Toronto Metropolitan University

The Chang School of Continuing Education

# FINAL PROJECT

## ANALYSIS OF GERMAN CREDIT CARD DATA

| | |
|---|---|
| COURSE NUMBER : | CIND 119-DJ0 |
| COURSE TITILE : | INTRODUCTION TO BIGDATA ANALYTICS |
| SEMESTER & YEAR : | WINTER 2024 |
| INSTRUCTOR : | ZEKIYE ERDEM, Ph.D |
| STUDENT NAME : | KHOA TRUONG NGUYEN |

# TABLE OF CONTENTS

# 1. INTRODUCTION

In the financial sector, credit risk assessment plays a crucial role in determining whether a bank should approve a loan application. The objective is to minimize the risk associated with loans while maximizing the profitability of good loans. To achieve this, loan managers evaluate the demographic and socio-economic profiles of applicants before making loan approval decisions. The "German Credit Card" dataset has been chosen by our group to study credit risk assessment and construct a predictive model that can assist bank managers in making informed loan approval decisions.

The "German Credit Card" dataset contains valuable information about loan applicants, enabling us to explore various factors influencing credit risk on 20 variables and classifies whether each of the 1,000 loan applications is deemed to be a **Good** or **Bad** credit risk. The dataset incorporates a range of demographic and socio-economic attributes such as age, employment status, income, existing loans, credit history, etc.

The primary objective of this project is to develop a predictive model for credit risk assessment. This model will assist bank managers in evaluating loan applications and making informed decisions. The predictive model will classify loan applicants into two categories: those with good credit risk and those with bad credit risk. By accurately identifying applicants likely to repay their loans, the bank can minimize the risk of financial loss and maximize potential business opportunities.

To achieve our objectives, we will follow a systematic approach:

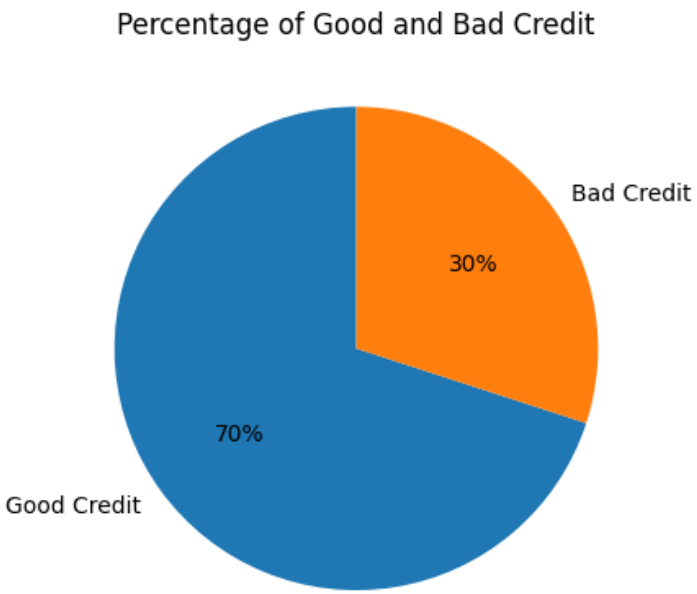
- Exploratory Data Analysis: We will perform a comprehensive analysis of the dataset to understand its structure, identify missing values, handle outliers, and gain insights into the distribution of variables.
- Model Selection: We will explore various classification algorithms suitable for credit risk assessment, such as decision trees, Naïve Bayes.
- Model Training and Evaluation: We will train the selected models on the dataset, validate their performance using appropriate evaluation metrics, and fine-tune them if required.
- Model Deployment: Once we have identified the best-performing model, we will deploy it as a tool for bank managers to aid them in loan approval decisions.

# 2. DATA PREPARATION

## 2.1. Data Information

The dataset utilized in our project, sourced from the German Credit Data, is provided in the ARFF (Attribute-Relation File Format). This dataset is renowned in the field of credit risk analysis and is widely used for predictive modeling tasks. The dataset contains ***1000 rows*** (applications) and ***21 columns*** (attributes), it offers a comprehensive overview of individuals' financial profiles, encompassing factors such as account balances, credit amounts, payment statuses, and personal demographics. By leveraging Python packages such as Pandas, NumPy, and Scikit-learn, we can efficiently load, preprocess, and analyze the data, empowering us to derive valuable insights and develop predictive models tailored to our specific objectives.

The attribute ***"Creditability***" as the classifier with *1 = good* credit risk and *0 = bad* credit risk. With 1000 applications in total, 70% or 700 applications are classified as having 'good' credit and 30% or 300 applications are classified as having 'bad' credit. The pie chart shows that 30% of the loan applicants defaulted. From this information, we see that this is an imbalanced class problem. Hence, we will have to weigh the classes by their representation in the data to reflect this imbalance.




The dataset is completeness with ***no missing*** values in any attributes. Regarding format, the datasets are integrated with both ***nominal and quantitative*** types. The dataset comprises a single set within a single time interval timeframe.

| | Attribute | Type | Min | Max | Mean | Std | Distinct |
|---|---|---|---|---|---|---|---|
| 1 | Creditability | Nominal | | | | | 2 |
| 2 | Account Balance | Nominal | | | | | 4 |
| 3 | Duration of Credit (month) | Quantitative | 4 | 72 | 20.903 | 12.058 | 33 |
| 4 | Payment Status of Previous Credit | Nominal | | | | | 5 |
| 5 | Purpose | Nominal | | | | | 10 |
| 6 | Credit Amount | Quantitative | 250 | 18424 | 3271.248 | 2822.751 | 923 |
| 7 | Value Savings/Stocks | Nominal | | | | | 5 |
| 8 | Length of current employment | Nominal | | | | | 5 |
| 9 | Instalment per cent | Quantitative | 1 | 4 | 2.973 | 1.118 | 4 |
| 10 | Sex & Marital Status | Nominal | | | | | 4 |
| 11 | Guarantors | Nominal | | | | | 3 |
| 12 | Duration in Current address | Nominal | | | | | 4 |
| 13 | Most valuable available asset | Nominal | | | | | 4 |
| 14 | Age (years) | Quantitative | 19 | 75 | 35.542 | 11.352 | 53 |
| 15 | Concurrent Credits | Nominal | | | | | 3 |
| 16 | Type of apartment | Nominal | | | | | 3 |
| 17 | No of Credits at this Bank | Quantitative | 1 | 4 | 1.407 | 0.577 | 4 |
| 18 | Occupation | Nominal | | | | | 4 |
| 19 | No of dependents | Quantitative | 1 | 2 | 1.155 | 0.362 | 2 |
| 20 | Telephone | Nominal | | | | | 2 |
| 21 | Foreign Worker | Nominal | | | | | 2 |

Figure 1: Descriptive Statistics Summary Table

## 2.2. Exploration of Quantitative Variables

Based on the correlation heatmap, three variables stand out for their significant impact on credit:

- ***Credit Amount***: The heatmap reveals a strong positive correlation between Credit Amount and Credit. This indicates that a higher credit amount correlates with greater creditworthiness. It's logical to assume that individuals with larger credit amounts are more likely to exhibit strong credit profiles and better management of credit obligations.
- ***Duration of Credit***: The heatmap suggests a moderate positive correlation between Duration of Credit and Credit. This implies that a longer credit duration is associated with higher creditworthiness. Individuals with longer credit histories may demonstrate better repayment records and financial stability, positively influencing their creditworthiness.

- ***Age:*** The heatmap displays a weak positive correlation between Age and Credit. Although not as pronounced as the previous variables, it still suggests that age plays a role in creditworthiness. Younger individuals may have limited credit histories and less experience managing credit, potentially impacting their creditworthiness. Conversely, older individuals may possess more established credit histories and financial stability, contributing to higher creditworthiness.

Observations (Histogram and Boxplot)

- The distribution of the continuous variables reveals that they span different ranges.
- The histogram indicates that many observations are concentrated within the first quantile of the variable. This observation can be further confirmed by examining box plots.
- The box plots illustrate that many credit amounts fall within the range of *1000 to 4500* dollars. Additionally, the distribution of credit amounts appears to be positively skewed.
- Regarding loan duration, most loans have durations between *15 to 30 months*.
- Most loan applicants fall within the age range of *28 to 43 years*.

Dealing with Outliers:

The dots that are outside of the 1.5 +/- IQR[Q3-Q1] are considered outliers. We decided to retain the outliers, as they were relatively few, and we aimed to maintain the integrity of our data as per its original source.

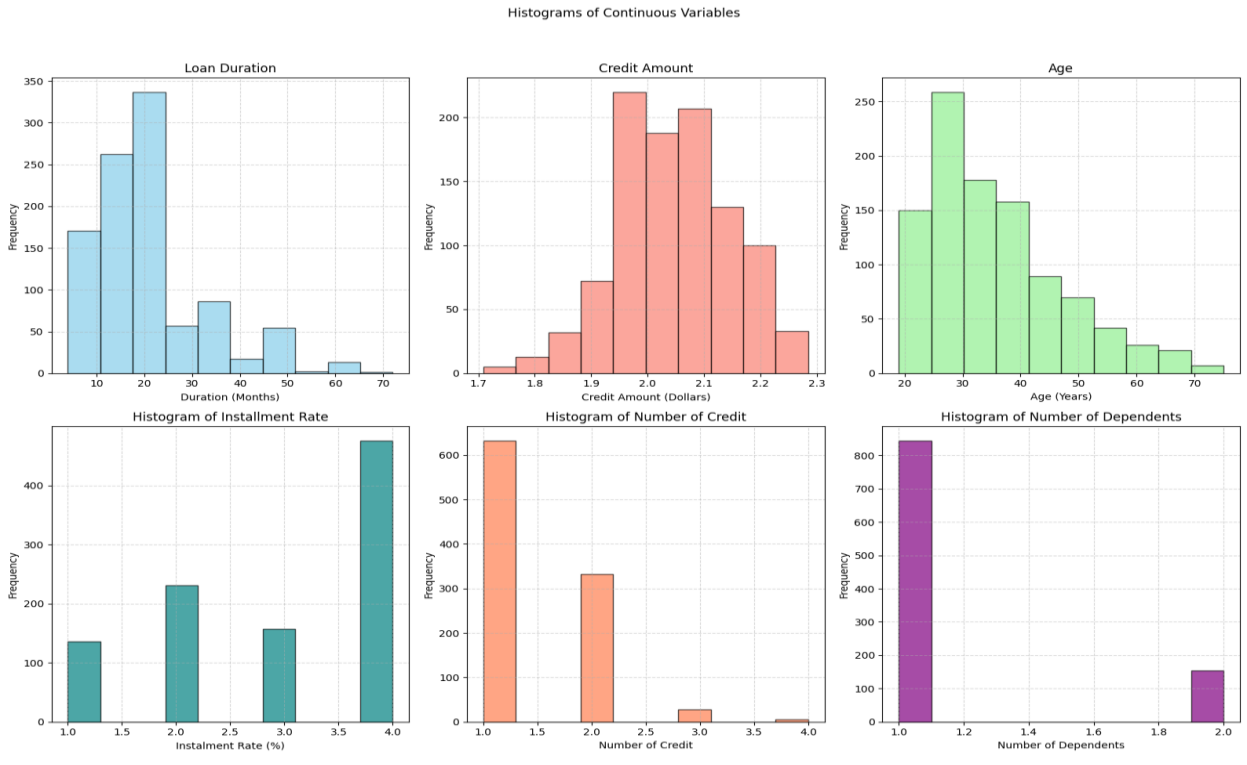



Figure 2: Histograms of Continuous Variables

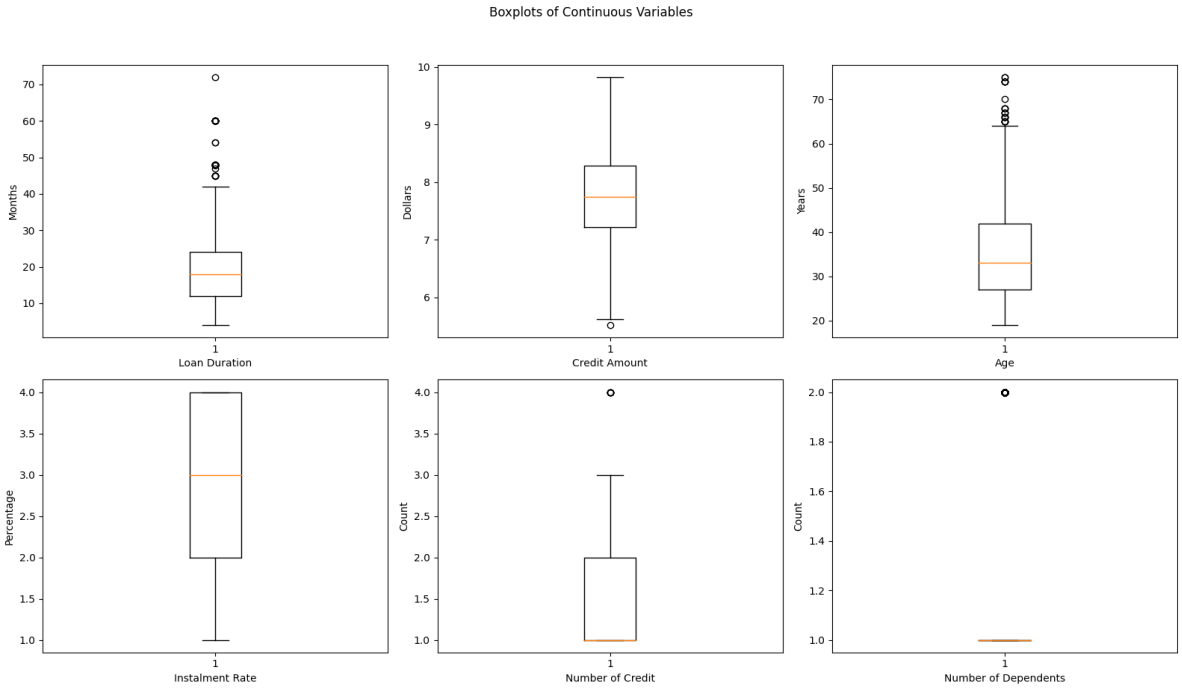



Figure 3: Boxplots of Continuos Variables

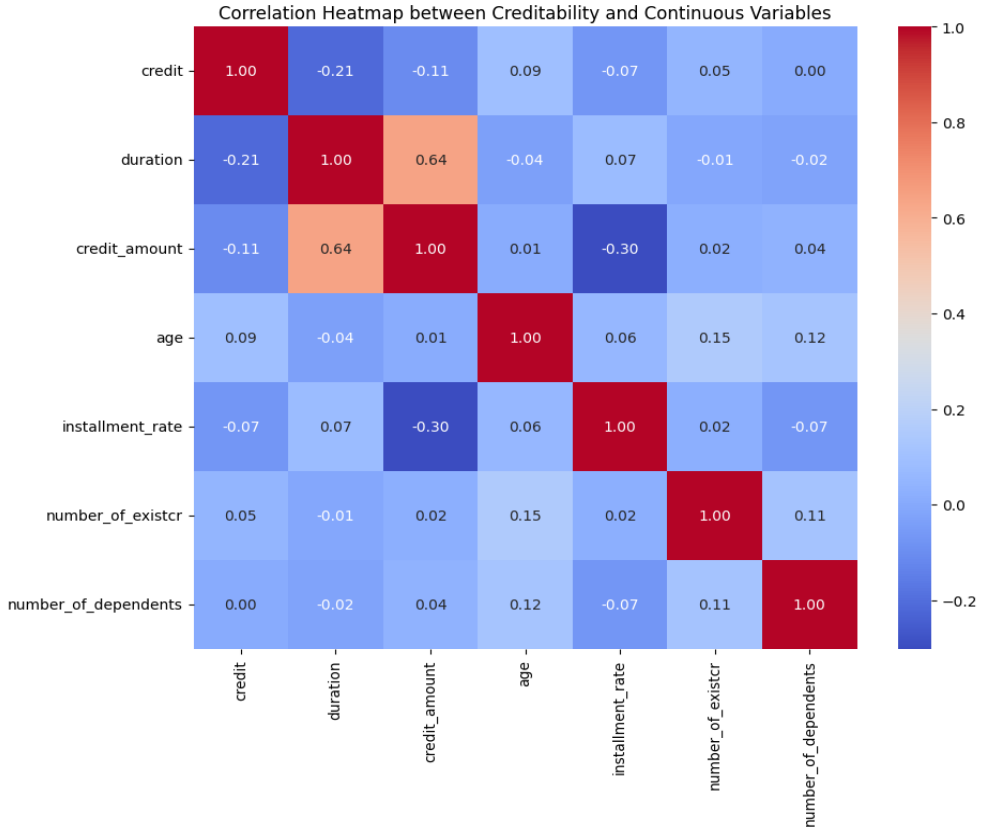



Figure 4: Correlation Heatmap between Creditability and Continuous Variables

## 2.3. Correlation between Nominal Variables and Creditability

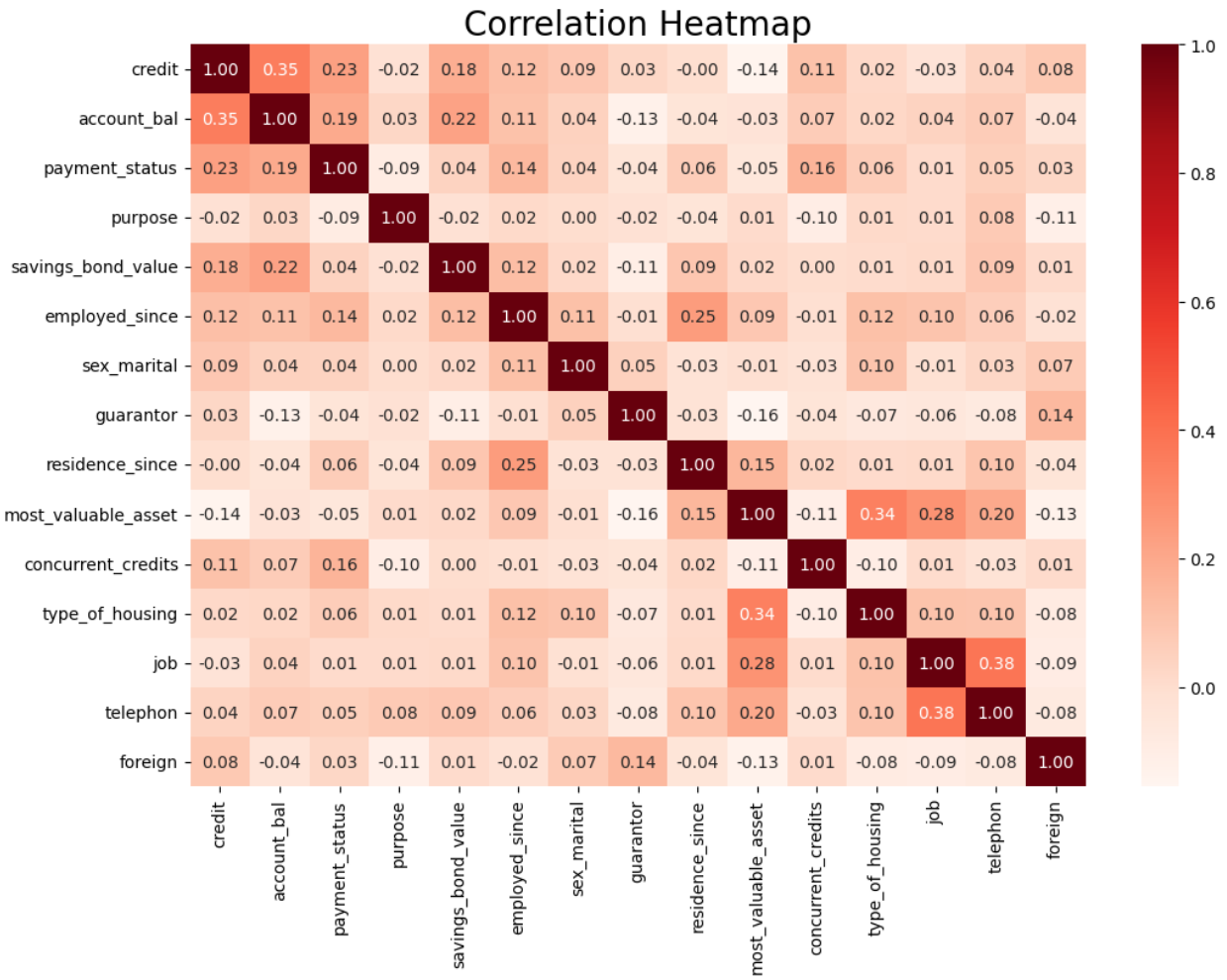


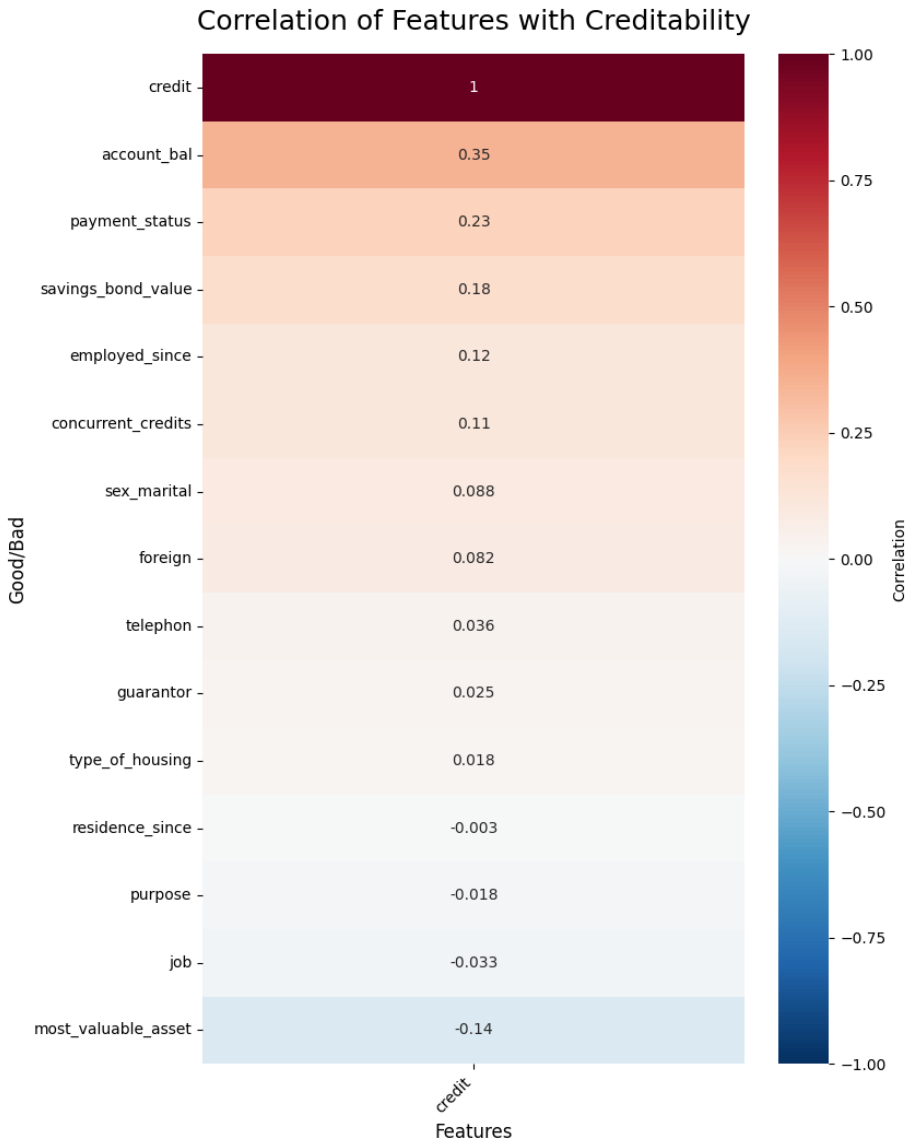

Figure 5: Correlation Heatmap Between Nominal Variables and Creditability

We can observe the relationship between nominal attributes and creditworthiness. The heatmap is divided into three sections, each indicating a different level of correlation strength.

In the upper section of the heatmap, there's a notable positive correlation between specific attributes and creditworthiness, indicating that an increase in these attributes significantly raises the probability of having good credit.

Middle section, a moderate positive correlation is observed, suggesting that these attributes moderately affect creditworthiness, albeit not as strongly as those in the upper section.

The lower section of the heatmap reveals a weak positive correlation, indicating that these attributes have a relatively minor impact on creditworthiness compared to those in the upper and middle sections.




Note: Positive values indicate a higher correlation with Good Credit, while negative values indicate a higher correlation with Bad Credit.

Figure 6: Correlation of Features with Creditability

## 2.4. Primary Attributes Linked to the Class Attribute

Quantitative Attributes:

- Credit Amount: This attribute shows a strong positive correlation with creditworthiness. The higher the credit amount, the more likely an individual is to have good credit. We selected this attribute because it suggests that individuals with larger credit amounts are more likely to have better credit profiles and be responsible in managing their credit obligations.

- Duration of Credit: This attribute demonstrates a moderate positive correlation with creditworthiness. A longer credit duration is associated with higher creditworthiness. It suggests that individuals with longer credit histories may have better repayment records and financial stability, positively impacting their creditworthiness.
- Age: Although the correlation is weak, the heatmap indicates a positive relationship between age and creditworthiness. Age plays a role in determining creditworthiness. Younger individuals may have limited credit histories and less experience managing credit, potentially affecting their creditworthiness. Conversely, older individuals may have more established credit histories and financial stability, contributing to higher creditworthiness.

Nominal Attributes:

- Account Balance: Higher account balances are more likely to have good credit, as indicated by the strong positive correlation between Account Balance and creditworthiness.
- Payment Status of Previous Credit: Having a good payment history on previous credit is also strongly correlated with creditworthiness, suggesting that individuals with a good payment status are more likely to have good credit.
- The purpose of the credit: It plays a role in creditworthiness, with a moderate positive correlation suggesting that the purpose of the credit may have an impact on an individual's creditworthiness.
- Individuals with higher savings or stocks value are more likely to have good credit, as indicated by the strong positive correlation between Savings/Stocks Value and creditworthiness.

## 2.5. Dataset Balancing

With 70% of applications classified as "good" credit and only 30% classified as "bad" credit, there is a significant class imbalance. Imbalanced datasets can pose challenges for machine learning models because they may become biased towards the majority class, leading to poor performance in predicting the minority class.

To address this imbalance, we may need to apply techniques such as resampling (e.g., oversampling the minority class or under sampling the majority class), using different evaluation metrics (e.g., precision, recall, F1-score), or employing algorithms specifically designed to handle imbalanced data.

## 2.6. Elimination of Attributes

From the observations made in the previous section, it becomes evident that certain attributes may be redundant or unsuitable for modeling due to various factors. Raw data often requires refinement before it can be effectively utilized in machine learning models. The process of determining which attributes to include in a machine learning model is known as feature selection. The feature selection can be done by using the BestFIRst algorithm in Weka.

```
# attributes: 3
attributes: [1 2 3 0]
result string:


=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 133
        Merit of best subset found:    0.076

Attribute Subset Evaluator (supervised, Class (nominal): 1 Creditability):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,4 : 3
                     Account Balance
                     Duration of Credit (month)
                     Payment Status of Previous Credit
```

The algorithm has identified three attributes as significant factors in determining creditworthiness: Account Balance, Duration of Credit (month), Payment Status of Previous Credit.

# 3. PREDICTIVE MODELING (CLASSIFICATION)

## 3.1. Data Split Strategy

The dataset was split into two subsets: a training set and a test set, with 66% of the data allocated to the training set and 34% to the test set. This split was performed within the WEKA environment using the 'train_test_split' function. A seed value of 1 was specified for the random number generator to ensure reproducibility of the split. The resulting training and testing sets were stored in variables named 'train' and 'test', respectively. This splits ratio and random seed configuration enable consistent and reliable model training and evaluation within the WEKA platform.

## 3.2. Decision Tree

We were generating a pruned C4.5 decision tree (J48 decision in WEKA) using a confidence factor of 0.25 for pruning. By setting it to 0.25, we were indicating that nodes with a confidence below this threshold will be pruned during the tree construction process. Once the tree is built, we will be able to visualize its structure and examine the decision rules it learned.

The decision tree generated offers a comprehensive understanding of how the model classifies data. It delineates various decisions based on attribute values, ultimately leading to classification outcomes at the leaf nodes. Here's an overview of the decision tree:

- The tree initiates by partitioning the data according to the "Account Balance" attribute.
- Subsequent splits occur based on different attributes for each "Account Balance" value.
- Conditions in each branch outline the criteria for guiding decision-making.
- At the terminal leaf nodes, the predicted class is provided for the associated attribute value set.

With a total of 104 nodes, including 74 leaf nodes, this tree demonstrates a complex decision-making process. The evaluation summary provides key performance metrics for the decision tree model when applied to the test set:

- 72.35% of instances were classified correctly.
- There are 340 instances in the test set.

| | Position 0 | Position 1 |
|---|---|---|
| TP Rate | 0.393 | 0.896 |
| FP Rate | 0.103 | 0.606 |
| Precision | 0.666 | 0.73 |
| Recall | 0.393 | 0.8896 |

## 3.3. Naïve Bayes

Naive Bayes is an algorithm that discovers every object's probability, characteristics, and groupings. An alternative name for it is a probabilistic classifier. Under supervised learning, the Naive Bayes Algorithm is primarily utilized for resolving classification issues.
For instance, there are numerous fish with identical characteristics, therefore you cannot recognize a fish just by its color and features. Nonetheless, you formulate a probabilistic forecast regarding the same, and the Naive Bayes Algorithm is utilized for that purpose. We started by loading data and checking it.
The task of risk prediction follows a typical supervised classification approach:

Supervised: This involves training the model with labeled data, aiming to teach it to predict these labels based on the features provided.
Classification: The labels in this case are binary, with '0' indicating no risk (loan expected to be repaid on time) and '1' indicating risk (difficulty foreseen in loan repayment).

In any modeling task, we initiate by establishing the baseline model. This model serves as the point of comparison for the models under development, enabling us to identify any enhancements achieved. The most basic prediction entails considering all loan applicants as having a favorable credit rating, resulting in approximately 70% accuracy. However, this figure lacks significance without proper context. Therefore, we compare this baseline with the models we construct, as well as with existing literature. Misclassification stemming from this baseline is then weighed by the associated penalty score to determine the total penalty incurred by the model. We analyze both the accuracy and mean penalties of the Baseline1 model to gain insights.
In this report, we develop and compare Bayesian Networks which attempt to predict whether loan applications will be successful or not.

When we created NB networks, using all twenty variables, we can see that it has an accuracy of almost 73%. Meaning, almost more than 700 out of the 1000 loans will be repaid.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| bad | 0.00 | 0.00 | 0.00 | 70 |
| good | 0.73 | 1.00 | 0.84 | 189 |
| accuracy | | | 0.73 | 259 |
| macro avg | 0.36 | 0.50 | 0.42 | 259 |
| weighted avg | 0.53 | 0.73 | 0.62 | 259 |

# 4. CONCLUSION AND RECOMMENDATIONS

We have modeled the German Credit Data set using naive and simple baseline models.

Results of the same data set available elsewhere show similar order of accuracy for prediction. Results from Applications of Data Mining in E-business and Finance also give similar accuracies.

Ultimately these statistical decisions must be translated into profit consideration for the bank. Let us assume that the correct decision of the bank would result in 35% profit at the end of 5 years. A correct decision here means that the bank predicts an application to be good or credit-worthy and it turns out to be creditworthy. When the opposite is true, i.e. the bank predicts the application to be good, but it turns out to be bad credit, then the loss is 100%. If the bank predicts an application to be non-creditworthy, then the loan facility is not extended to that applicant and the bank does not incur any loss.

The current model has an accuracy of 73% on unseen data.
The Model makes a trade-off i.e., to reduce False-Negative, the False-Positive prediction rate increases.