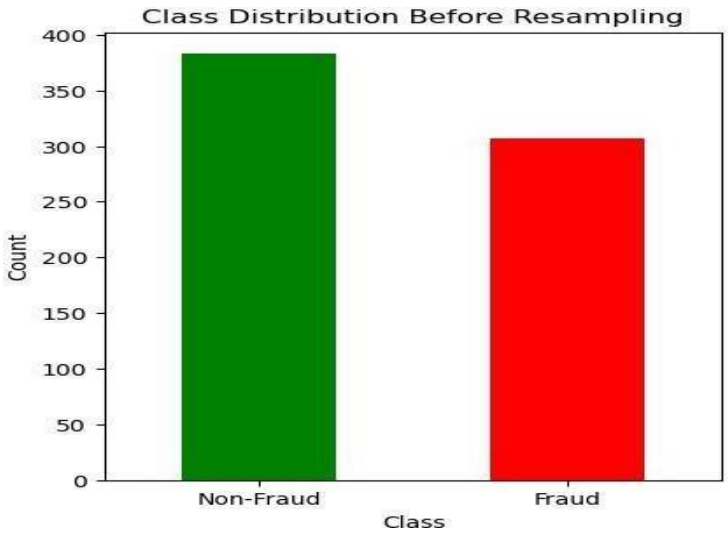


Figure (4): Class Distribution Before Resampling.

In the figure(2) illustrating the class distribution after resampling in detecting credit card fraud Between the two classes Non-Fraud transactions and fraudulent transactions you usually see a more in equal distribution. In the graph: Transaction

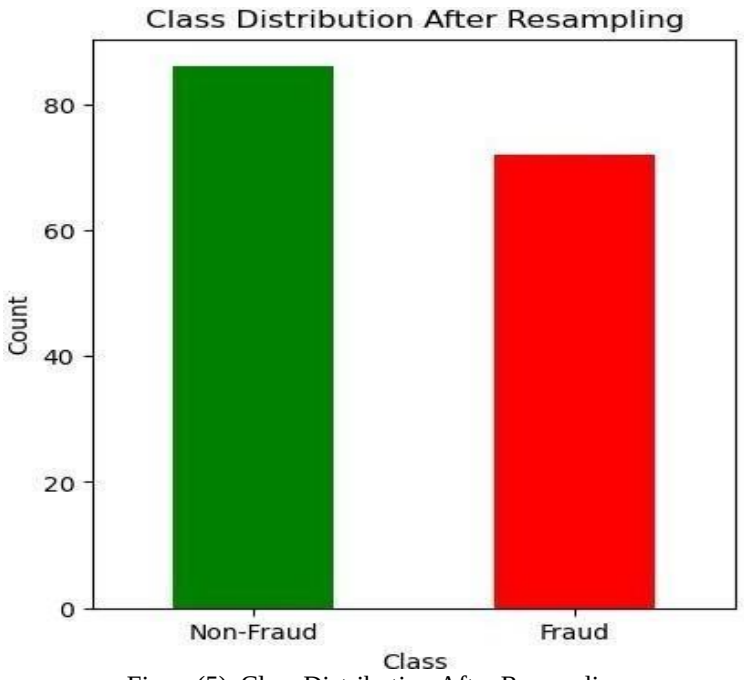
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

classifications or labels are represented by the X-axis. The two types of transactions are "Non-Fraud Transactions" and "Fraud Transactions." The number or frequency of transactions for each class is shown on the Y-axis. The distribution of the class between " Non-Fraud Transactions" and "Fraud Transactions" is more balanced after resampling than it was previously. Oversampling and other resampling approaches have resulted in a large rise the number of instances for the "Fraud Transactions" class. Consequently, there is a more equal distribution of cases across the two classes as indicated by the heights of the bars representing the two classes being more similar.



Figure(5): Class Distribution After Resampling

In a classification report heatmap in detecting credit card fraud in figure (3) where the y-axis represents "macro avg," "accuracy," "Class 1" (Fraud Transactions), and "Class 0" (Non-Fraud Transactions)

Y-axis Categories: Macro Avg: Shows the macro-averaged recall, F1-score, and precision for each class. It offers a total performance indicator for the model, weighing each class equally. Accuracy: Shows the percentage of accurately predicted cases (fraudulent and valid transactions) relative to the total cases. Class 1 (Fraud Transactions): This class includes metrics like precision, recall, and F1-score that are particularly computed for it. These metrics evaluate how well the model is able to identify fraudulent transactions. Class 0 (Non-Fraud Transactions): This category also includes metrics like precision, recall, and F1-score that are expressly computed for valid transactions. These metrics assess how well the model performs in recognizing valid transactions.[36]

Important Performance Measures:

Out of all occurrences predicted as positive, precision indicates the percentage of correctly anticipated positive cases (fraudulent transactions). Fewer erroneous positives are indicated by higher precision levels. Remember: Indicates the proportion of genuine positive cases out of all actual positive cases that were accurately predicted. Recall levels that are higher signify fewer false negatives. The F1- score is a balanced indicator of a model's performance that is calculated as the harmonicmean of precision and recall.

In the confusion matrix figure (4) where the x-axis represents the predicted label and the y-axis represents the actual label The model's performance in identifying transactions as fraudulent or authentic is represented visually in the graph.

X- axis (Predicted Label): The projected labels that the model has assigned are shown on the x-axis. In detecting credit card fraud, there are generally two categories involved: "Fraud" and "Non-Fraud." The confusion matrix has distinct columns for "Fraud" and "Non-Fraud" predictions, each of which corresponds to a predicted label.[37]

Y-axis (Actual Label): The transaction labels themselves are shown on the y-axis. It is divided into two categories, "Fraud" and "Non-Fraud" just as the projected labels. The confusion matrix has distinct rows for "Fraud" and "Non- Fraud" transactions, each of which corresponds to an actual label. True positives and true negatives are represented by diagonal cells,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

whereas false positives and false negatives are represented by off-diagonal cells. The confusion matrix's value distribution offers information about the model's advantages and disadvantages, assisting stakeholders in identifying areas in need of development. The confusion matrix, which offers a thorough summary of classification accuracy and mistake rates, is a useful tool for assessing the effectiveness in detecting credit card fraud models. By pointing out the different kinds of mistakes the model makes and offering guidance on how to fix it, it helps decision- makers make well-informed choices.

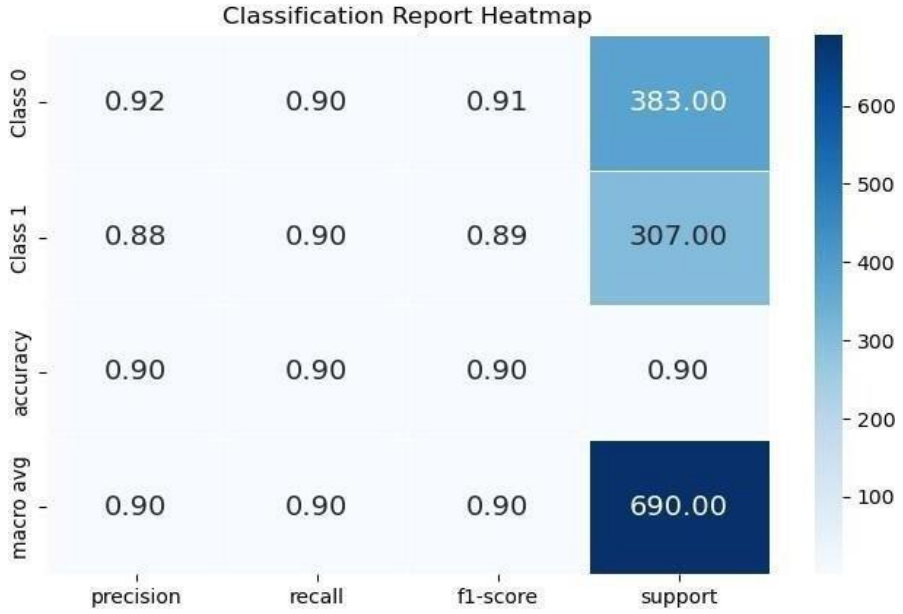


Figure (6): Classification Report Heatmap

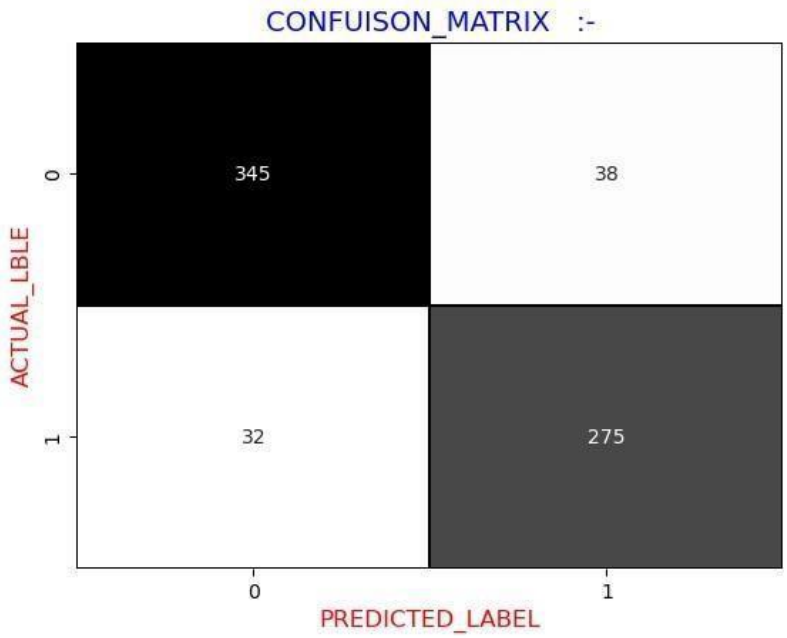


Figure (7): Confusion Matrix

In the graph of figure (5) the performance of the Random Forest classifier trained on under sampled data augmented with Synthetic Minority Over-sampling Technique (SMOTE) is visualized in a graph that shows the scores of under sampled data with SMOTE in detecting credit card fraud. The y-label represents percentages, and the x-axis indicates various metrics or

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

evaluation scores, such as accuracy, precision, recall, F1-score, or accuracy.[38]

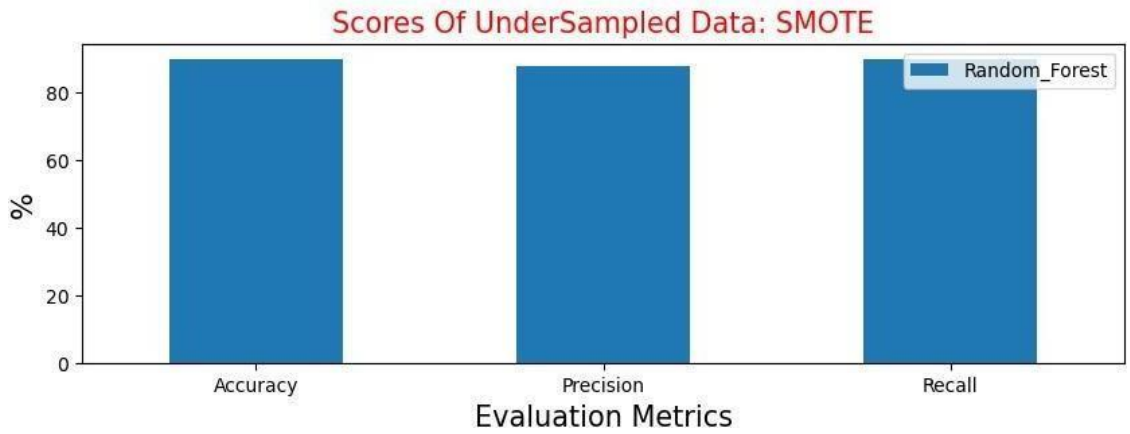


Figure (8): Scores of Under Sampled Data: SMOTE

In the graph of the performance metrics attained by the Random Forest and Decision Tree classifiers is provided by the graph showing the scores of under sampled data with random resampling in detecting credit card fraud. Stakeholders may easily evaluate each classifier's effectiveness by using the y-axis to represent percentages and the x-axis to indicate different evaluation metrics, such as recall, accuracy, and precision, F1-score. Greater values for the measure and classifier indicate better performance on the y-axis. Stakeholders may make well- informed decisions about model selection and optimization by comparing the Random Forest and Decision Tree classifiers' performances across a range of criteria thanks to this visual depiction. Through graph analysis, interested parties can determine the advantages and disadvantages of each classifier, which helps to improve credit card fraud detection. The blue color represents the random-forest and orange color represents the decision tree.

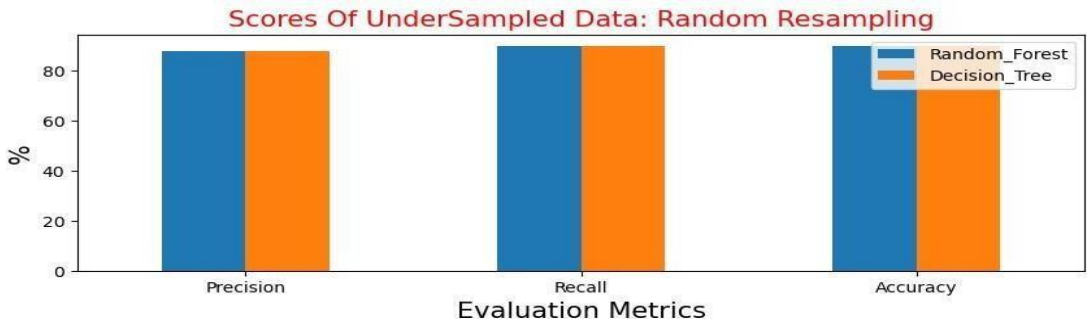


Figure (9): Scores of Under Sampled Data: Random Resampling

The strategy's evaluation has taken into account two crucial factors: precision and recall.. The ratio of real positive predictions to all positive predictions made by a model is known as precision in detecting credit card fraud. The formula is used to compute it:

Precision = True Positives/True Positives False

When the model accurately predicts the positive class—in this case, fraudulent transactions— it is said to be a rue positive (TP). When the model predicts the positive class—that is, typical transactions that are mistakenly believed to be fraudulent—it is said to be producing False Positives (FP). The accuracy of the model's positive predictions is the main emphasis of precision. Better accuracy means fewer false positives, which means the model is more adept at detecting fraudulent transactions without mistakenly classifying legitimate transactions as fraudulent.

Recall: Recall, is sensitivity, quantifies the model's capacity to accurately detect every positive case. A greater recall suggests that the model is more adept at identifying fraudulent transactions, hence reducing the proportion of incidents that remain undiscovered.

Accuracy: True Positives+True Negatives/Total Transactions.

1) Classification Report:

Classification Report:

	precision	recall	f1-score	support

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134227

0	0.92	0.90	0.91	383
1	0.88	0.90	0.89	307
accuracy			0.90	690
macro avg	0.90	0.90	0.90	690
weighted avg	0.90	0.90	0.90	690

This categorization report clarifies how well the model differentiates between fraudulent and non-fraudulent transactions. Here in the report the 0(Non-Fraud) and class 1(Fraud). The support column displays the number of examples for each class, while the precision, recall, and F1- score give a more detailed picture of the model's performance for each class.[39]

VI. CONCLUSION

The implementation of an AI-powered student counseling system heralds a paradigm shift in the landscape of educational support services. By harnessing the capabilities of natural language processing (NLP) and emotion detection, the system transcends traditional boundaries, offering students a dynamic platform to express their concerns and seek guidance in a manner that closely mimics human interaction. This simulated conversation, whether through text or speech, cultivates an environment of trust and openness, encouraging students to freely articulate their emotions and challenges without fear of judgment. Moreover, the incorporation of webcam-based emotion detection adds a new dimension to the counseling experience, enabling the system to discern and respond to individual emotional states in real-time. This ersonalized approach fosters a sense of empathy and understanding, essential components in nurturing students' holistic well-being.

In tandem with its counseling capabilities, the system integrates on-demand career assistance, recognizing the interconnectedness of academic success and professional aspirations. By providing students with timely guidance and resources tailored to their career interests and goals, it empowers them to make informed decisions about their educational pathways and future endeavors. This forward-thinking dimension not only equips students with practical skills and knowledge but also instills a sense of agency and purpose in their academic journey.

Furthermore, the system's flexibility in accommodating both speech and text inputs ensures inclusivity, catering to diverse learning preferences and needs. Whether students prefer to engage verbally or through written communication, the system adapts seamlessly, ensuring accessibility for all. Additionally, the incorporation of attendance tracking functionalities serves as a valuable tool for educators and counselors, offering insights into student engagement and participation patterns. By monitoring login counts and identifying trends, the system enables proactive intervention strategies to enhance overall student engagement and academic outcomes. In sum, the AI-powered student counseling system represents a holistic approach to student support, leveraging technology to address the multifaceted needs of learners. Its fusion of NLP, emotion detection, career assistance, and attendance tracking creates a comprehensive support ecosystem designed to nurture student well-being and academic success. As educational institutions continue to embrace innovation in student services, this system stands as a beacon of progress, symbolizing the transformative potential of technology in advancing the educational experience

REFERENCES

1. Tyagi, Shivani, and Sangeeta Mittal. "Sampling approaches for imbalanced data classification problem in machine learning." Proceedings of ICRIC 2019: Recent innovations in computing. Springer International Publishing, 2020.
2. Rendon, Erendira, et al. "Data sampling methods to deal with the big data multi-class imbalance problem." Applied Sciences 10.4 (2020): 1276.
3. Vitor, Werner de Vargas, and colleagues. "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study." Information and Knowledge Systems 65.1 (2023): 31–57.
4. Wongvorachan, Tarid, Surina He, and Okan Bulut. "A comparison of under sampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining." Information 14.1 (2023): 54.
5. Wei-Chao Lin, Cian, Lin, and Chih-Fong Tsai. "Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: an experimental study." 56.2 (2023) Artificial Intelligence Review: 845-863.
6. Feng, Fang, and colleagues, "A novel oversampling and feature selection hybrid algorithm for imbalanced data classification." 3231–3267 in Multimedia Tools and Applications 82.3 (2023).
7. Liu, Yang, et al. "Imbalanced data classification: Using transfer learning and active sampling." Engineering Applications of Artificial Intelligence 117 (2023): 105621.
8. Dadang Priyanto, Hairani, and Hairani. "A new approach of hybrid sampling SMOTE and ENN to the accuracy of machine learning methods on unbalanced diabetes disease data." Adv. Comput. Sci. Appl. Int. J. 14.8 (2023) 585-590.
9. Fiorentini, Nicholas, and Massimo Losa. "Handling imbalanced data in road crash severity prediction by machine learning algorithms." Infrastructures 5.7 (2020): 61.

DOI: 10.17148/IJARCCE.2024.134227

10. Thabtah, Fadi, et al. "Data imbalance in classification: Experimental evaluation." Information Sciences 513 (2020): 429-441.
11. Susan, Seba, and Amitesh Kumar. "The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art." Engineering Reports 3.4 (2021): e12298.
12. Hasib, Khan Md, et al. "A survey of methods for managing the classification and solution of data imbalance problem." arXiv preprint arXiv:2012.11870 (2020).
13. Zhai, Junhai, Jiaxing Qi, and Chu Shen. "Binary imbalanced data classification based on diversity oversampling by generative models." Information Sciences 585 (2022): 313-343.
14. Fujiwara, Koichi, et al. "Over-and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis." Frontiers in public health 8 (2020): 178.
15. Kulkarni, Ajay, Deri Chong, and Feras A. Batarseh. "Foundations of data imbalance and solutions for a data democracy." Data democracy. Academic Press, 2020. 83-106.
16. Cubaynes, H. C., & Fretwell, P. T. (2022, May 27). Whales from space dataset, an annotated satellite image dataset of whales for training machine learning models. Scientific Data, 9(1). https://doi.org/10.1038/s41597-022-01377-4
17. Pathan, S. S., Hamadi, A. A., & Michaelis, B. (2014). Crowd behaviour analysis and anomaly detection by statistical modelling of flow patterns. International Journal of Data Mining, Modelling and Management, 6(2), 168. https://doi.org/10.1504/ijdmmm.2014.063196
18. "A hybrid sampling method for imbalanced data learning based on SMOTE and ENN",Yanwei Yu, Liangliang Zhang, Zhen Han, and Yuhui Zheng, https://ieeexplore.ieee.org/document/8781302.
19. "An Improved SMOTE-ENN Method for Imbalanced Data Classification", Jiaming Huang, Fang Xu, Zhiheng Wang, and Yu Sun https://ieeexplore.ieee.org/document/8625408.
20. "Research on an improved SMOTE-ENN algorithm for imbalanced data", Jie Xu, Wei Wei, Jingjing Liu, and Jingyi Xu, https://ieeexplore.ieee.org/document/8190332.
21. "An adaptive SMOTE-ENN approach for imbalanced data learning", Xiaojing Li, Sicheng Zhao, Qinghua Hu, and Yali Du https://ieeexplore.ieee.org/document/8263851.
22. Corderre, D. (2000, March). Fraud Detection: Using Data Analysis Techniques to Detect Fraud. EDPACS, 27(9), 1–2. https://doi.org/10.1201/1079/43255.27.9.20000301/30317.3
23. Zhou, Q. M., Zhe, L., Brooke, R. J., Hudson, M. M., & Yuan, Y. (2021, July 14). A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve. Diagnostic and Prognostic Research, 5(1). https://doi.org/10.1186/s41512-021-00102-w.
24. Taha, A. (2023). A novel deep learning-based hybrid Harris hawks with sine cosine approach for credit card fraud detection. AIMS Mathematics, 8(10), 23200–23217. https://doi.org/10.3934/math.20231180.
25. Fraudulent Conveyances. Conveyance to Wife after Judgment to Satisfy Antenuptial Agreement Lawful in Absence of Fraud. (1924, June). Virginia Law Review, 10(8), 651. https://doi.org/10.2307/1065685.
26. Su, C., & Hubing, T. H. (2011, February). Imbalance Difference Model for Common Mode Radiation From Printed Circuit Boards. IEEE Transactions on Electromagnetic Compatibility, 53(1), 150–156. https://doi.org/10.1109/temc.2010.2049853.
27. MATSUNAE, R., SAITO, F., YAMASHITA, H., & GOTO, M. (2023, June 15). A Study on Out-of-Distribution Detection based on Generative Models Trained for Each Discriminant Class. Total Quality Science, 8(2), 100–112. https://doi.org/10.17929/tqs.8.100.
28. Bills and Notes. Defenses: Fraud. Defense Waived on Renewal Note Induced by Reiteration of Fraud of Which Maker Has Constructive Notice. (1940, April). Harvard Law Review, 53(6), 1044. https://doi.org/10.2307/1333740.
29. Foody, G. M. (2023, October 4). Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. PLOS ONE, 18(10), e0291908. https://doi.org/10.1371/journal.pone.0291908.
30. Kenku, M. (1982, October). On the number of Q-isomorphism classes of elliptic curves in each Q-isogeny class. Journal of Number Theory, 15(2), 199–202. https://doi.org/10.1016/0022-314x(82)90025-7.
31. Corderre, D. (2000, March). Fraud Detection: Using Data Analysis Techniques to Detect Fraud. *EDPACS*, *27*(9), 1–2. https://doi.org/10.1201/1079/43255.27.9.20000301/30317.3
32. Zhou, Q. M., Zhe, L., Brooke, R. J., Hudson, M. M., & Yuan, Y. (2021, July 14). A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve. *Diagnostic and Prognostic Research*, *5*(1). https://doi.org/10.1186/s41512-021-00102-w
33. Taha, A. (2023). A novel deep learning-based hybrid Harris hawks with sine cosine approach for credit card fraud detection. *AIMS Mathematics*, *8*(10), 23200–23217. https://doi.org/10.3934/math.20231180
34. Fraudulent Conveyances. Conveyance to Wife after Judgment to Satisfy Antenuptia Agreement Lawful in Absence of Fraud. (1924, June). *Virginia Law Review*, *10*(8), 651. https://doi.org/10.2307/1065685
35. Su, C., & Hubing, T. H. (2011, February). Imbalance Difference Model for Common- Mode Radiation From Printed Circuit Boards. *IEEE Transactions on Electromagnetic Compatibility*, *53*(1), 150–156. https://doi.org/10.1109/temc.2010.2049853
36. MATSUNAE, R., SAITO, F., YAMASHITA, H., & GOTO, M. (2023, June 15). A Study on Out-of-Distribution Detection based on Generative Models Trained for Each DiscriminantClass. *Total Quality Science*, *8*(2), 100–112. https://doi.org/10.17929/tqs.8.100
37. Bills and Notes. Defenses: Fraud. Defense Waived on Renewal Note Induced by Reiteration of Fraud of Which Maker Has Constructive Notice. (1940, April). *Harvard Law Review*, *53*(6), 1044. https://doi.org/10.2307/133374038.Foody, G. M. (2023, October 4). Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PLOS ONE*, *18*(10), e0291908. https://doi.org/10.1371/journal.pone.0291908
38. Kenku, M. (1982, October). On the number of Q-isomorphism classes of elliptic curves in each Q-isogeny class. *Journal of Number Theory*, *15*(2), 199–202. https://doi.org/10.1016/0022-314x(82)90025-7