

**M2 IFM**  
**2022-2023**



**UNIVERSITÉ  
DE LORRAINE**



# **Fraud Detection Classification :**

*Encadré par :*

***Marianne Clausel***

*Travail élaboré par :*

***Abdouramane Amadou Ismael***

***Ben Yahmed Ryma***

***Hannoun Zoubaida***

## Introduction :

Detecting bank fraud is extremely important for financial institutions, as fraud has a significant financial impact on both banks and their customers. Banks process millions of financial transactions every day, and detecting fraudulent activity in real-time without automated analysis is difficult.

Bank fraud can take many forms, including identity theft, credit card fraud, check fraud, electronic transfer fraud, and many others. Fraudsters often use sophisticated techniques to conceal their fraudulent activities, so it is essential for banks to implement advanced fraud detection systems to protect their customers and their own reputation.

The financial consequences of bank fraud can be huge. Banks can suffer significant financial losses if funds are diverted or if fraudulent activities are not detected in time. Additionally, bank fraud can also affect customer trust in the banking institution, which can result in a loss of customers and a decrease in the bank's reputation.

In summary, detecting bank fraud is essential to protect financial institutions, their customers, and their reputation. Banks must invest in advanced fraud detection technologies to quickly identify fraudulent activities and minimize financial losses.

*Given the importance of bank fraud detection and the need for advanced technology solutions, we have chosen to focus on this topic and will explore the development of a fraud detection model using Python programming language.*

To develop our fraud detection model using Python, we will start by using a publicly available dataset from Kaggle. The database we are using is called "**Synthetic Financial Datasets For Fraud Detection**". It contains simulated financial transactions that have been used for bank fraud detection. The database consists of over one million transactions, each labeled as fraudulent or non-fraudulent.

The primary objective of this database is to provide a large dataset for training and validating bank fraud detection models. The transactions are simulated to mimic real data, but for confidentiality reasons, names and account numbers have been anonymized.

By using this database, analysts and data scientists can train fraud detection models to detect suspicious activities in financial transactions and help financial institutions prevent fraud and protect their customers' funds.

In summary, this database is a useful tool for training bank fraud detection models and helping financial institutions identify suspicious activities in financial transactions.

## Methodology :

To properly address the above problem, we will follow a well-structured methodology that consists of: First **selecting a dataset and an output variable of interest**. Once identified, we will proceed to **describe the dataset and its characteristics, as well as the manner in which it was compiled**, which will lead us to obtain a **descriptive visualization**. Finally, we will **conclude**.

- ***Selecting a dataset and an output variable of interest :***

The dataset is the "Synthetic Financial Datasets For Fraud Detection" . The output variable of interest for this dataset is the "isFraud" column, which indicates **whether a transaction is fraudulent (value of 1) or not (value of 0)**. This is the variable that we will aim to predict using the other variables in the dataset.

- ***Describe the dataset and its characteristics, as well as the manner in which it was compiled:***

The applied problem involves analyzing financial transaction data to detect fraudulent activities : Our dataset is a synthetic dataset that simulates mobile money transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country. The dataset consists of over 6 million transactions, which are labeled as either fraudulent or legitimate.

The dataset includes features such as transaction type, transaction amount, account balance, old balance, new balance, and whether or not the transaction is fraudulent. The features are encoded and anonymized for privacy reasons, so their exact meanings are not provided.

It was created by a team of researchers from the University of Cagliari and the University of Milano-Bicocca, who used a combination of statistical methods and machine learning algorithms to simulate realistic financial transactions.

Overall, our dataset provides a valuable resource for detecting fraudulent transactions and developing predictive models for financial fraud detection.

***The dataset is composed of 11 columns:***

**step:** The step of the simulation, representing one hour of real time (total of 743 steps).

**type:** The type of transaction (4 possible types: CASH-IN, CASH-OUT, DEBIT, TRANSFER).

**amount:** The amount of the transaction in local currency.

**nameOrig:** The identifier of the account holder of the origin account.

**oldbalanceOrig:** The initial balance of the origin account before the transaction.

**newbalanceOrig:** The final balance of the origin account after the transaction.

**nameDest:** The identifier of the account holder of the destination account.




**oldbalanceDest:** The initial balance of the destination account before the transaction.

**newbalanceDest:** The final balance of the destination account after the transaction.

**isFraud:** Indicates if the transaction is fraudulent (1) or not (0).

**isFlaggedFraud:** Indicates if the transaction has been marked as fraudulent due to a transaction limit (1) or not (0).

***The specifics of this dataset are:***

-  The transactions involve mobile accounts, rather than traditional bank accounts.
-  The transactions are made in local currency, rather than foreign currency.
-  The transactions are simulated rather than real.

- ***Compilation of Dataset and descriptive Visualization :***

*This is the essential step of our project as it encompasses all the relevant steps leading to the compilation of our code that will allow the bank to identify fraudulent financial transactions and prevent them, of course, as well as visualizing graphs.*

## ***1- Performing an exploratory data analysis (Importation of packages + Importation of data):***

Firstly, we will proceed with a classic schema that is done in data analysis, particularly a brief statistical study of our variables, with the aim of learning more about the characteristics of our variables, which could potentially give us an idea about the preprocessing that we might face (outliers, missing values, etc.).

## ***2-Data Scrubbing (Handling of outliers + Management of missing data):***

In case of missing values in our database, and based on literature review, we decide to set a threshold of 25% of missing values to consider it as a major issue. To address this issue, we decide to use the median as a metric to fix it. Indeed, we consider that the median is a fairer metric than the mean, for example, in the sense that it gives us an idea of the first 50% of our sample, while the mean could potentially be biased by the presence of outliers.

## ***3- Summary statistics of our study data and Construction of graphs:*** See Notebook.

## ***4- Implementation of the correlation matrix:***

In the statistical analysis of this dataset, it would be relevant to create a correlation matrix as it allows us to visualize the relationships between the different variables in the dataset. Specifically, the correlation matrix can highlight linear relationships between variables, which can be useful in understanding how different variables may influence fraudulent transactions.

Additionally, the correlation matrix can help identify variables that have a strong correlation with the target variable (fraud or non-fraud), which can be useful in selecting variables to include in a fraud detection model. Finally, the correlation matrix can help detect potential issues of multicollinearity, which can affect the accuracy of fraud detection models.

## ***5- Implementation of bank fraud detection models :***

we need to know how to choose the right technique or method to proceed with our dataset. In this modelling phase, we will rely on the literature and what we saw in class. First of all, when we train an algorithm, we must always check its performance. In other words, we must check/evaluate its ability to generalize on new data. This is why we are going to divide our original data into training and test data, allowing us to evaluate the algorithm that has been trained. Thus, we will proceed to the division of the data into a train set that will be used to train our data set, and a test set that will be used to evaluate the ability of the model to generalize new data.

In other words, this will allow us to see how well the constructed model performs. We will use a method called cross validation, which evaluates machine learning models by training several machine learning models on the train subsets of the available input data, and applies them on the complementary test subset of the available input data. With this method, it will be possible to detect overfitting, for example, the failure of generalization of a trend. Its rows represent the actual classes the outcomes should have been, while the columns represent the predictions we have made. Using this table, it is easy to see which predictions are wrong. It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves. It is presented in the form of a square as seen in class during the course, so this matrix allows us to observe the number of true positives, the number of true negatives, the number of false positives, and the number of false negatives.

***One important thing is to choose the appropriate classification metric, especially when dealing with an imbalanced dataset. Using a metric such as accuracy would not be very relevant, and therefore metrics such as F1 score or recall would be more interesting to analyze.***

After validating these steps, we can propose bank fraud detection algorithms. In the context of this project, we propose **Decision Tree** and we will also try to implement a **KNN algorithm** to see a second method.



A popular method for detecting bank frauds due to their ability to model complex decisions. Decision trees are classification models that use a top-down approach to divide the population into groups based on characteristics. In the case of bank fraud detection, the characteristics can be variables such as transaction amount, transaction location, or account type.

The decision tree is constructed by choosing the variable that best divides the population into homogeneous groups. The process is repeated until the groups are as homogeneous as possible. Once the decision tree is built, it can be used to classify new observations.

Decision trees are particularly useful for bank fraud detection because they can model complex interactions between variables. For example, a decision tree can be used to detect fraudulent transactions that occur at a certain time of day, with a certain amount, and from a certain location. Decision trees are also interpretable, which means that the decisions are easily understandable and can be explained to auditors or regulators.

In summary, decision trees are a powerful tool for bank fraud detection due to their ability to model complex decisions and interpret results. Decision trees can be used to identify fraudulent transactions based on characteristics such as transaction amount, transaction location, or account type.



Is another classification algorithm that can be used for bank fraud detection. KNN is a simple yet effective approach that uses the Euclidean distance to find the K nearest samples to the input data point. The data point is then assigned to the majority class of the K nearest samples.

By using the KNN algorithm for bank fraud detection, transactions can be classified based on their similarity to past fraudulent transactions. This can help detect suspicious activity that may otherwise go unnoticed. KNN can also be used for identifying groups of similar transactions, which can help detect recurring fraud patterns.

However, the KNN algorithm has some limitations, such as its sensitivity to the size and quality of input data, as well as the distance and selection criteria of K. Additionally, KNN can be expensive in terms of computation time and memory, especially for large datasets.

In summary, the KNN algorithm can be a useful tool for bank fraud detection, especially when used in combination with other machine learning techniques. It can help detect suspicious activity and identify recurring fraud patterns, but it's important to consider its limitations and properly tune the algorithm for accurate and reliable results.



## Conclusion

---

*Fraud is a significant challenge facing the banking industry, and it is essential to develop effective strategies to combat it. The use of data analysis and machine learning techniques has emerged as a promising approach to detect and prevent fraud in banking.*

*The Synthetic Financial Datasets For Fraud Detection database is an excellent resource for developing and testing these strategies. By using advanced algorithms and techniques, it is possible to identify patterns of fraud and take appropriate measures to mitigate the risks associated with them. While there is no one-size-fits-all solution to fraud prevention, continued investment in innovative technologies and approaches can help mitigate the risks of fraud and protect the interests of banks and their customers.*

*Ultimately, the success of fraud prevention efforts will depend on a collaborative approach that involves stakeholders across the industry working together towards a common goal of protecting against fraudulent activities.*

---