**FRAUD DETECTION SYSTEM**

**CONCEPT PAPER**

**Recess Year 2**

**GROUP MEMBERS**

|  |  |
| --- | --- |
| **NAME** | **REGISTRATION NUMBER** |
| TUMUSIIME ARNOLD | 17/U/10549/EVE |
| TUSIIME ALLAN MUHANGUZI | 17/U/10630/EVE |
| MUYANJA RODNEY | 17/U/6819/EVE |
| SHAKA STEPHEN | 16/U/20180/EVE |

**CONCEPT PAPER**

The fraude\_detection dataset was obtained from a tradition data source called Kaggle.com. That is obtained from a loan laundering firm. The loans are provided with the ‘customerid’ so as to identify the customer, the system identifies these loans using the ‘systemloanid’.

The customer’s loans are given numbers ‘loannumber’ along with ‘approveddate’ so as to indicate the dates the were approved and the ‘creationid’ to show the creation date.

The loan amount is ‘loanamount’ is indicated to specify the amount of money loaned out, this money is paid back to the firm with interest ‘totaldue’ within a certain timeframe ‘termdays’.

The dataset also indicates the individual who referred the person acquiring the loan using ‘referredby’.

Finally, the dataset includes a column to shows ‘good\_bad\_flag’ column that is our target, which we’ll base on to detect the fraud ie the fake or real loan accounts.

Deployed Model

Training

Train data

Test data

­

Feature engineering

Evaluation

Our data pipeline is going to consist of the following.

1. **Data Loading:** This will be simply the importing the dataset into working environment.
2. **Data wrangling:** This will involve:

* Cleaning of data if it contains noise.
* Label Encoding of the dataset.
* Feature scaling.
* Removing the outliers.

1. **Data Visualization:** This will involve:

* Presenting our analysis data using graphs such as bar graphs, histograms, and pie charts.
* We will also determine the skewedness of data and correlation of the features.
* Determining the correlation of features

1. **Transformation ML Algorithm**: Among the many Algorithms, we shall pick one after evaluating them. These may include: Logistic and Linear Regression, Simple Vector Machine, Decision Trees, K-Nearest Neighbors, among others.
2. **Data Modelling:** This will involve creating models that correlate the data with outcomes. We will have to use our train data to enable the modelling procedure.
3. **Evaluation** **of the ML Algorithm**: Here we shall choose the Algorithm with the best score.
4. **Deployment** shall be done by a Dashboard.