

**MZUZU UNIVERSITY**

**FACULTY OF SCIENCE, INNOVATION AND TECHNOLOGY**

**DEPARTMENT OF ICT**

IMPROVING CREDIT RISK PREDICTION: BUILDING A SEMI-SUPERVISED LEARNING FRAMEWORK

DETAILED DESIGN DOCUMENT SUBMOTTED TO FACULTY OF SCIENCE INNOVATION AND TECHNOLOGY

SUBMITTED BY

JEREMIA NKOSI (BSDS2422)

MAXWELL MWALA (BSDS2322)

QUEEN SOSOLA (BSDS2822)

TAMANDANI YONA (BSDS3022)

ELIZABETH MFUNE (BSDS1922)

SAMUEL CHIZUMA (BSDS0821)

SUPERVISORMR REUBEN MOYO

Contents

[Project overview 3](#_Toc199861021)

[Description 3](#_Toc199861022)

[Problem statement 3](#_Toc199861023)

[Objective: 3](#_Toc199861024)

[Dataset overview 3](#_Toc199861025)

[Size and format: 3](#_Toc199861026)

[Description 3](#_Toc199861027)

[Features: 3](#_Toc199861028)

[Data issues: 5](#_Toc199861029)

[Data preprocessing and feature engineering 5](#_Toc199861030)

[Data pipeline design 7](#_Toc199861031)

[Model design and selection 8](#_Toc199861032)

[Modes considered 8](#_Toc199861033)

[ROC / AUC 8](#_Toc199861034)

[Selected model 9](#_Toc199861035)

[Justification 9](#_Toc199861036)

[Hyperparameters 9](#_Toc199861037)

[Evaluation metrics 9](#_Toc199861038)

[Model deployment plan 9](#_Toc199861039)

[Tools and technologies 9](#_Toc199861040)

[Risks and mitigation 9](#_Toc199861041)

[Timeline and milestones 10](#_Toc199861042)

[Appendices 11](#_Toc199861043)

[Glossary: 12](#_Toc199861044)

[References 13](#_Toc199861045)

# Project overview

Description: This project aims to explore the use of semi-supervised learning to improve credit risk prediction

Problem statement:

Supervised machine learning models rely on labeled data to make predictions. In African financial markets, labeled data is often scarce and hard to obtain. Assigning labels to this data is expensive according to Chawla and Karakoulas ­­[1]. This limits the scope of credit risk prediction using supervised machine learning models in such markets. Our project aims to explore the use of semi-supervised learning techniques namely self-training and pseudo labelling to overcome the unlabeled data challenge and improve credit risk prediction for higher education students using supervised learning models. We aim to answer the question, “can semi-supervised learning improve credit risk prediction?”.

## Objective:

To apply self-training and pseudo labeling on unlabeled data to achieve supervised classification model accuracy above 80%.

# Dataset overview

## Size and format:

The data is a 255 kilo byte comma separated value file.

Sources:

* Higher Education Students Loans and Grants Board (HESLGB)

Description:

The dataset has twelve thousand records and ten features including the target feature.

The format of the data is comma separated. It is unbalanced in terms of loan repayment status whereby more students have either fully or partially repaid their loan.

## Features:

The dataset the following features as some of the variables; gender, marital status, age range, informed consent, convene of repayment process, highest education, program of study and region.

|  |  |  |
| --- | --- | --- |
| Feature | Type | Description |
| BAge | int64 | Borrower's age in years. |
| BMaritalStatus | Object | Borrower's marital status (e.g., Married, Single, Divorced). |
| BEduLevel | Object | Borrower's education level (e.g., Undergraduate, Postgraduate). |
| Householdsize | float64 | Number of members in the borrower's household. |
| Dependents | float64 | Number of dependents the borrower supports financially. |
| BStateR | int64 | Borrower's region/state code. |
| BTimeToComp | int64 | Time (in years) the borrower took to complete their education. |
| BMajor | Object | Borrower's academic major (e.g., Bachelor of Science in Midwifery). |
| BCollege | Object | Name of the institution the borrower attended. |
| CollegeYrOfEntry | int64 | Year the borrower entered college. |
| BGradYear | int64 | Year the borrower graduated. |
| OntimeGraduation | Object | Indicates whether the borrower graduated on time (e.g., Yes, No). |
| BEmpStatus | Object | Borrower's employment status (e.g., Employed, Unemployed, Self-employed). |
| RegionUnemploymentRate | float64 | Unemployment rate in the borrower's region at the time of loan application. |
| GradRate | float64 | Graduation rate of the borrower's institution. |
| CurrentLintRate | float64 | Current loan interest rate applicable to the borrower. |
| OgLIntRate | float64 | Original loan interest rate at the time of disbursement. |
| LTerm | int64 | Loan term in years. |
| Loangraceperiod | int64 | Grace period (in years) before loan repayment begins. |
| Repayment\_status | String | Loan repayment status |
| LPaymentDate | Object | Date of the borrower's last loan payment. |
| Repaymentplan | Object | Borrower's repayment plan type (e.g., Salary, Voluntary). |
| RTime | int64 | Time (in months) borrower has been repaying the loan. |
| ColEfforts | Object | Collection efforts made by the lender (e.g., Notices, Legalities). |
| LAmount | int64 | Loan amount originally disbursed. |
| FPaymentDate | Object | Date of the borrower's first loan payment. |
| RAmnt | int64 | Total repayment amount made by the borrower to date. |
| DeliqDate | Object | Date of loan delinquency occurrence, if applicable. |
| DeliqStatus | int64 | Status of loan delinquency (1 = Delinquent, 0 = Not Delinquent). |
| lossDefault | int64 | Loss given default as a percentage. |

Table 1: Data desription

## Data issues:

* Missing values in FPymentDate, BEmStatus
* Class imbalance in Repayment\_status
* Dataset size too small for model training
* Bad formatting of column names and values

# Data preprocessing and feature engineering

|  |  |  |
| --- | --- | --- |
| Step | Method | Justification |
| Handling missing values | Mode imputation in FPaymentDate  Removing columns based on a 50% threshold of missing values | Improve model predictive accuracy |
| Formatting | Converting all column names to lower case including their values | To allow easy variable manipulation and avoid mistakes due to typing errors |
| Oversampling | Use borderlineSMOTE to mitigate the class imbalance | Improve model performance |
| Feature creation | Create age\_group from age range | Explicitly show to which group a loan applicant belongs to |
| Created feature probability of default | Replace reapayment status values (fully\_paid: low, partially\_paid: medium, unpaid: high) | Clear defined target feature |
| Created feature loss given default | LGD = EAD \* (1 - Recovery Rate) | Good predictor for credit risk |
| Encoding | Label encoding for target variable, one hot encoding for marital status, target encoding for other categorical features | Ease model training and testing |
| Data splitting | 70 – 30 split into training and testing sets | To avoid data leakage after semi-supervised learning |

Table 2: data preprocessing and feature engineering

# Data pipeline design

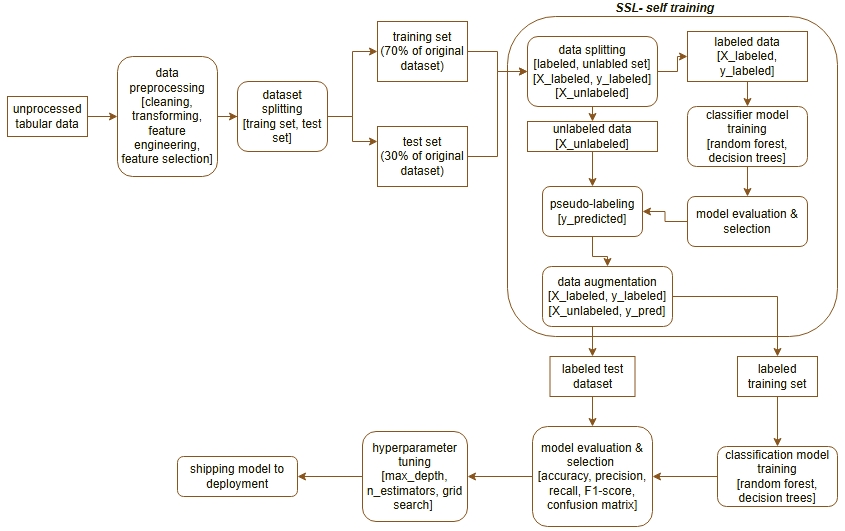


Figure 1: SSL pipeline

The pipeline will be implemented using python functions.

Steps in the pipeline

1. Load csv data file
2. Perform data preprocessing (handle missing values: dropping based on threshold of 50%, impute using mode)
3. Feature engineering
4. Feature transformation
5. Dataset splitting (into training set and testing set using the 70 – 30 split)
6. Applying semi-supervised learning for both training and testing sets separately
   1. Data splitting (into labeled and unlabeled sets)
   2. Train a classifier model using the labeled set (random forest classifier)
   3. Use the classifier model to predict target labels
   4. Augment data. Combine the initially labeled set and the latter labeled sets
7. Use the training set to train a classifier model (ensemble models, decision trees)
8. Use the test set to evaluate and validate the models
9. Select best performing model and use it to predict credit risk

# Model design and selection

Modes considered:

1. Decision trees
2. Random forest classifier
3. Artificial neural networks

Artificial neural networks were dropped from our proposed models of choice due to its computational complexity and implementation time.

ROC / AUC:

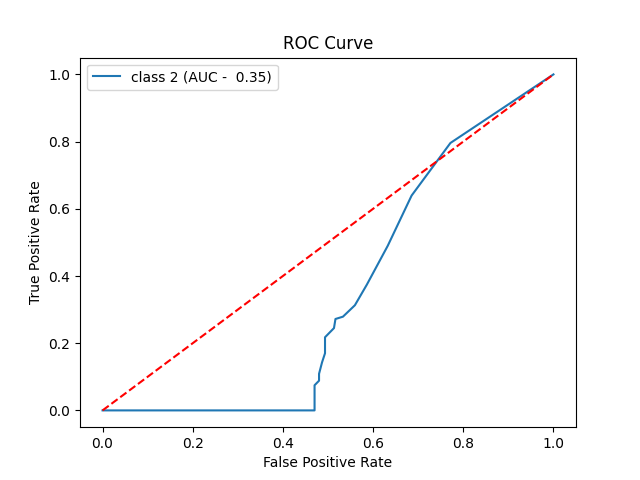


Figure 2: ROC Curve

Selected model: Random forest classifier

Justification: Better performance in terms of accuracy compared to decision trees.

Hyperparameters: n\_estimators = 120, max\_depth = 15

Evaluation metrics: Accuracy, Precision, F1 score, Recall, ROC, AUC

Model deployment plan: The final model will be saved into a joblib file and be incorporated using python’s fastAPI.

# Tools and technologies

|  |  |
| --- | --- |
| Category | Tool/library |
| Programming language | Python 3.11 |
| Libraries:  Visualization  Data preprocessing  Modeling  Evaluation  Hyperparameter tuning  Data generation | Matplotlib, seaborn  Pandas, numpy  Scikit-learn  Scikit-learn  Scikit-learn  SMOTE, sdv |
| IDE | Jupyter notebook, visual studio code |
| Version control | Git, GitHub |
| Storage | Local storage, google drive, csv files |
| Saving models and pipelines | joblib |

Table 3: tools and technologies

# Risks and mitigation

|  |  |
| --- | --- |
| Risk | Mitigation |
| Small dataset size | Synthetic data generation |
| Class imbalance | Oversampling |
| Privacy concerns | Using data stripped of names and identification |
| Categorical data | Limited scope of solutions to handle categories properly |

Table 4: risks and mitigation

# Timeline and milestones

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Start date | End date | Status |
| Data collection | 20 / 01 / 2025 | 17/01/2025 | Completed |
| Data cleaning | 10/02/2025 | 25/02/2025 | Completed |
| Feature engineering | 10/02/2025 | 25/02/2025 | Completed |
| Feature selection | 10/02/2025 | 25/02/2025 | Completed |
| Synthetic data generation | 26/02/2025 | 30/03/2025 | In progress |
| Semi-supervised learning | 05/05/2025 | 16/06/2025 | In progress |
| Model training | 05/05/2025 | 16/06/2025 | In progress |
| Design document submission | 05/05/2025 | 03/06/2025 | Due |

Table 5: timeline and milestones

# Appendices

Appendix A: confusion matrix, accuracy and classification report

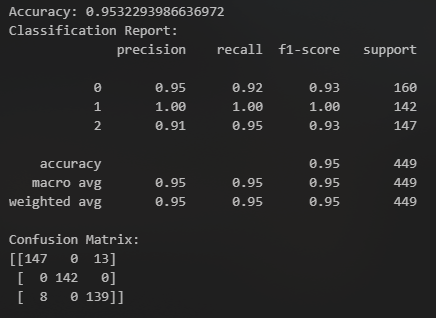


Figure 3: confusion matrix and classification report

Appendix B: data cleaning (dropping missing values based on threshold 50%)

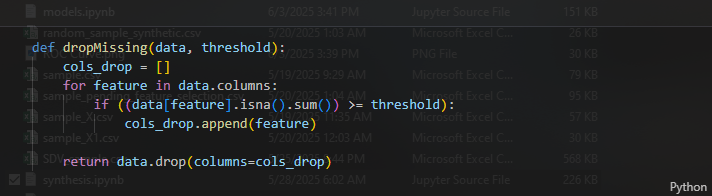


Figure 4: code snippet for dropping missing values

## Glossary:

* SSL: semi-supervised learning, a technique of making use of a small amount of labeled data and a larger amount of unlabeled data in modelling.
* SMOTE: synthetic minority oversampling
* ROC curve: Receiver operating characteristic

# References

[1]Bhateja, V. (n.d.). Information System Design : AI and ML Applications: Proceedings of Eighth International Conference on Information System Design and Intelligent Applications(2024)

[2]Langer, A. M. Analysis and Design of Next-Generation Software Architectures: 5G, IoT, Blockchain, and Quantum Computing. Germany: Springer International Publishing.(2020)

[3]Automated Machine Learning: Methods, Systems, Challenges. . Germany: Springer International Publishing.(2019)

[3]Zhou, Z. Machine Learning. Singapore: Springer Singapore.(2021)

[4]Geetha, T. V., Sendhilkumar, S. Machine Learning: Concepts, Techniques and [5]Applications. United States: CRC Press.(2023).

<https://ihsanulpro.medium.com/complete-machine-learning-project-flowchart-explained-0f55e52b9381>