Group 30: Fraud Detection with Graph Databases and Machine Learning

Ahmet Kapkic
School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ
akapkic@asu.edu

Pulkit Kulshrestha
School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ
pkulshr1@asu.edu

Karthik Murugesan
School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ
kmuruge1@asu.edu

Andrew Murwin
School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ
amurwin@asu.edu

Sneha Kurunthil Veedu
School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ
skurunt1@asu.edu

***Abstract*—As the world grows accustomed to online based transactions, the number of fraudulent transactions continues to increase, leading to substantial losses in the financial industry. Although credit card fraud is not the only crime in the industry, the lack of a consistent, accurate method of detection has led to it becoming one of the most prevalent and lucrative crimes. As such, it also positions itself as one of the most important issues to resolve. Traditional research has focused on perfecting detection techniques using traditional machine learning techniques. More recently, additional techniques including the use of derived graph data and graph-based neural networks have begun being tested. Each approach however, is only tested on a single dataset. This project will attempt to generalize the techniques of data mining and graph based-property derivation to create a method that can be used on multiple fraud-related, graph-based datasets. This will be done through the attribute-based preprocessing, graph and statistical based feature extraction, developing multiple models, and model evaluation based on multiple minority class metrics.**

***Index Terms*—fraud detection, data mining, machine learning, graph databases, graph structures, classification, BankSim, community detection, support vector machine**

I. Introduction

As we witness a swift global shift towards digitization and cashless transactions, credit card usage has been on a significant rise. The convenience of online shopping has helped people purchase items using a click of a button and hence e-payment methods have been on an ever increasing demand. However, this has also resulted in a rise in fraudulent activities, causing glaring losses to financial institutions. Credit card frauds are one of the demanding issues faced by the US populations. It is imperative that credit card companies identify and block such fraudulent transactions so that the customers do not end up paying for the items that they have not purchased.

Although the huge volume of these transactions make this task daunting. With the help of Machine learning models for fraud detection, we aim to help detect and identify such fraudulent transactions and assist to curb them.

II. Problem Statement

Credit card payments seem to be the key drivers in the online payment system. During the year 2022, US credit card transaction volume hit 3.089 trillion, and as these transactions increase, so does the chance of fraudulent transactions. It was estimated that the total card fraud losses have climbed up to 12 billion since 2021 [18]. Given the financial liability these transactions bring, the ability to employ techniques to identify and curb such fraudulent transactions is indispensable. The field of data mining helps to identify these transactions based on the normal and anomalous data, and the field of machine learning helps classify the transactions as genuine and non genuine by identifying the patterns in the data using the various algorithms. In this project we aim to identify and classify the fraudulent transactions by employing various machine learning algorithms and graph based features, and identifying the features that help recognise the fraudulent patterns. Detailed below are the steps which will be implemented in this project:

1. Background research on standard and graph based fraud detection
2. Data Collection and Selection
3. Data Preprocessing and feature extraction for use in classification models
4. Model development based on standard and graph-based features
5. Evaluation of models

III. Related Works

Currently, one of the prominent methods in fraud detection is using standard classification techniques combined with additional preprocessing [2]. For preprocessing, techniques like Principle Component Analysis are used to reduce the dimensionality of the data. Multiple models are then tested against the transformed data to classify fraudulent transactions. The paper utilizes Bayesian classifiers K2, Tree Augmented Naïve Bayes, Naïve Bayes logistics, and J48 classifiers. Analysis of the models' performance was done using precision,

recall, f-measure, and accuracy. The use of PCA to trim the data led to a minimum accuracy of 95% across the models.

Aside from standard classification techniques and models, there is currently research into the inclusion of graph and community-based techniques to improve accuracy. Alongside provided data, techniques like stable group detection, k-step relation, and similarity metrics have been shown to improve performance in predictive models [3].

Additionally, research is being conducted into the use of graph neural networks due to their inherent ability to quickly process and work with graph-based data like transactions [4]. Current focus regarding the topic is on making graph-based models more resilient to camouflaged fraudsters, who attempt to use multiple transactions to take advantage of the graph and avoid detection.

Also, there are also unique fault detection approaches that use different methods, extraction of temporal patterns and creating a forecasting model [5], being one of them. Even though some of these methods are used in forecasting, they are able to provide significant insights into a ranking of variates.

While most previous works involved using supervised data, unsupervised and hybrid techniques have also gained traction. This is due to the need for constantly evolving detection algorithms since as techniques for avoiding detection are recognized and dealt with, new hybrid ones pop up. Unsupervised learning techniques have been used to augment existing methods to separate out outlier samples as fraudulent cases [6]. Although there were varying degrees of success in the dataset used, the process itself has potential.

Utilization of these methods alongside graphs can help us characterize authentic and fraudulent transactions, which can be used to realize a fault detection/transaction classification with regard to any given transaction's characteristics. Our research and development plan section will utilize these methods and algorithms to create a baseline methodology.

IV. Dataset

For this project, the Banksim dataset is used [1]. It is a collection of 594,643 transactions, out of which 7,200 are of fraudulent nature. It has the following fields.

- Customer: Contains the customer id of the customer.
- Age: Contains the age of the credit card.
- Gender: Contains the gender of the customer.
- Merchant : Contains the merchant id , place where card is used
- Category: It contains the category of the transaction. for example fashion, transport etc.
- Amount: It contains the transaction amount.
- Fraud: Whether or not the transaction was fraudulent

V. System Architecture and Algorithms

The architecture of the system we have developed is depicted in Fig 1. The raw Banksim dataset goes through 4 steps which are described in detail below:

Basic Preprocessing : Unnecessary features such as zip code are removed as the value is the same for all transactions in the Banksim dataset. Also the string category names of all category features are converted to integers.

Build Graph Database and Extract Features : Build a graph database using the Python package 'networkx'. In this graph a node will be a merchant or a customer, and an edge between 2 nodes represents a transaction between them. After constructing the graph the following features are extracted for each record for both merchant and customer.

1. Degree Centrality (DC): Fraction of nodes it is connected to.
2. Betweenness Centrality (BC): Sum of the fraction of all-pairs shortest paths that pass through a node.
3. No. of standard deviations from mean DC: Get the mean DC of all transactions and calculate how many standard deviations each value is from the mean.
4. No. of Standard deviations from mean BC: Get the mean BC of all transactions and calculate how many standard deviations each value is from the mean.

Before building the graph, the dataset was split into train/test portions. The graph was then built using only the training data, excluding the fraud column. This was to ensure the graph is only built with "known" data, and prevent data leakage, both from the testing set and from the classification information. The existing graph was then used to add the betweenness and degree centralities to the testing data, using the existing customer and merchant ids as keys, and filling in any missing values with zero.

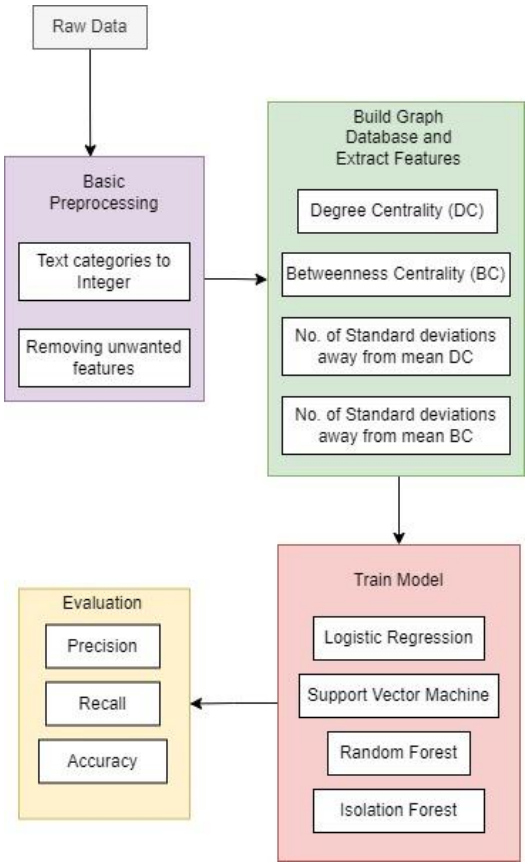

Fig 1. System Architecture

Train Model : After extracting the graph based features, train different machine learning models to classify each record as fraudulent or not. We experimented with Logistic Regression [7], Support Vector Machines [8], Random Forest [9], Decision Trees [10] and Isolation forest [11].

Evaluation : Evaluate the performance of each model by calculating the precision, recall, F1 score and false positive rate.

VI. Evaluations

We have compared five different classification methods to provide empirical results on fraud detection. The models are evaluated on only the fraudulent data, to ensure we detect faults on an imbalanced dataset. There are some experiment choices taken to ensure the results are comparable and provide a fair comparison:

- The models are not hyperparameter tuned. This will result in potentially worse results, but it helps to compare the models fairly. These results may be improved in a further study using hyperparameter tuning.
- The dataset has the same split across the models to ensure the models have the same data and the results are replicable.

The results are presented in Table I.

TABLE I
Work Schedule

	Precision	**Recall**	**F1 Score**	**FP Rate**
Logistic Regression	85.73%	66.45%	74.87%	0.13%
Decision Tree	72.94%	72.11%	72.25%	0.32%
Random Forest	92.73%	69.33%	79.34%	0.06%
Isolation Forest	6.45%	100%	12.13%	17.40%
Support Vector Classifier	92.81%	62.19%	74.48%	0.06%

Random Forest Regression provides the best F1 score, with minimal False Positive Rate with a high precision value. Support Vector Classification (SVC) and Logistic Regression has similar, but not better results to Random Forest. Isolation Forest has a 100% Recall value, but has very low precision. This shows that the Isolation Forest was able to predict all fraudulent cases, although it also falsely flags a lot of legitimate cases.

Other than Random Forest, we can observe with a detailed perspective that Logistic Regression and SVC suffer a lot from detecting False Negatives, but they are doing well on avoiding False Positives.

Even though its results are not the best, Decision Tree has the most consistent results with very close Precision and Recall scores. This shows that the choices made are balanced-even on errors, it will benefit the most from any improvements on the dataset regardless of the FP/FN focus.

Isolation Forest is a unique case, achieving 100% Recall means that it can detect all the fault cases, but it fails predicting non-fault cases by a huge margin. Even though it looks good on paper if looked at only the Recall and True Positive values, on a practical system such a system will most likely guess 90% of the transactions as fraudulent. This especially shows the importance of using different metrics to compare model performances.

Support Vector Classification does the opposite of IF, although on a smaller scale. High precision means it's able to detect non-fraudulent cases well, but the lowest recall shows that it also passes a lot of fraudulent cases as legitimate. This is a "genuinely happy" system, compared to "trigger-happy"ness of IF.

To determine how our generated variables affect the prediction process, Gini importance rankings for methods (1,2,3 and 5) are also included below in Fig 2., Isolation Forest (4) is not included as the inherent randomness does not provide the same Gini metrics for IF.

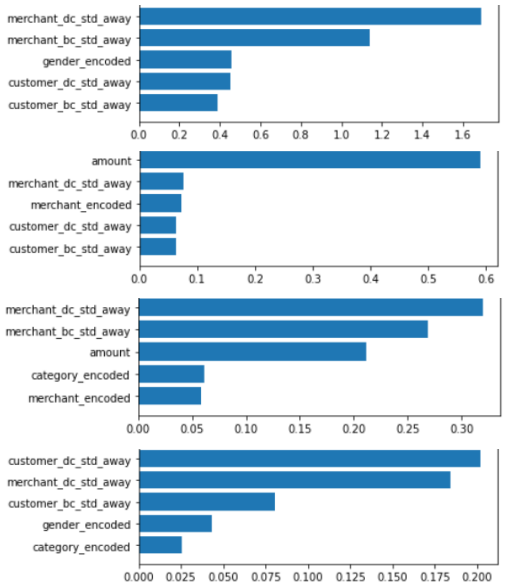

Fig 2.a-d. Variable Importance Rankings

In general, two of the graph features we have generated have the highest Gini importance values across different methods. This shows that our generated features provide more robust results. Merchant based graph features have higher importance compared to customer based ones, but both feature types still provide a significant impact on the predictions.

Recall values are low - this shows a potential for improvement by increasing recall. Understanding that a low recall value is caused by a high number of False Negatives provides a target to focus on. There are several ways to achieve this, doing a post-exploratory data analysis, or introducing new features focusing on detecting false negatives into the dataset are some approaches that can be followed.

An ideal solution would be in a single round of prediction, doing data exploration, and introducing new features that emphasize their fraud values by utilizing other variables in the X axis. If we are sure of such an imbalance, we can also utilize class weights or increase the prediction threshold lines, although it's important to note that these will mitigate the problem, decreasing False Negatives at the cost of increasing the False Positive rate.

VII. UI/Visualization Interface Design

To put the performance of our models in perspective, it is necessary to discuss the data distribution in terms of fraudulent data and non-fraudulent data. The data provided to us contains 594643 records. Out of these records, 587443 are non fraudulent records whereas 7200 are fraudulent records. We analyzed the records on various parameters.

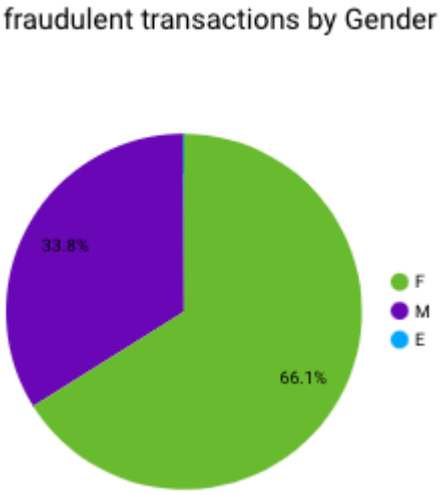

Fig 3. Fraudulent transactions by gender

The pie chart in Figure 3 visualizes the distribution of the 7200 fraudulent records from the dataset. In the full dataset, female transactions made up 54.5% of the records. While the distribution would be expected to persist when looking at the fraudulent data specifically, it was discovered that 66% of the transactions were performed by a female. Given this distribution ratio, females were more likely than males to have a fraudulent transaction committed under their name. The other gender label of interest, 'E', was also only fraudulent 0.006% of transactions, when the average across all records was 0.012%, suggesting that transactions with minority labels are less susceptible to fraud.

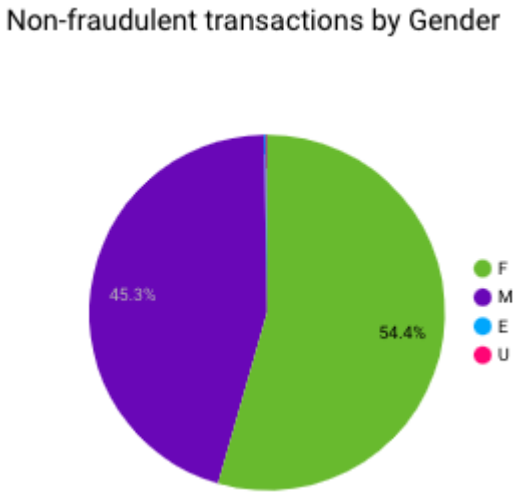

Fig 4. Non - Fraudulent transactions by gender

As the counterpart to Figure 3, Figure 4 shows the distribution of legitimate transactions. For both of the majority labels, the ratio of legitimate transactions to total transactions is approximately one to one, which is expected due to the significant majority of transactions being legitimate. In this case, the minority labels provide significantly more insight. As the inverse of its Figure 3 results, the label 'E' had a 99.4% legitimacy rating, 0.6% higher than the average legitimacy of 98.7%. Meanwhile, the label 'U' showed a 100% legitimacy rating for its records, further cementing the idea that both majority and minority gender labels can be used as indicators.

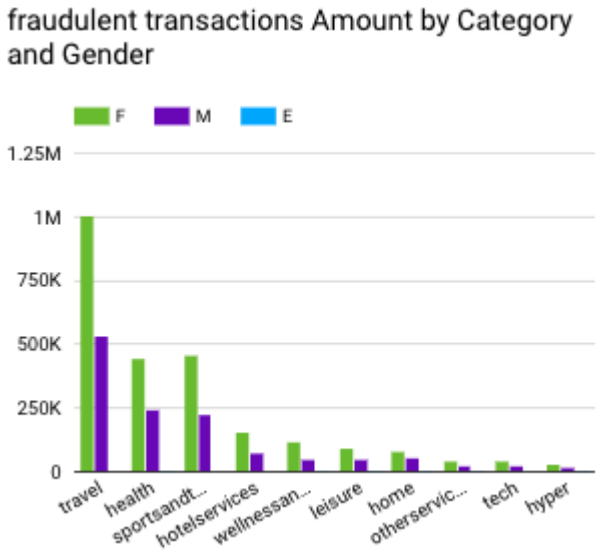

Fig 5. Fraudulent transactions amount by gender and category

While gender itself provides insight into some of the indicators of fraudulent transactions, it is not the only feature that shows indicators of fraudulent activity. Independent examination of transaction categories also showed promise in providing insight. Through a combined investigation of the two labels, further information can be derived. Examining Figure 5 shows that, across all categories, the amount spent by males is usually 50% of the amount spent by females, which fits the expected 2:1 ratio. On the other hand, when we look at the legitimate transactions in Figure 6, the amount spent by males is usually 70%-80% of the total amount spent by females in all categories. This also correlates to the 6:5 ratio seen in Figure 4. While these results show the lack of correlation between gender and category, the categorical data itself provides insight into its own trends.

Travel and transportation industries are heavy cash flow oriented businesses, meaning that a significant amount of financial transactions occur in the industry. When we look at Figure 5 and Figure 6, we can see the travel category is involved more in fraudulent transactions while transportation is involved in non fraudulent transactions. When examining the travel transactions, we see that while it is the most populous fraudulent category, the number of legitimate transactions is so small that it did not even fit in the top 10 categories shown on Figure 6. Meanwhile, fraudulent transportation transactions did not show on Figure 5 due to the fact that there were none of them. All transportation transactions were legitimate. Fraudulent transactions are unlikely to occur in the

transportation industry because it is focused on goods and physical evidence is critical, reducing the likelihood of fraud.

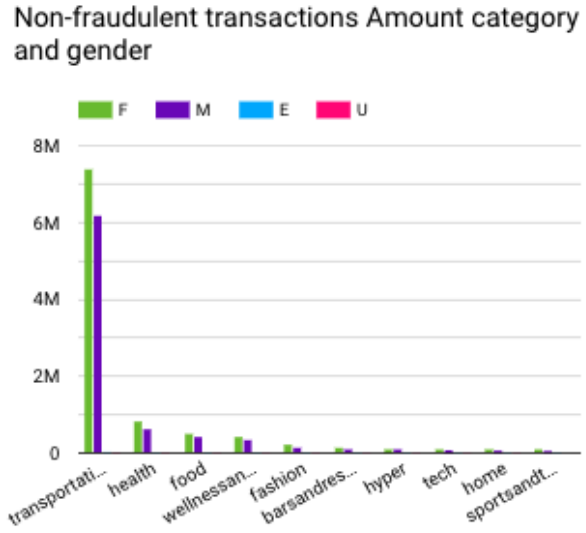

Fig 6. Non-fraudulent transactions amount by gender and category

The second category in both Figures 5 and 6 is health. We can conclude that health is an important category for transactions, both fraudulent and legitimate. Out of the health category transactions, legitimate transactions make up $1.5 million in transactions, while fraudulent ones make up around $700 thousand. Given that the overall fraud rate is around 1.5%, this quantity indicates a significantly higher fraud rate for the category, which turns out to be true at around 10%. In this category, females are more susceptible to being involved in high value fraudulent transactions than males. These large quantities make this category very exciting for scammers.

Interestingly if we look at category sports and toys, the cost of fraudulent transactions is at par with the health category at $675 thousand, but for legitimate transactions the spending is only $175 thousand. Like health, this split makes the category very prone to fraud, as 80% of the transactions are fraudulent.

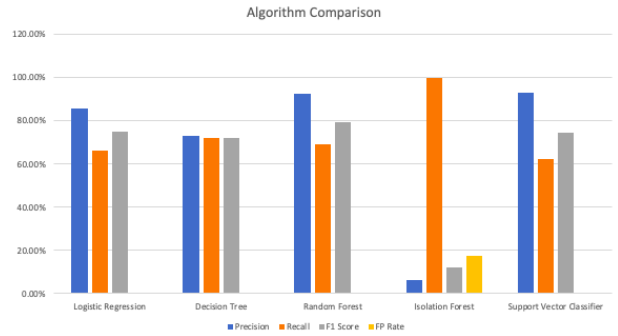

Fig 7. Comparison of Precision, Recall, F1 Score and FP Rate for different algorithms.

Figure 7 provides further insight into the differences between models, as well as how we decided our "optimal model". While a table to percentages can be difficult to balance, graph visualizations have the advantage of providing visual comparisons, allowing the person using the models to quickly differentiate and decide what metrics are a priority. By rejecting the Isolation Forest, which is dysfunctional due to its poor false positive rate, and focusing on the well-performing models, the benefits of each other model show. Random Forest and SVC work best for high precision with decent recall, while Decision Trees have the best overall balance of evaluation metrics.

VII. Division of Work and Team Member Contributions

A work schedule is created with regard to the project timeline and the proposed evaluation plan. The timeline reflects a division of work and who is responsible for the respective tasks.

TABLE II
Work Schedule

Name	Task	Proposed Deadline
All	Background research : Literature survey of the current practices used for fraud detection and the type of data they use.	Feb 15, 2023
All	Data Collection : Search and collect different datasets that could be used for the purpose of fraud detection	Feb 22, 2023
Pulkit	Data Preprocessing : Preprocess each collected dataset based on the features present and the requirements of the models that will be developed in the next phase.	Mar 8, 2023
Karthik Sneha Drew Ahmet	Model development : Explore different architectures that could be used for classification based on both the standard and graph based features of each dataset collected	Mar 29, 2023
Ahmet Pulkit	Evaluation : Evaluate the performance of each model developed based on the metrics described in the evaluation plan.	Apr 24, 2023
All	Final report preparation : Describe in detail each of the above phases as well as the result obtained from simulations in the report.	May 1, 2023

Fig 2.a-d. Variable Importance Rankings

VII. Conclusion

Like many topics, the optimal model to use ultimately comes down to the use of the focus of the model. We found that overall, each model had both desirable and undesirable traits. The Isolation Forest perfect recall was negated by its abysmal precision. The decision tree had the best balance between precision and recall, but in general performed mediocre overall. The logistic regression was a marginally worse version

of the support vector, which was marginally worse than the random forest model. In conclusion, the Random Forest model was determined to be the best available model based on its high precision and its above average recall.

Ultimately though, the preprocessing steps proved to be just as impactful as the model choice itself. Through removing negligible features, such as secondary keys, the data could be minimized and used to derive new, more impactful features. In all of the models tested, the derived betweenness and degree centrality standard deviation features were found to be high impact features that improved model performance. Through combining these features with model selection, we were able to generate models with both high precision and recall and a low false positive rate.

References

[1] BankSim, Synthetic data from a financial payment system, Kaggle, 2020 [Online], Available: https://www.kaggle.com/datasets/ealaxi/banksim1

[2] O. S. Yee, S. Sagadevan, and N. H. Ahamed Hassain Malim, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique", *JTEC*, vol. 10, no. 1-4, pp. 23–27, Jan. 2018.

[3] G. Deshpande. "Finding Needles in a Haystack with Graph Databases and Machine Learning." DZone. https://dzone.com/articles/finding-needles-in-a-haystack-with-graph-databases (accessed Feb. 14, 2023).

[4] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, "Enhancing Graph Neural Network-based Fraud Detectors against Camouflaged Fraudsters," Proceedings of the 29th ACM International Conference on Information & Knowledge Management. ACM, Oct. 19, 2020 [Online]. Available: http://dx.doi.org/10.1145/3340531.3411903

[5] Tiwaskar, M., Garg, Y., Li, X., Candan, K.S., Sapino, M.L., Selego: robust variate selection for accurate time series forecasting. *Data Min Knowl Disc* **35**, 2141–2167 (2021). https://doi.org/10.1007/s10618-021-00777-1

[6] F. Carcillo, Y. Borgne, O. Caelen, Y. Kessaci, F. Oblé,G. Bontempi, "Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection," in Information Science, vol. 557, 2021, pp. 317-331, doi: https://doi.org/10.1016/j.ins.2019.05.042.

[7] "What is logistic regression?, IBM, [Online], Available:https://www.ibm.com/topics/logistic-regression

[8] "Support Vector Machines", Wikipedia, [Online], Available:https://en.wikipedia.org/wiki/Support_vector_machine

[9] "Random Forest", Wikipedia, [Online], Available:https://en.wikipedia.org/wiki/Random_forest

[10] "Decision tree", Wikipedia, [Online], Available:https://en.wikipedia.org/wiki/Decision_tree

[11] Isolation Forest", Wikipedia, [Online], Available:https://en.wikipedia.org/wiki/Isolation_forest