الجمهورية الشعبية الديمقراطية الجزائرية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

المدرسة العليا للإعلام الآلي ٠08 ماي 1945٠ بسيدي بلعباس

École Supérieure en Informatique

-08 Mai 1945- Sidi Bel Abbès

Mémoire de Fin d'étude

En Vue de l'obtention du diplôme de **Master**

Filière : **Informatique**

Spécialité : **Ingénierie des Systèmes Informatiques (ISI)**

Thème

Loan Approval Prediction Using Machine Learning

Présenté par :

- Mlle. CHERIF Achouak
- Mlle. BERKANE Naoul

Soutenu le : **30/09/2023** Devant le jury composé de :

Mme. BELALIA Amina	Présidente
M. KHALDI Miloud	Encadreur
M. CHAIB Souleyman	Encadreur
M. BEKKOUCHE Mohammed	Examinateur

Année Universitaire : 2022/2023

Acknowledgments

We would like to begin by expressing our gratitude to **ALLAH** for granting us the strength and patience to undertake this journey.and also the courage to continue despite the encountered difficulties. We extend our heartfelt thanks to our families and friends, with a

special mention to our parents, whose Unchanging support has been our anchor throughout this remarkable journey and our entire lives. Their support, which words cannot adequately convey, has been a constant source of inspiration.

We would like to express our sincere appreciation to our supervisors **Dr. Khaldi Miloud** and **Dr. Chaib Souleyman** for their invaluable assistance, unwavering patience, and constant encouragement.as they steered us in the best possible direction to establish and complete our thesis successfully.

Special thanks to the administration and working team of our school, especially the school director **Pr. Benslimane Sidi Mohammed** and **Pr. Amar Bensaber Djamel**, and all the teachers who provided us with great knowledge throughout all the five years.

This thesis would not have been possible without the support and guidance of these remarkable individuals, and for that, we are truly grateful.

Dedication

I dedicate this work to my dear parents, thank you for giving me the support to be where I am today. Accomplishing this would hopefully make you proud of me as much as I am proud of having you as my parents. I would also like to thank my sister **Sondos** and my two brothers **Zakaria** and **Mahmoud** for all the things you have done for me.

I also dedicate it,to each member of my father's family and espacially I want to thank my beloved uncle **Cherif Bensadek** for always being there to help and give support and encourage me throughout my journey. And I dedicate it to each member of my mother's family(Grine) which I consider my second family and espacially mentioning my uncle **Grine Cherif** you are truly the best.

I would like to thank the friends I met in ESI SBA. The memories we shared during these wonderful five years filled with joy, late night talks and the unforgettable moments will remain with me for good. So thank you for being part of this journey.

Whithout Forgetting all the teachers and professors in the higher school of computer science, and my high school, middle school and primary school, I dedicate this to all of you, for you have shared your knowledge and effective teachings to me.

Finally, I would like to thank everyone who contributed to the success of this work from near or far.
(Achouak Cherif)

Dedication

I dedicate this work to my beloved family, with a special dedication to my source of inspiration, my big brother **BERKANE YOUCEF** ,dear brother. Your constant support, belief in me, and never-ending encouragement have always pushed me forward. While your name may find a place on these pages, your influence runs much deeper, woven into every word and every effort. I hope this work makes you as happy and as proud as I am to have you as my brother. It is dedicated to With deep love and gratitude.

To my friends Thank you for being with me on this journey. Your presence, late-night study sessions, laughter, and shared experiences have made this chapter of my life truly memorable.

To the teachers and professors, I extend my heartfelt thanks for sharing your knowledge and wisdom. Your guidance has been instrumental in shaping my academic path and I am forever grateful.

Lastly, I want to express my appreciation to all those who have contributed, near and far, to the success of this work. Your support and encouragement have been invaluabl

(BERKANE Naoul)

Abstract

Incorrect decision-making in financial institutions can have severe consequences, leading to financial crises.In recent years, numerous studies have highlighted the potential of artificial intelligence techniques as alternative methods for credit scoring.

Some researches has shown that prediction models utilizing hybrid approaches outperform single approaches. Furthermore, incorporating feature selection techniques into prediction models can enhance their performance.

The objective of this work is to explore the field of credit risk and compare a set of researches that propose approaches for loan default prediction. We then present a summary of the approaches and a comparative table of them, showing their results and the different techniques used.

Keywords : Digital Lending, Credit Risk, Credit Scoring, FinTech, Machine learning, Features selection.

Résumé

Une prise de décision incorrecte dans les institutions financières peut avoir des conséquences graves, pouvant entraîner des crises financières. Ces dernières années, de nombreuses études ont mis en évidence le potentiel des techniques d'intelligence artificielle en tant que méthodes alternatives pour l'évaluation du crédit.

Certaines recherches ont montré que les modèles de prédiction utilisant des approches hybrides surpassent les approches individuelles. De plus, l'incorporation de techniques de sélection des caractéristiques dans les modèles de prédiction peut améliorer leurs performances.

L'objectif de ce travail est d'explorer le domaine du risque de crédit et de comparer un ensemble de recherches qui proposent des approches pour la prédiction des défauts de paiement des prêts. Nous présentons ensuite un résumé des approches et un tableau comparatif les illustrant, en exposant leurs résultats et les différentes techniques utilisées.

Mots-clés : Prêt numérique, Risque de crédit, Évaluation du crédit, FinTech, Apprentissage automatique, Sélection des attributs.

Contents

List of Figures

List of Tables

List of acronyms

1. **AI :** Artificial Intelligence.
2. **ML :** Machine Learning.
3. **DL :** Deep Learning.
4. **SVM :** Support Vector Machines.
5. **KNN :** K-Nearest Neighbor.
6. **NB :** Naive Bayes.
7. **RF :** Random Forest.
8. **LR :** Logistic Regression.
9. **RL :** Reinforcement Learning.
10. **TP :** True Positives.
11. **TN :** True Negatives.
12. **FP :** False Positive.
13. **FN :** False Negative.
14. **MAE :** Mean Absolute Error.
15. **MSE :** Mean Squared Error.
16. **RMSE :** Root Mean Square Error.

Part I

General Introduction

Chapter 1

General Introduction

1.1 Context

With the enhancement in the banking sector lot of people are applying for bank loans everyday, and because the bank has limited resources, and not all the borrowers are eligible for receiving a loan, the bank can not approve every loan application, which it has to grant to limited people only, so the main concern of the banks is to assign the loan to a safe condidate. And this is a very challenging task due to various conditions such as the huge amount of data which needs to be filtered to make the final decision, especially in the past where the evaluation primarily depended on manual review, which was very exhausting and time-consuming. But, recently, banks have opted for machine learning approaches to automatically predict the loan default since it can highly enhance the accuracy and the efficiency of the prediction.[Wu, 2022]

However the ML industry made a lot of improvement in the loan defaults area and could be a highly effective solution for loan default prediction to help both banks and borrowers. In this work, we are interested in creating a web application that contains an AI model using an approach based on machine learning to find a solution to this problem.

1.2 Problem Statement

The good operation of this mechanism must prevent loan defaults and calculate the possible default risk of borrowers before issuing loans .

Loan lending plays an important role in our everyday life. Applying for a loan has been inevitable for people since individuals around the world depend on loans to overcome financial constraints to achieve their personal goals, and organizations rely on loans to expand their production.[Zhang et al., 2020] In most cases, loan lending is beneficial to both the borrowers and the lenders. However, loan default is still unavoidable, which carries a great risk. So preventing loan defaults and calculating the possible default risk of borrowers before issuing loans will save lots of bank efforts and resources.

So how could we know that a borrower is eligible for receiving the loan using Machine Learning loan prediction algorithms?

1.3 Objective

In this master study, our objective is to explore the field of loan defaults, and to represent the different techniques and approaches proposed and their results.

We also intend to present the approaches of Machine Learning, its applications and to develop our own model as a contribution to the problem of loan default prediction.

1.4 Thesis Organisation

- **Chapter 01:** The first chapter will be a presentation of the context of the work, as well as a description of the problem.
- **Chapter 02:** The second chapter will be a presentation and a brief definition of loan approval problem.
- **Chapter 03:** The third chapter will be a definition of Artificial Intelligence, Machine Learning as well as the methods of each one of them.
- **Chapter 04:** In the fourth chapter we're going to presented the resources and papers needed for our problem, as well as a state of the art of existing methods.
- **Chapter 05:** The last chapter will present the conclusion of this work.

Part II

Background

Chapter 2

Basic concepts

2.1 Introduction

In the fast-changing world of digital lending,credit risk emerges as a pivotal concern. This exploration delves into the world of digital lending, highlighting the crucial role of credit risk assessment and management in shaping the future of finance.

2.2 Digital Lending

2.2.1 What is Digital Lending

‘Digital lending is the process of offering loans that are applied for, disbursed, and managed through digital channels, in which lenders use digitized data to inform credit decisions and build intelligent customer engagement’.[Ravikumar, 2019]

2.2.2 Digital Lending Models

The major models of digital lending are presented in table 2.1 :

Online Lender	Finantcial Service Provider (FSP) that provides end-to-end digital lending products via a website or mobile application.
P2P Lender	Digital Platforms that facilitate the provision of digital credit between many borrowers and lenders, typically playing an ongoing central role in the relationship between these parties.
e-commerce and Social Platforms	Digital Platforms wherein credit is not their core business, but that leverages their digital distribution, strong band, and rich customer data to offer credit products to their customer base.
MarketPlace Platforms	Digital Platforms that originate and match one borrower with many lenders for an origination fee; the lender and borrower then enter into a bilateral agreement.
Supply Chain Lender	No-cash digital loans for specific asset financing, invoice financing, or pay-as-you-go asset purchase within a supply chain or distribution network.
Mobile Money Lender	Partnership model wherein lenders work with mobile network operators (MNOs) to offer mobile money loans to their customer base, leveraging mobile phone data for scoring.
Tech-enabled Lender	Traditional FSPs that have digitized parts of the lending process, either in-house or through partnership.

Table 2.1: Digital lending models
[Ravikumar, 2019]

2.3 Credit risk

2.3.1 Credit loans

Credit loans, are financial products offered by banks and other financial institutions to individuals or customers. These loans are provided with the expectation that they will be paid back within a specific period which can vary based on the terms and conditions agreed upon, either with or without interest.[Anand et al., 2022]

Individuals and businesses worldwide rely on Credit loans to overcome financial constraints and achieve their goals. Moreover, financial institutions consider it a major profit-making opportunity and vital for their smooth functioning. However, loan lending also carries significant risks that need to be managed effectively.[Aslam et al., 2019]

2.3.2 Credit Risk In Banking

Credit risk, which is also commonly known as default risk, represents the most significant financial risk within the banking system. It denotes the inability or unwillingness of the borrower to repay the loan within the mutually agreed time frame, as decided by the lender and the borrower during the loan origination process.[Aslam et al., 2019]

As [Himberg, 2021] mentionned their are various financial risks faced by banks as , which can be broadly categorized as credit risk, market risk, liquidity risk, and interest rate risk. Credit risk, refers to the risk of a debtor failing to repay their loan, resulting in losses for the lender. This risk encompasses situations where borrowers, counterparties, or investments fail to meet their obligations. Bandyopadhyay (2016) further highlights that loans are the most common source of credit risk for banks.

Default risk is a specific subclass, representing the probability that a borrower will not comply to the loan terms and fail to repay the loan. [Bandyopadhyay, 2016] emphasizes that estimating default risk is a fundamental aspect of assessing credit risk. Banks employ various methods for estimating default probabilities, including using their own historical data and experience, linking internal default information to external data, and utilizing default models.[Bessis, 2015]

2.3.3 Loan Default and Its Impact

Loan default occurs when a borrower fails to fulfill their payment obligations to the lender, resulting in financial losses, including capital, interest, and increased collection costs. Loan defaults can be categorized as temporary or indefinite. Temporary defaults occur when the borrower manages to repay the overdue amount, resulting in partial loss. Indefinite defaults, on the other hand, occur when the payment is overdue for 90 days or more. In such cases, the losses are higher as no interest is paid for an extended period, and collection costs continue to rise. [Himberg, 2021]

The Basel II Agreement, provides a definition for loan default based on two conditions. Firstly, the lender considers that the borrower is unlikely to repay the loan in its entirety. Secondly, the borrower's past due amount must exceed 90 days on any credit. According to the Basel II Agreement, meeting either of these conditions qualifies as a loan default. The Basel II Agreement holds significant importance as a regulatory framework, ensuring the resilience of banks in the face of potential risks. It mandates financial institutions to calculate credit risk components which are exposure at default (EAD), loss given default (LGD) and default probability (DP), in order to ensure that lenders are prepared for potential defaults. [Himberg, 2021]

2.3.4 Credit risk Management

Credit risk management in banking refers to the process employed by banks to regulate their financial exposures. It has gained paramount importance not only due to the ongoing financial industry challenges but also because it profoundly influences a bank's survival, growth, and profitability [Abiola et al., 2014],. The effectiveness of credit risk management stands as a pivotal predictor of a bank's financial performance, making it a foundation for a bank's success. Managing credit risk is one of the most critical tasks for ensuring the financial liquidity and stability of the banking sector, particularly given the heightened sensitivity of banks to credit risks. To achieve effective credit risk management in commercial banks, the development of credit risk assessment techniques is essential, as these techniques play a primary role in the management and minimization of credit risk.[Konovalova et al., 2016]

2.3.5 Credit risk Assessment

Credit assessment encompasses a set of decision models and techniques that assist lenders in making informed choices when extending consumer credit. It involves evaluating the risk associated with lending to various consumers, ensuring responsible lending practices[Wu et al., 2010]. Credit assessment can be broadly categorized into two types: quantitative assessment, which relies on data-driven methods, and qualitative assessment, which involves judgment-based approaches.[Uthayakumar et al., 2020]

2.3.5.1 Qualitative Assessment:

Often referred to as subjective models, draw on experts' problem-solving knowledge. These models enrich credit risk assessment with qualitative insights and expert judgment.
[Uthayakumar et al., 2020]

2.3.5.2 Quantitative Assessment:

Play a pivotal role in credit risk assessment, This data-driven approach leverage data mining techniques like discriminant analysis,neural networks, and more. These models focus on the analysis of extensive financial data to learn classification functions.quantitative Assessment harness the power of data to objectively evaluate credit risk . A key component of quantitative assessment is credit scoring.[Uthayakumar et al., 2020]

- **Credit Scoring:**
  Credit scoring is a way to assess the risk associated with lending money by using statistical or data-driven methods on historical information. Its main goal is to predict whether a person applying for credit is likely to be a reliable borrower (creditworthy) or not (non-creditworthy).Credit score classification using spiking extreme learning machine.
  In the past, this process relied on both experts and statistical algorithms to decide whether to approve or reject a loan application. However, recent advancements have led to the use of machine learning and deep learning techniques. These technologies

automatically analyze a person's credit history and other relevant data to predict their credit score. This makes it much easier to identify eligible candidates for loans before approval.[Madaan et al., 2021]

2.4 Conclusion

In the world of digital lending , credit scoring stands as a crucial tool for responsible finance. As we wrap up this section , it's clear that mastering credit scoring and risk management is essential for success in this rapidly evolving financial landscape.

Chapter 3

Artificial Intelligence

3.1 Introduction

Artificial Intelligence (AI) is a transformative field with a rich history spanning 60 years. This exploration outlines the evolution of AI, its types, applications, and delves into Machine Learning (ML) and its process while highlighting the distinction between AI and ML.

3.2 Definition of Artificial Intelligence

As the father of artificial intelligence John McCarthy, who created the term of "Artificial intelligence" in 1956, said that "It is the combination of science and engineering to make intelligent devices for human welfare." "Artificial intelligence is an intellect that is much smarter than the best human brain in practically every field, including computer science and linguistic logic".[Rupali and Amit, 2017]

In simple terms, AI aims to extend and augment the capacity and efficiency of mankind in tasks of remaking nature and governing the society through intelligent machines, with the final goal of realizing a society where people and machines coexist harmoniously together.
[Liu et al., 2018]

3.3 The 60-year developmental history of AI

Over the last 60 years, artificial intelligence (AI) has experienced continuous development, marked by significant progress as well as challenges and setbacks. By reviewing the valuable insights gained from this journey, we can better understand the trends and patterns that have shaped the development of AI. [Pan, 2016]

3.3.1 The birth of AI :

Back to 1956, a group of scholars including Professor J. McCarthy, Professor M. L. Minsky, Professors H. Simon and A. Newell, C. E. Shannon, N. Rochester, and others, established the concept of "artificial intelligence" at Dartmouth College in the US. Their definition referred to machines possessing the ability to understand, think, and learn, similarly to human beings, thereby opening the door to computer simulations of human intelligence. Since 1970s, AI has expanded into numerous research fields such as mechanical theorem proving, machine translation, expert systems, game theory, pattern recognition, machine learning, robotics, and intelligent control. The exploration within these fields has led to the development of various technologies. [Pan, 2016]

3.3.2 Three setbacks to AI development :

AI development has encountered significant challenges throughout its 60-year history, with three major :

- **The first setback :** occurred in 1973, accompanied by a report authored by James Lighthill and published in England. The report assessed the progress of fundamental AI research in areas like automaton, robots, and the central nervous system. It concluded that while research on the automaton and central nervous system showed some value but fell short of expectations, research on robots was deemed worthless and marked as a spectacular disappointment, leading to its cancellation. This setback happened during the early stages of AI's development, highlighting the initial challenges faced by the field.

- **The second setback :** concerned a project undertaken by Japan, which aimed to develop an intelligent computer known as the fifth-generation, launched in 1982. The aim of the project was to develop an intelligent machine capable of inference and knowledge processing. With a goal to construct a parallel-inference machine, the project failed to achieve breakthroughs by 1992, despite an expenditure of 850 million dollars. This setback highlighted the need to prioritize innovation and software development in AI, recognizing hardware as a supportive element.

- **The third setback :** In 1984, when they tried at Stanford University to manually create Cyc, an encyclopedia of knowledge, aiming to achieve human-like inferential ability. However, with the rapid advancement of the Internet and big data, Cyc couldn't keep up and its development declined by the end of the 1990s . Even though external knowledge bases were integrated, the project highlighted the impracticality of relying solely on human exports for vast knowledge. The lesson learned emphasized the significance of gaining knowledge from the environment. [Pan, 2016]

3.3.3 Artificial intelligence 2.0 :

Analyzing the setbacks in AI development reveals a consistent cause: the inability of AI to adapt to changes in the information environment. The progress of AI is influenced by both research efforts and the information environment, but the latter has a stronger impact. With the wide spread use of the Internet, the presence of sensors, the growth of e-commerce, the emergence of big data, and the integration of data and knowledge across physical and cyber space, the information environment has experienced a profound transformation. This has given rise to a new phase of AI, called AI 2.0, where new technologies play a crucial role in advancing AI capabilities.[Pan, 2016]

3.4 Types of Artificial Intelligence

Artificial Intelligence (AI) has been a topic of significant interest in recent years due to its potential to revolutionize many aspects of human life [Krenn et al., 2022],AI systems can be classified into three distinct types: Artificial Super intelligence (ASI), Artificial General Intelligence (AGI), and Artificial Narrow Intelligence (ANI)[Kaplan and Haenlein, 2019],these classifications are based on the degree of a machine's cognitive capabilities compared to human intelligence .[Chiu et al., 2021]

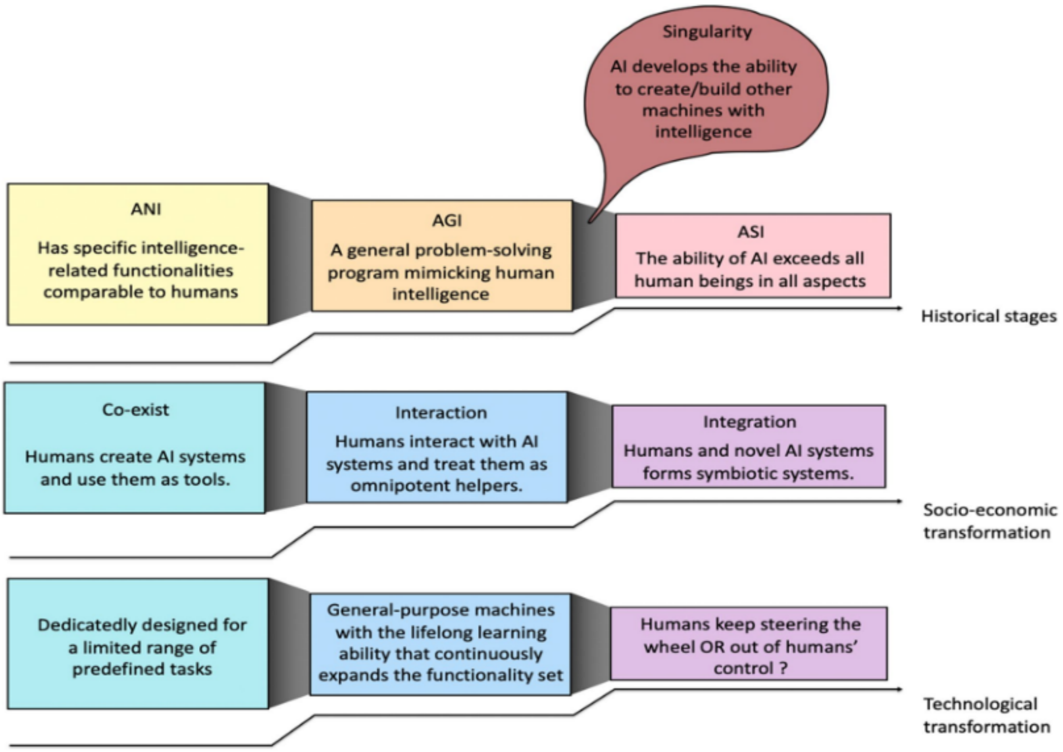

Singularity
AI develops the ability to create/build other machines with intelligence
ANI
Has specific intelligence-related functionalities comparable to humans
AGI
A general problem-solving program mimicking human intelligence
ASI
The ability of AI exceeds all human beings in all aspects
Historical stages
Co-exist
Humans create AI systems and use them as tools.
Interaction
Humans interact with AI systems and treat them as omnipotent helpers.
Integration
Humans and novel AI systems forms symbiotic systems.
Socio-economic transformation
Dedicatedly designed for a limited range of predefined tasks
General-purpose machines with the lifelong learning ability that continuously expands the functionality set
Humans keep steering the wheel OR out of humans' control ?
Technological transformation

Figure 3.1: Types of Artificial Intelligence
[Jiang and Zhu, 2022]

3.4.1 Artificial narrow Intelligence :

Artificial Narrow Intelligence (ANI) or Weak Artificial Intelligence refers to computer systems that are able to perform only one single ,precise task extremely well. These systems are not capable of defining new problems or exploring new areas, but rather rely on well-defined tasks assigned to them by humans. Therefore, the effectiveness of ANI is dependent on how accurately and clearly these tasks are defined for the machine to execute [Jiang and Zhu, 2022]. All current examples of AI fall under this category, including the most complex machines that use advanced techniques like deep learning and machine learning.[Chiu et al., 2021]

3.4.2 Artificial General Intelligence :

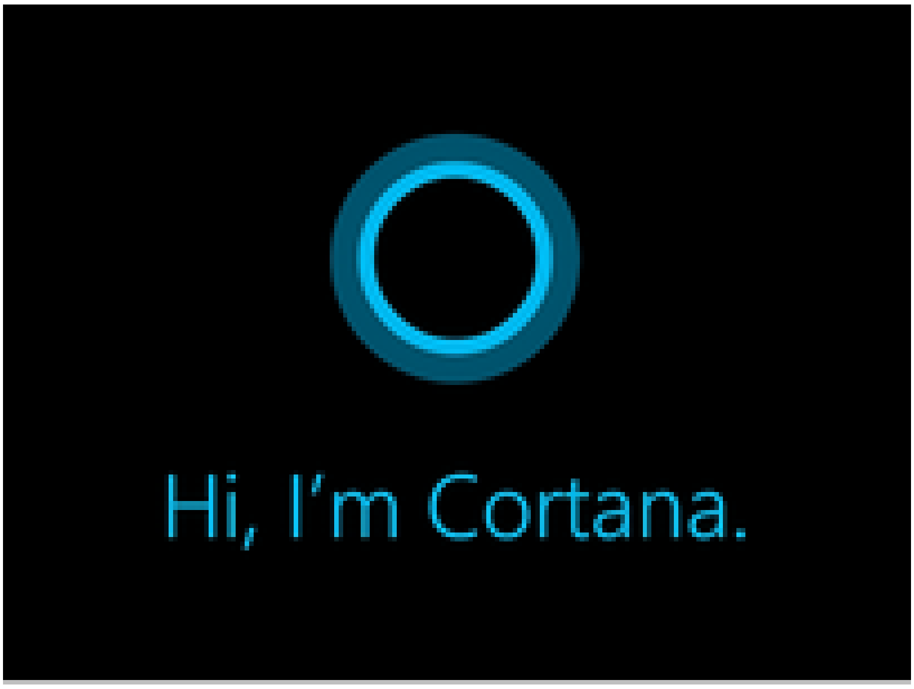
Artificial General Intelligence (AGI) or Strong Artificial Intelligence refers to machines that are expected to have abilities such as reasoning, problem-solving, planning, and learning, as well as the capacity for imagination and creativity, allowing them to function like humans. Such systems would be capable of developing diverse competencies and forming connections across different domains on their own ,as well as performing a range of tasks for which they have not been specifically trained. However, as of now, no machine possessing AGI has been created.[Chiu et al., 2021]

3.4.3 Artificial super Intelligence :

ASI surpasses AGI by enabling machines to possess a level of intelligence that exceeds human's intelligence . This implies that machines with ASI will outperform mankind in all aspects , including learning, prediction, creativity, productivity, decision-making, organization, management, as well as survival.Additionally, ASI systems may possess the remarkable capability to reprogram the firmware and design hardware from scratch, thereby expanding their potential beyond the limitations of current technologies.[Jiang and Zhu, 2022]

3.5 Application of Artificial Intelligence

Nowadays, AI is the most useful topic in human life. There are many examples of AI. There is Siri by apple, google now by google, Watson by IBM and cortana by windows mobile for various operating systems which are intelligent digital personal assistants (which have gesture and speech recognition) to help the users to find and sort out all the needed things without any physical appearance.[Rupali and Amit, 2017]



Hi, I’m Cortana.

Figure 3.2: Application of Artificial Intelligence

3.6 Machine Learning

3.6.1 Definition

It is a concept that enables machines to learn from real-world interactions and observations and behave like human beings and improve their ability to learn and perform using data given as input. In the recent years, ML has gained a huge focus and interest of researchers and technologists that they are trying to implement various machine learning models and algorithms in fields which will make various important tasks and lives of common man a lot easier. Two popular examples are the banking sector and finance. With the help of various ML models, banking authorities and financial firms are observing patterns and making conclusions in areas like credit card frauds, loan default prediction. It has made the process much easier now and more accurate.[Madaan et al., 2021]

3.6.2 Examples of Machine Learning Applications

3.6.2.1 Learning Associations :

[Alpaydin, 2020] gives the example of a supermarket chain which is finding associations between products bought by customers: if we can see that in most of the cases customers who buy X typically also buy Y, and if there is a customer who buys X and does not buy Y, he or she is a potential Y customer. Once we find such customers, we can target them for cross-selling.

3.6.2.2 Classification :

When a bank loaned some amount of money to someone to be paid back with interest this is called credit. In this case it is so important for the bank to predict in advance the risks of this loan and the possibility of the customer to default and not pay the whole amount back. The bank can calculate the risk through the amount and information about the customer. This is an example of a classification problem where there are two classes: low-risk and high-risk customers. The information about a customer makes up the input to the classifier whose task is to assign the input to one of the two classes.[Alpaydin, 2020]

3.6.3 Types of machine learning

According to [Janiesch et al., 2021] there exist three types of Machine learning :

- Supervised Learning.
- Unsupervised Learning.
- Reinforcement Learning.

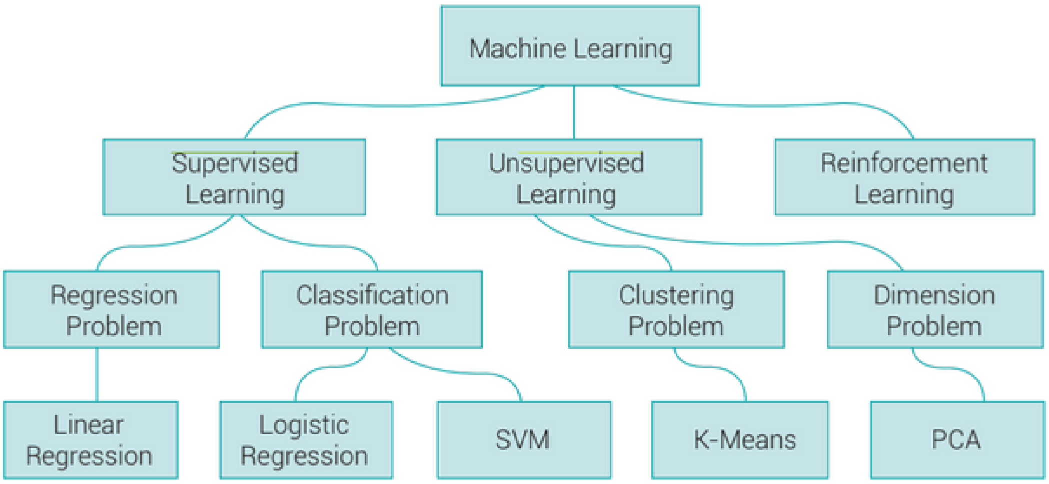


Figure 3.3: Different types of machine learning
[Prasad et al., 2020]

3.6.3.1 Supervised Learning :

In supervised learning, the training data (or the example data) which is a set of paired input-output training samples where the output is regarded as the label of the input data or the supervision [Qiong Liu,Ying Wu ; 2012] is used in teaching computers in order to build an artificial system that can learn the mapping between the input and the output. After the learning process the model is used with unseen data to predict the output.
There are two major problems of supervised learning: Classification or regression problems.

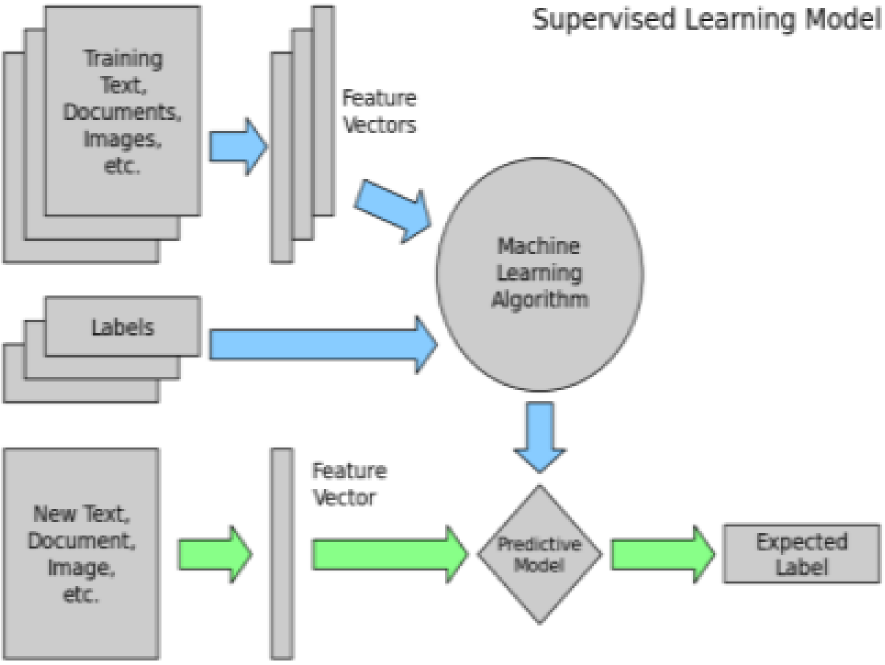

Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Figure 3.4: Supervised learning
[Rupali and Amit, 2017]

Classification versus regression algorithms

Regarding the type of supervised learning, we can further distinguish between classification and regression problems, Classification focuses on predicting categorical outcomes, while regression aims to predict continuous outcomes.Many machine learning algorithms that were developed to perform classification have been adapted to also address regression problems, and vice versa.[Janiesch et al., 2021]

1. **Classification :**
   is the process of finding a function which helps in dividing the dataset into classes based on different parameters. In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes. The following are some Classification algorithms :

   - **Decision tree :** One of the most popular algorithms used for classification. As the name implies, The algorithm generates a structure like a tree by classifying the instances which comprise several branches, root nodes, and leaf nodes ,and utilizing an (RPA) or recursive portioning algorithm. A class label is represented by a leaf node and the branches represent test results, these tests are represented by internal nodes for an attribute.[Aslam et al., 2019] Decision trees can be used in regression and classification analysis, and regression trees and classification trees are similar, the only difference is that the classification tree is used to predict a qualitative response and regression tree a quantitative one.[Himberg, 2021]

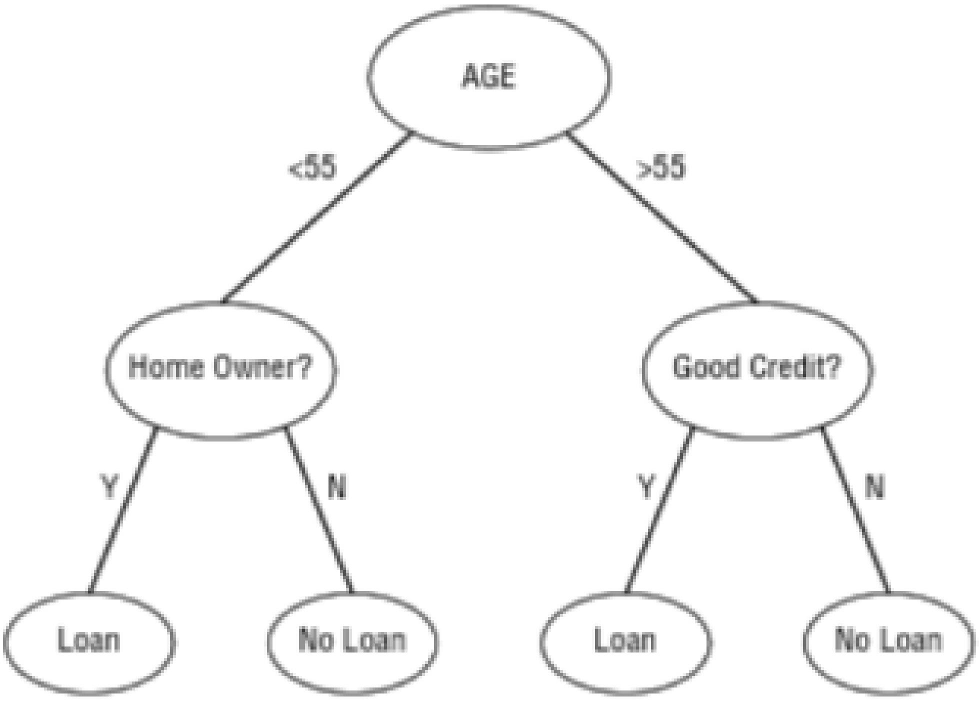


Figure 3.5: Decision Tree (Example)
[Himberg, 2021]

As shown in the figure a decision tree where “Age” is the root node, “Home owner?” and “Good credit?” are the decision nodes. And at the bottom of the tree, “Loan” and “No Loan” are the result value and they are contained in the leaves. The tree at first checks if the person applying for a loan is over or under 55 years old. If he or she is under 55 years old, the tree checks if the applicant is a home owner or not. If not, loan is not granted and if yes, loan is granted. The same decision is made for over 55 years old’s, based on good or bad credit.[Himberg, 2021]

- **Random Forest :** A Random Forest Algorithm is used for Classification and Regression problems in Machine Learning.[Madaan et al., 2021]

  The accuracy of a random forest depends on the strength of the individual tree classifiers and the measure of the dependence between them.[Addo et al., 2018]

  Random Forest is built with several decision trees that expand in randomly selected subsets of the given dataset[Madaan et al., 2021] , it takes the average to improve the predictive accuracy of that dataset. This classifier is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

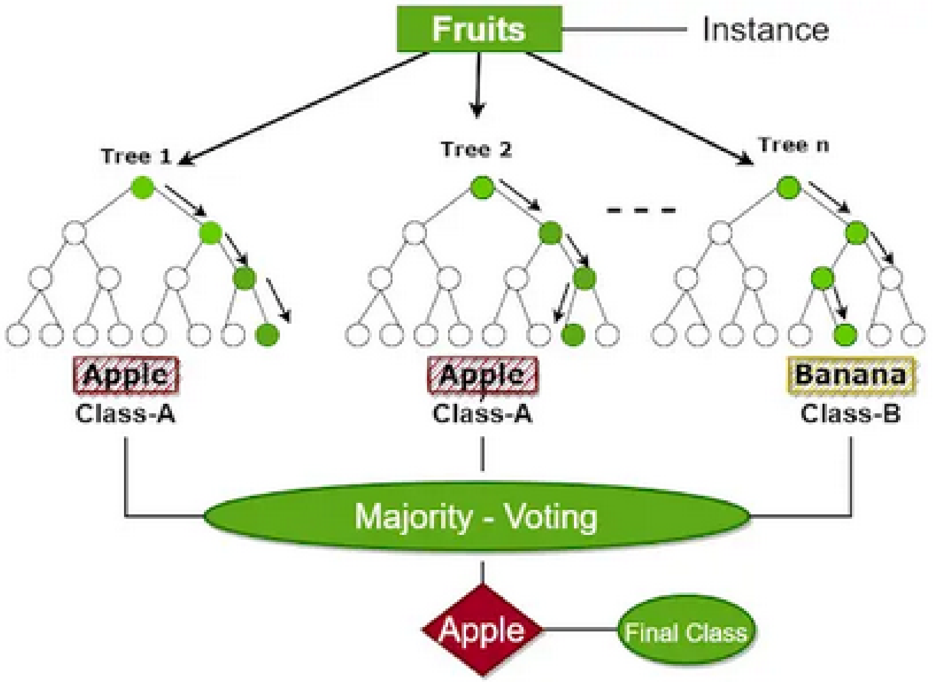


Figure 3.6: Random Forest Classifier
[?]

- **KNN (k-nearest neighbors)** The k-nearest neighbors (kNN) algorithm is a widely used algorithm in statistical pattern recognition. It has been applied in various domains, including credit scoring, where it is commonly used to construct predictive models.

  The fundamental concept behind the kNN algorithm is to predict the label of a new sample based on the labels of its k nearest neighbors. In other words, the algorithm identifies the k training samples that are closest to the new input sample and assigns the label that is most frequently represented among these k neighbors to the new sample.

  To measure the distance between the new sample and the previous training samples, the Euclidean distance metric is commonly used in kNN models. The Euclidean distance between two points in n-dimensional space is the length of the straight line that connects these two points.

  One of the benefits of the kNN algorithm is its simplicity. The only parameter that needs to be specified is the size of the neighborhood (i.e., the value of k). This makes it relatively easy to implement and understand. However, choosing the appropriate value of k is crucial in ensuring that the model achieves optimal performance.[Bao et al., 2019]

- **Support Vector Machines** Support Vector Machines (SVM) is a widely used supervised machine learning algorithm that was first introduced by Vapnik in 1995 [Bao et al., 2019], within the context of statistical learning theory.

(SVM) has emerged as a competitive classifier due to its successful applications in various pattern recognition tasks.The fundamental concept behind SVM is to project the input data into a high-dimensional feature space and then determine the optimal hyperplane supported by the support vectors. This hyperplane acts as the decision boundary, separating the two classes with a maximal margin. [Goh and Lee, 2019]

SVM predicts the label of a new input sample based on the characteristics of the support vectors. The flexibility of SVM is further enhanced by the availability of multiple kernel functions, such as linear and polynomial, which map the input data into the high-dimensional feature space.

SVM has been applied to numerous real-world problems and has demonstrated high accuracy and robustness, making it a valuable tool in the realm of machine learning research.[Bao et al., 2019]

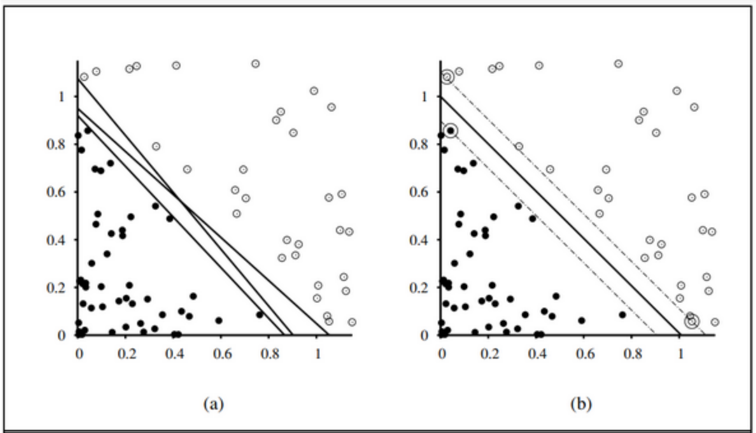


Figure 18.30 Support vector machine classification: (a) Two classes of points (black and white circles) and three candidate linear separators. (b) The maximum margin separator (heavy line), is at the midpoint of the **margin** (area between dashed lines). The **support vectors** (points with large circles) are the examples closest to the separator.

Figure 3.7: Support Vector Machine Classification
[Chen and Li, 2010]

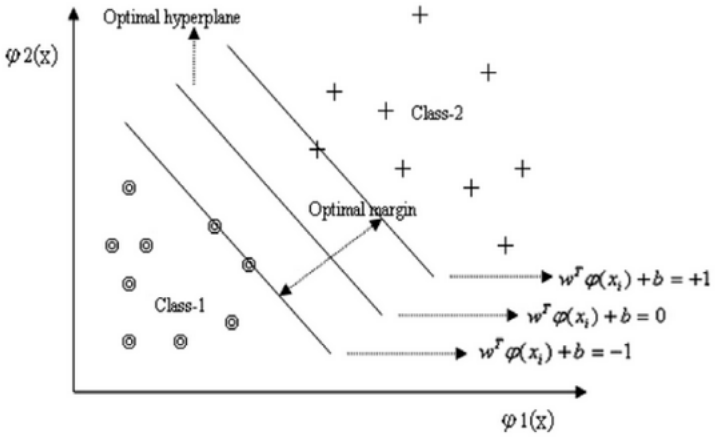


Fig. 2. An example of a separable problem in a two dimensional space.

Figure 3.8: An example of separale problem in a two dimensional space [Chen and Li, 2010]

- **Naive Bayes** The Naive Bayes classifier is a linear classifier that is reliable, fast, and accurate. It is based on the Bayesian theorem and focuses on conditional probability. The classifier is particularly suitable for high-dimensional inputs, as it assumes independence between features and attributes, simplifying computation. This assumption is considered naive, but despite its simplicity, the Naive Bayes classifier often outperforms alternative classifiers, especially for small sample sizes. This classifier is widely used in various fields. However, in practice, independence may be violated, and non-linear problems may result in poor performance of this classifier.[Chen et al., 2020]

  The Naive Bayes classifier works by calculating the prior probability for each class label, finding the likelihood probability for each class, and then using Bayes' formula to compute the posterior probability.

  Finally, the input is assigned to the class with the highest probability. This classifier is very effective in supervised learning environments and always uses the maximum likelihood method to estimate parameters.[Chen et al., 2020]

2. **Regression :** is the process of finding the correlations between dependent and independent variables.Used to predict the continuous variables such as prediction of Market Trends, House prices, weather forecasting, etc..

   - **Linear Regression :** is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques.[Kumari et al., 2018]

- **Logistic regression :** is one of the most popular statistical techniques among others in the financial world for credit risk assessment models. The strengths of a logistic regression model lie in its easy implementation, simple understanding and sturdy performance[Aslam et al., 2019]. The logistic regression model differs from linear regression as in logistic regression the outcome variable is binary.[Himberg, 2021]

  As [Aslam et al., 2019] explained, logistic regression performs better than linear regression as it overcomes multiple issues, such as in linear regression the output of the regression can be a negative value or greater than the value 1, which is not possible for probability. Logistic regression solves this by providing a continuous range of grades between 0 and 1 and keeping the output limited to values between 0 and 1.

3.6.3.2 Unsupervised Learning :

According to [Siadati et al., 2018] unsupervised learning can be defined as “a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision". Unsupervised learning differs from supervised learning as it has only input data, and no labels are given to the learning algorithm. [Rupali and Amit, 2017] As we know supervised learning uses output data so it can find a mapping from output data, but in the case of unsupervised there is no output data, and the model is left on its own to find hidden patterns and regularities from input.

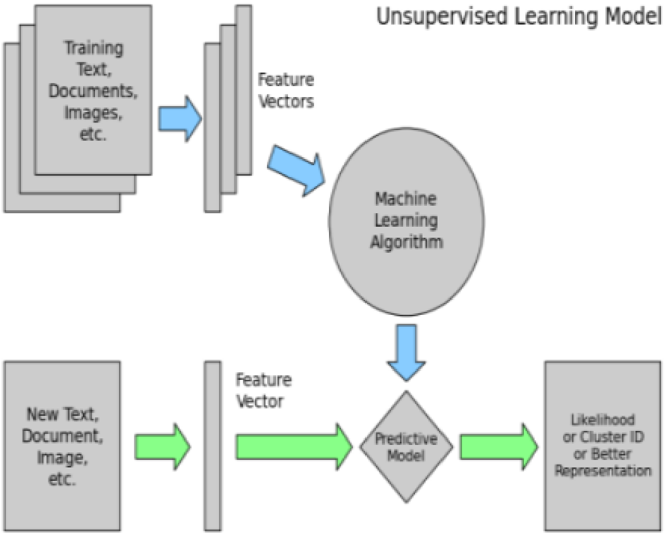

Unsupervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Likelihood or Cluster ID or Better Representation

Figure 3.9: Unsupervised learning
[Rupali and Amit, 2017]

There are four types of unsupervised tasks: Clustering, Principal component analysis, Anomaly detection, and Autoencoders[Dridi, 2021]

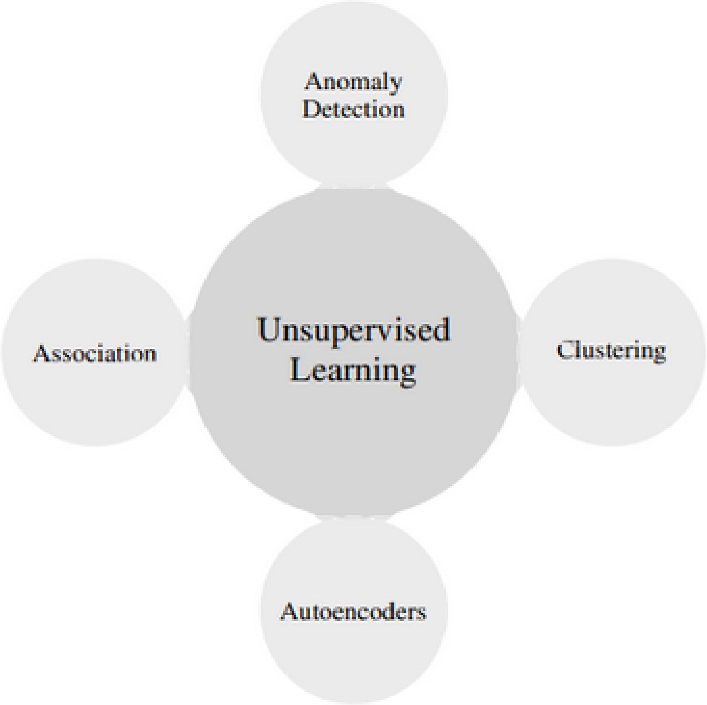


Figure 3.10: Types of Unsupervised Learning
[Dridi, 2021]

3.6.3.3 Reinforcement Learning :

Reinforcement learning (RL) systems represent a paradigm shift from traditional systems that rely on input-output pairs. Rather than providing such pairs, an RL system describes the current state of the system, specifies a goal, lists allowable actions and their environmental constraints, and allows the machine learning (ML) model to experience the process of achieving the goal via trial and error while maximizing a reward. RL models have proven to be successful in closed world environments, such as games , and have been applied to more complex multi-agent systems such as electronic markets . This demonstrates the potential for RL models to address more complex real-world problems. [Janiesch et al., 2021]

3.6.4 Machine Learning Process

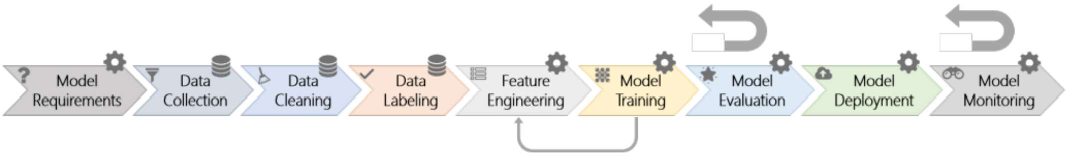


Figure 3.11: Machine Learning Process
[Amershi et al., 2019]

3.6.4.1 Model Requirements:

In this step designers figure out which features can be handled effectively with machine learning and which ones would be beneficial for a product. They also decide on the best types of models to use for the problem at hand[Amershi et al., 2019]. Data Collection :In the data collection phase, professionals either bring together and utilize existing datasets, including those obtained internally or from open-source repositories, or they embark on the process of collecting their own data as needed for their specific project [Amershi et al., 2019].

3.6.4.2 Data Cleaning :

Data cleaning involves a meticulous examination of the data to identify and subsequently remove any inaccurate, noisy, or erroneous records to, ensuring that the data is free from inconsistencies and anomalies[Amershi et al., 2019]

3.6.4.3 Data Labeling :

Data labeling assigns accurate labels to dataset records, a crucial step for supervised learning techniques. Labels can come from engineers, domain experts, or online crowd workers, enabling the training of effective machine learning models.[Amershi et al., 2019]

3.6.4.4 Feature Engineering :

Encompasses all the tasks and processes undertaken to extract and choose relevant and meaningful features for machine learning models[Amershi et al., 2019].it serves as an essential component. It not only accelerates data mining algorithms but also enhances predictive precision and contributes to improved model interpretability . [Kumar and Minz, 2014]
There are generally three types of feature engineering methods: embedded, wrapper and filter methods.

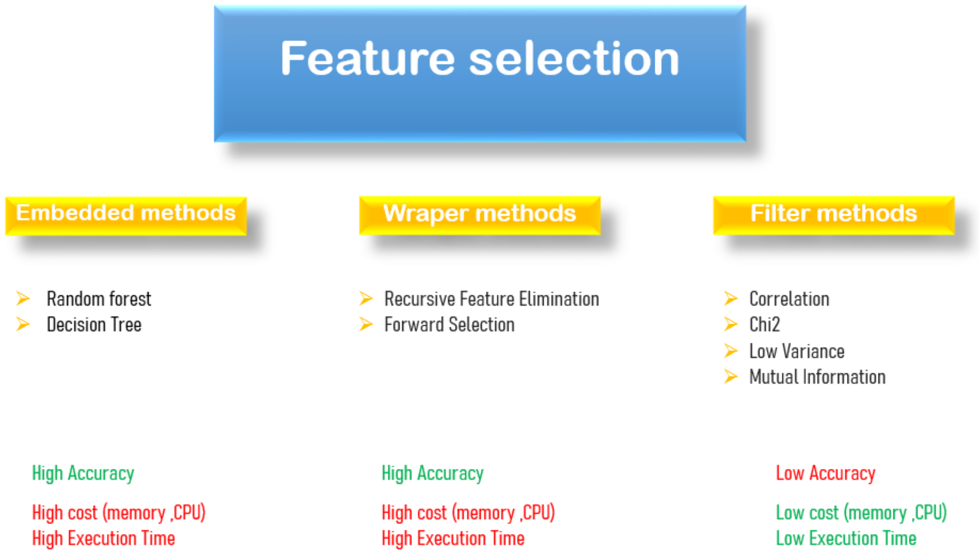


Figure 3.12: Feature Selection types

- **Embedded methods:**
  Embedded methods incorporate feature selection within the model training process. They optimize an objective function that balances classification accuracy with the use of fewer features, streamlining the selection of relevant attributes for improved model performance and efficiency.[Kotsiantis, 2011]

  - **Random Forest Importance:**
    Tree-based technique that assesses feature importance by measuring their impact on the target variable. It ranks features based on their performance and reduction in impurity across multiple decision trees within the Random Forest algorithm.[Garg, 2022]

- **Wrapper methods:**
  Use a classifier to assess feature subsets, training one machine learning model for every feature subset considered. While computationally intensive and dependent on the choice of classifier, they often yield high accuracy due to their tailored feature selection approach.[Kotsiantis, 2011]

  - **Recursive Feature Elimination (RFE):**
    is a step-by-step optimization technique that operates in a recursive and greedy manner. It progressively narrows down the feature set by iteratively selecting a smaller subset of features with each round of evaluation. .[Garg, 2022]
  - **Forward selection:**
    Forward selection is an iterative feature selection method that starts with no features

included. In each step, it adds one feature and assesses whether this addition enhances the model's performance. This process repeats until including a new feature no longer improves the model's performance. . .[Garg, 2022]

- **Filter methods:**
  Filter methods independently assess and rank features or feature subsets based on predetermined criteria. These methods are known for their speed and ease of interpretation, making them efficient for initial feature selection in large datasets.[Kotsiantis, 2011]

  - **Mutual Information :**
    In the context of feature selection, it helps assess the relevance of individual features with respect to the target variable. Higher mutual information values indicate a stronger relationship between a feature and the target variable, making the feature more likely to be selected for inclusion in a predictive model ,it aims to retain features that provide the most information or predictive power for the task at hand..
  - **Correlation:**
    Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features. .[Vishal, 2022]

3.6.4.5 Model Training:

During the model training stage, the selected machine learning models, utilizing the chosen features, undergo training and fine-tuning using the clean, collected data and their corresponding labels.[Amershi et al., 2019].

3.6.4.6 Model Evaluation:

the output model is assessed using pre-defined metrics on separate test or validation datasets, ensuring its performance and generalization.[Amershi et al., 2019].

- **Evaluation Metrics:**

  - **Confusion Matrix :**
    The confusion matrix is a widely used tool in the field of machine learning to evaluate the performance of a classification model on test datasets. It takes the form of a square matrix table that provides a comprehensive overview of all instances in a dataset [**?**], classified into four distinct categories. Its purpose is to display the model's classification outcomes, which include:

- ∗ **True Positives :** Goods that were classified correctly.
- ∗ **True Negatives :** Bads that were classified correctly.
- ∗ **False Positives :** Bads that were incorrectly classified as Goods.
- ∗ **False Negatives :** Goods that were incorrectly classified as Bads.

– **Accuracy :**

Accuracy is defined as the proportion of total instances that were classified correctly by the model[Fu et al., 2017]. It is calculated by dividing the number of correctly classified instances by the total number of instances . The resulting value ranges from 0 to 1, with higher values indicating better performance.[Schilling et al., 2019]

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

– **F1 Score :** The F1 score is a special case of the F score, where is set to 1.It is defined as the harmonic mean of precision and recall, It is widely used in statistical analysis as a classical metric for evaluating the performance of binary classification models.[Li et al., 2023]

$$\text{F1\_score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

– **Recall :** Recall, also referred to as sensitivity or true positive rate, is a measure of the classifier's ability to correctly identify positive instances. It represents the proportion of all actual positive instances that the classifier correctly identifies as positive[Schilling et al., 2019]

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

– **Precision :**It measures the accuracy of positive predictions made by a model it quantifies how well a model correctly identifies relevant instances (true positives) out of all instances it predicts as positive

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

- **ROC :** ROC curve is a graphical representation that plots the true positive (TP) rate which represents the proportion of correctly classified positive outcomes, against the false positive (FP) rate which represents the proportion of incorrectly classified negative outcomes, across a range of cut-off values. This trade-off between TP rate and FP rate is the main feature of the ROC curve, and is a useful means of evaluating the performance of a binary classification model. In general, more favorable models are those that have ROC curves situated towards the top-left hand corner of the (TP, FP) space.[Fu et al., 2017]

$$\mathrm{FPR} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}$$

$$\mathrm{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

3.6.4.7 Model Deployment:

The inference code of the model is deployed onto the designated device(s) and undergoes continuous monitoring for potential errors during its real-world execution.[Amershi et al., 2019].

3.7 Difference between ML and AI

A. Holzinger et al. discuss the long-standing tradition of artificial intelligence (AI) in computer science, back to 1950. During the initial three decades, there were high expectations for achieving human-level machine intelligence, but the expectations were too high, and the field experienced a setback known as the "AI winter" when these expectations couldn't be met. However, recent advancements in machine learning and knowledge extraction have sparked renewed interest in the field.[Holzinger et al., 2018]

The success of machine learning has been highlighted in prestigious journals like Science and Nature, with applications spanning various domains such as healthcare and manufacturing. Despite this progress, many scientists remain skeptical about the term "intelligence" due to its lack of clear definition, and the goal of human-level AI remains distant.

One common question is often asked is the difference between AI and machine learning, as well as the role of deep learning within them. A. Holzinger et al. answered this question in a simple way : Deep Learning is part of Machine Learning is part of Artificial Intelligence [Holzinger et al., 2018]. See figure 2.10 :

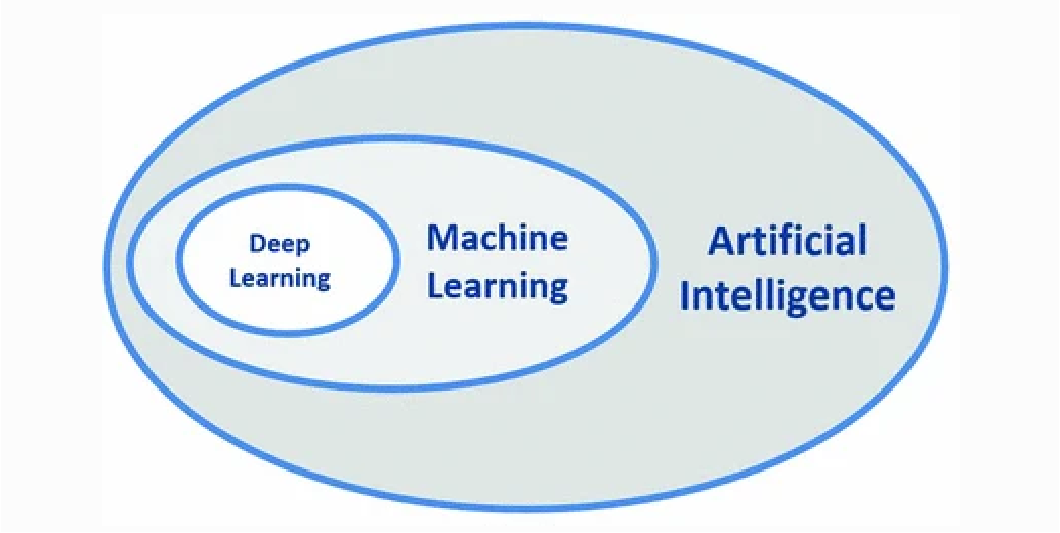


Figure 3.13: difference between AI, ML and DL
[Holzinger et al., 2018]

3.8 Conclusion

In the realm of Artificial Intelligence, understanding its history, types, and applications is vital. Additionally, Machine Learning serves as a powerful subset of AI, underlining its practical applications. The clear differentiation between AI and ML underscores their unique roles in shaping the future of technology and innovation.

Part III

State of the art

Chapter 4

State of the art

4.1 Introduction

In this chapter, we are going to talk about three main sections in order to have a clear and good idea about our work. These three main sections are :

- **Machine Learning based Approaches for Loan Approval Prediction :** In this section we will do a reviews literature of some articles and paper which talk about our theme.
- **Comparative table :** Then we will going to do a comparative table which contains different feature to compare between our reviews.
- **Synthesis :** Finally we end this chapter with some conclusions and results concluded from this reviews.

4.2 Machine Learning based Approaches for Loan Approval Prediction: State of the art

In this section we will present the various research works carried out on Loan Ap- proval Prediction. It contains about 14 articles talking about our the subject of our research :

1. [Trivedi, 2020] worked on Credit Scoring Models with different Feature Selection and Machine Learning Approaches.

   they started by collecting the data they used a publically available German Credit scoring data called corpora,to preprocess the collected data they used number of techniques such as cleaning, filtering, integration, discretization, and normalization.

   after that they conducted a comparative analysis of various combinations of classifiers and feature selection techniques they used three distinct feature selection techniques, namely Chi-Square, Information-gain, and Gain-Ratio, and five machine learning classifiers, including Bayesian, Naïve Bayes, SVM (support Vector Machine), Decision Tree, and Random Forest., Decision Tree, and Random Forest).

   The performance of these models was evaluated using various metrics such as False Positive rate, F-Measure, and Training time. Based on this analysis, three significant observations were derived.

F-Measure (in%)	Chi-Square	Gain-Ratio	Info-Gain
Bayesian	72.70	72.70	72.70
NB	78.30	78.30	78.30
SVM	75.70	74.90	75.70
J48	72.80	72.80	72.80
RF	79.20	80.10	80.70

Table 4.1: F-Measure of all classifiers

FP Rate (in%)	Chi-Square	Gain-Ratio	Info-Gain
Bayesian	21.90	21.90	21.90
NB	20.00	20.00	20.00
SVM	20.60	20.70	20.60
J48	27.10	27.10	27.10
RF	19.80	18.20	17.80

Table 4.2: False Positive Rate of all classifiers

Time Taken (Sec)	**Chi-Square**	**Gain-Ratio**	**Info-Gain**
Bayesian	0.08	0.04	0.03
NB	0.06	0.02	0.02
SVM	0.87	0.62	0.61
J48	0.24	0.17	0.17
RF	2.38	2.52	2.69

Table 4.3: Training Time of all classifiers

The results of this study indicate that utilizing a Random Forest classifier in combination with Information Gain feature selection is a favorable choice for developing a credit scoring model that is robust, accurate, and sensitive.

2. [Zhang et al., 2020] have worked on Loan Default Forecast using Random forest classifier. This study uses the idea of non-equilibrium data classification to statistically analyze the loan data and then establish a random forest model using the Sklearn-ensemble-Random Forest Classifier in Python. The loan default data set used in this article named "Give Me Some Credit" is from the Kaggle data science competition platform.

   They used the AUC (Area under the ROC curve) value to evaluate the model. [AUC is defined as the area under the ROC (Receiver Operating Characteristic) curve].Where the horizontal axis of the ROC curve is False Positive Rate (FPR), the vertical axis is True Positive Rate (TPR), and since the ROC curve is generally above the line y = x,The AUC value ranges between 0.5 and 1. And as the ROC curve does not always clearly indicate which classifier works better, the AUC value is used for evalutaion. The random forest model is compared with the logistic regression classification model and the decision tree classification model. The following table shows the results of the comparison :

Algorithm	Accuracy
Random Forest	0.86
Decision Tree	0.80
Logistic Regression	0.80

Table 4.4: Comparaision of Random Forest and other algorithms

   And to get the importance of each feature, This experiment uses the feature importance method of sklearn-ensemble-Random Forest Classifier as the following table indicates :

Variables	feature_importance_
Revolving Utilization Of Unsecured Lines	0.3411
NumberOfTime30-59DaysPastDueNotWorse	0.1694
NumberOfTime90DaysLate	0.1594
NumberOfTime60-89DaysPastDueNotWorse	0.0727
age	0.0677
Debt Ratio	0.0625
Monthly Income	0.0488
Number Of Open Credit Lines And Loans	0.0442
Number Real Estate Loans Or Lines	0.0223
Number Of DEpendents	0.0117

Table 4.5: Variable Importance

We can see from the above table that we need to pay special attention to these three characteristics of the borrower's total loan-to-credit ratio, the number of overdue 30–59 days in the past two years and the number of overdue over 90 days in the past two years when processing the loan application. This experiment shows that the random forest algorithm has better classification performance than the decision tree and logistic regression model.

3. [Ahmed and Rajaleximi, 2019] An Empirical Study on Credit Scoring and Credit Scorecard for Financial Institutions The main objective of this paper is to study the credit risk management using credit scoring strategies. According to this article we can define credit scoring as followed : " Credit scoring or credit rating computes the probability of repaying the loan for the customers based on their credit history or payment history along with their background history. With the calculated points or scores, the lenders can identify the ability of the customer's creditworthiness ". The credit risks have a few characteristics which are :

   - Expected Loss (EL)
   - Unexpected Loss (UL)
   - Probability of Default (PD)
   - Exposure at Default (EAD)
   - Loss Given Default (LGD)
   - Capital Adequacy

   CREDIT RISK MANAGEMENT There are several strategies in managing credit risks, Credit scoring is the most widely used technique. In measuring the risks some criterias are to be considered, and they are :

- Credit history
- Capacity to repay
- Capital
- Outstanding Debts
- Types of credit
- Loan's conditions
- Associated collateral

The credit scoring system collects information about the borrowers either internally from the organization and financial institution or externally from the other agencies. These collected details may include the past history of bill payment, number of accounts, transactions encountered, age of the account, current debts,then points are allocated for each information represented as parameters based on their significance, and if it reaches the given threshold value the applications will be accepted or else the applications get rejected.

CREDIT SCORECARD

The scorecards are created based on analyzing the past records. We can distinguish various types of scorecards based on the type of data given as an input and the type of information collected as an output. These are some of the important scorecards :

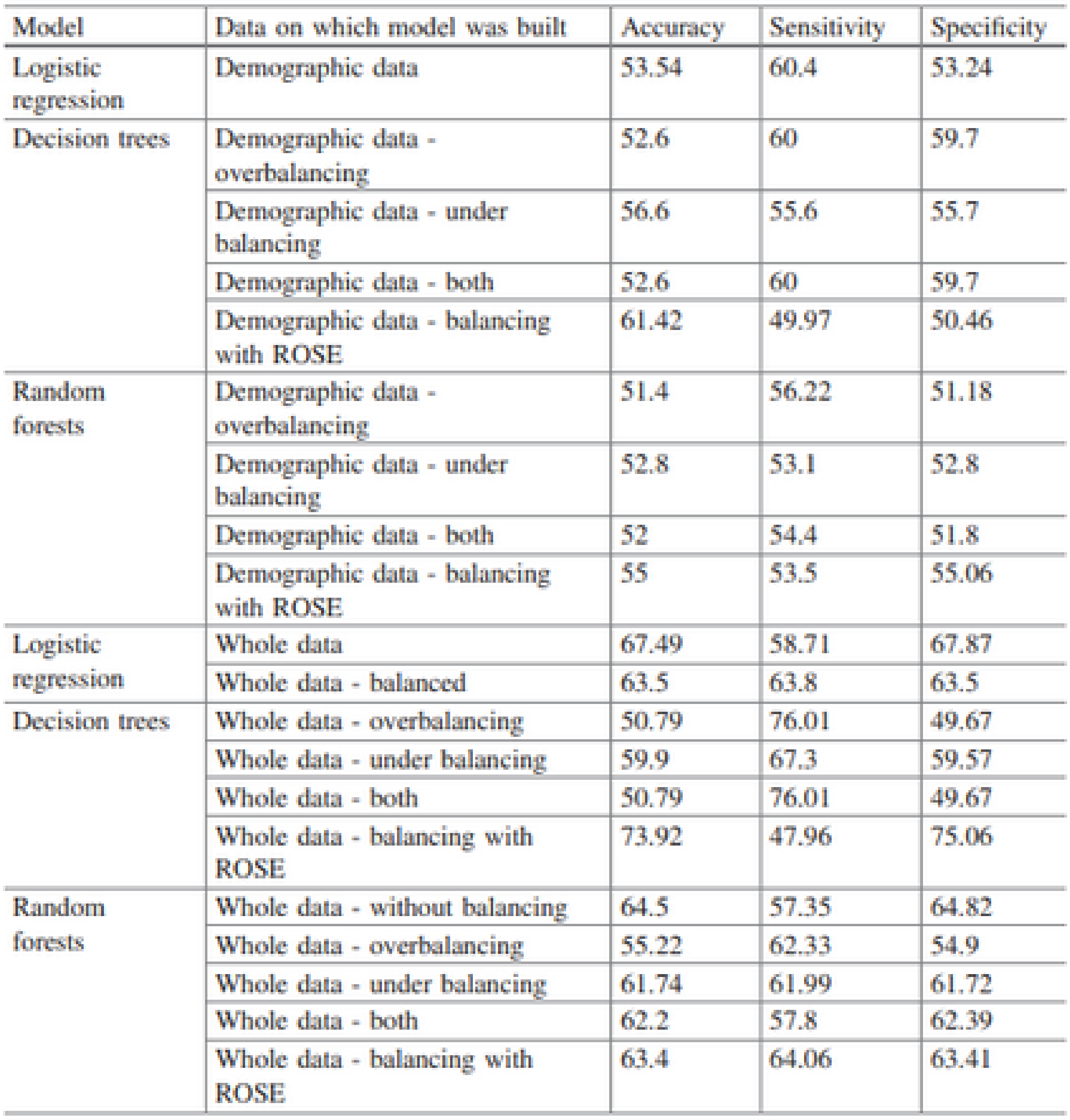
- Application Scorecards
- Behavioral Scorecards
- Collection Scorecards
- Fraud Scorecards
- Bankruptcy Scorecard
- Desertion Scorecard
- Profit and Non-profit Scorecards

We can see that the scorecards to be implemented can be chosen based on the application, and there are several methods to implement the scorecards, such as Discriminant analysis, Linear regression model, Probit and Poisson model, decision trees, and machine learning techniques.

4. [Kovvuri and Cheripelli, 2020] Have worked on Credit Risk Valuation Using an Efficient Machine Learning Algorithm. The goal is to calculate the credit score and categorize customers into good or bad In this paper Two data sets are required for the analysis, Demographic data and Credit bureau data. Demographic Data: Demographic data has simple variables. Credit Bureau data: Credit bureau data has variables obtained from previous history of the customer. After the data cleaning and exploratory Data Analysis, WOE (Weight of Evidence Analysis) analysis on the data has been performed and replaced demographic and credit data with WOE values.

   Model Building

   Several models were built using Logistic Regression, Decision Trees and Random Forests and best for this data was picked.



Model	Data on which model was built	Accuracy	Sensitivity	Specificity
Logistic regression	Demographic data	53.54	60.4	53.24
Decision trees	Demographic data - overbalancing	52.6	60	59.7
	Demographic data - under balancing	56.6	55.6	55.7
	Demographic data - both	52.6	60	59.7
	Demographic data - balancing with ROSE	61.42	49.97	50.46
Random forests	Demographic data - overbalancing	51.4	56.22	51.18
	Demographic data - under balancing	52.8	53.1	52.8
	Demographic data - both	52	54.4	51.8
	Demographic data - balancing with ROSE	55	53.5	55.06
Logistic regression	Whole data	67.49	58.71	67.87
	Whole data - balanced	63.5	63.8	63.5
Decision trees	Whole data - overbalancing	50.79	76.01	49.67
	Whole data - under balancing	59.9	67.3	59.57
	Whole data - both	50.79	76.01	49.67
	Whole data - balancing with ROSE	73.92	47.96	75.06
Random forests	Whole data - without balancing	64.5	57.35	64.82
	Whole data - overbalancing	55.22	62.33	54.9
	Whole data - under balancing	61.74	61.99	61.72
	Whole data - both	62.2	57.8	62.39
	Whole data - balancing with ROSE	63.4	64.06	63.41

Figure 4.1: Result of analysis using different models on data

The best model was chosen by looking at the three metrics : Accuracy, Sensitivity and Specificity. Some models gave +70 accuracy but didn't perform well in Sensitivity. Finally it was the choice between Logistic Regression and Random Forest models because they have equal numbers for all the three parameters. Chosen Random Forests because of two reasons: Sensitivity is slightly more compared to logistic regression, and as we know Random Forests will perform good on unseen data. So Random Forest with balanced data is the Final Model.

5. [Madaan et al., 2021] This paper does a comprehensive and comparative analysis between two algorithms Random Forest, and Decision Trees. They found that using Random Forest over other machine learning algorithms has so many advantages like:

- Immunity to overfitting.
- Accurate classification or regression.
- More efficient on large databases.

Dataset Description The dataset used in this paper is the publicly available Lending Club dataset from Kaggle. The data covers approximately 22 lakh loans funded by the platform between 2007 and 2015.

Proposed Model Which was in four steps:

Project Pre-Work :

where an exploratory data analysis (EDA) has been done of the given data to examine its features and answer some questions like : What are the characteristics of each loan? What features make them different or similar? etc..

Data Cleaning :

In order to get rid of null values in the dataset before moving to EDA.

Exploratory Data Analysis :

Exploratory Data Analysis (EDA) played an integral part in understanding the Lending Club dataset. This section investigates data distribution and asks specific questions about the data lying within the dataset. So many plots have been done such as : Count plot to show the number of values in each Loan status category, Count plot for Loan Grades showing the number of values in each grade, Violin plot of Loan Amount vs Home Ownership type, Box plot of Loan Amount vs Loan purpose, etc…

Modelling :

they used 2 algorithms for the modeling purpose: Decision Tree The Decision Tree classifier gave an accuracy score of 73%.

	Precision	Recall	F1 score	Support
0	0.82	0.85	0.83	312588
1	0.29	0.24	0.26	78504
Accuracy			0.73	391092
Macro Avg	0.55	0.55	0.55	391092
Weighted Avg	0.71	0.73	0.72	391092

Table 4.6: Classification Report for Decision Tree

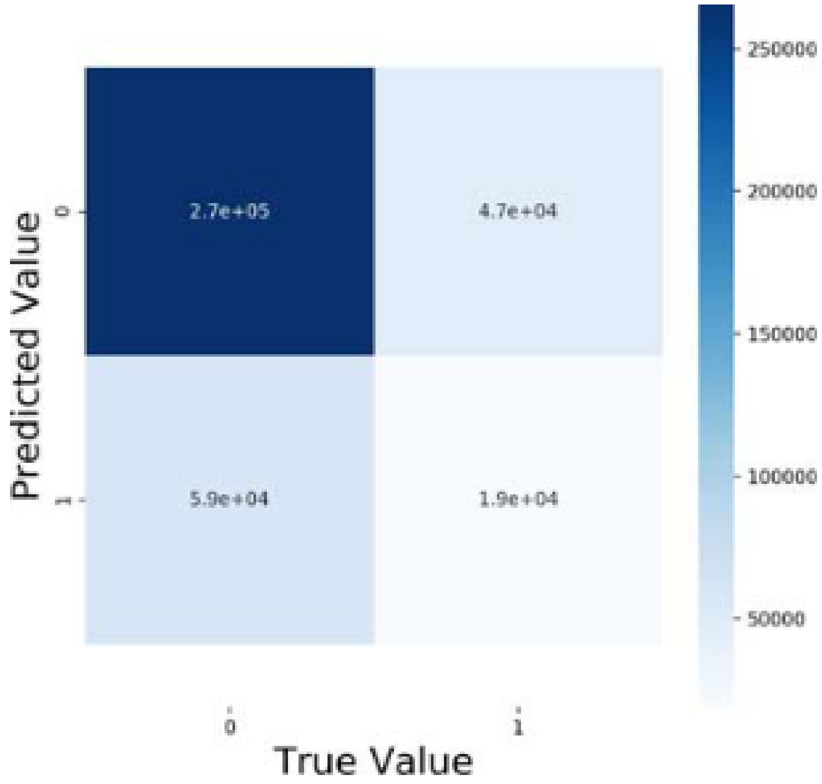


Figure 4.2: Confusion Matrix of Decision Tree
[Madaan et al., 2021]

Random Forest The Random Forest Classifier gave an accuracy score of 80%.

	Precision	Recall	F1 score	Support
0	0.81	0.98	0.89	312588
1	0.50	0.08	0.14	78504
Accuracy			0.80	391092
Macro Avg	0.65	0.53	0.51	391092
Weighted Avg	0.75	0.80	0.74	391092

Table 4.7: Classification Report for Random Forest

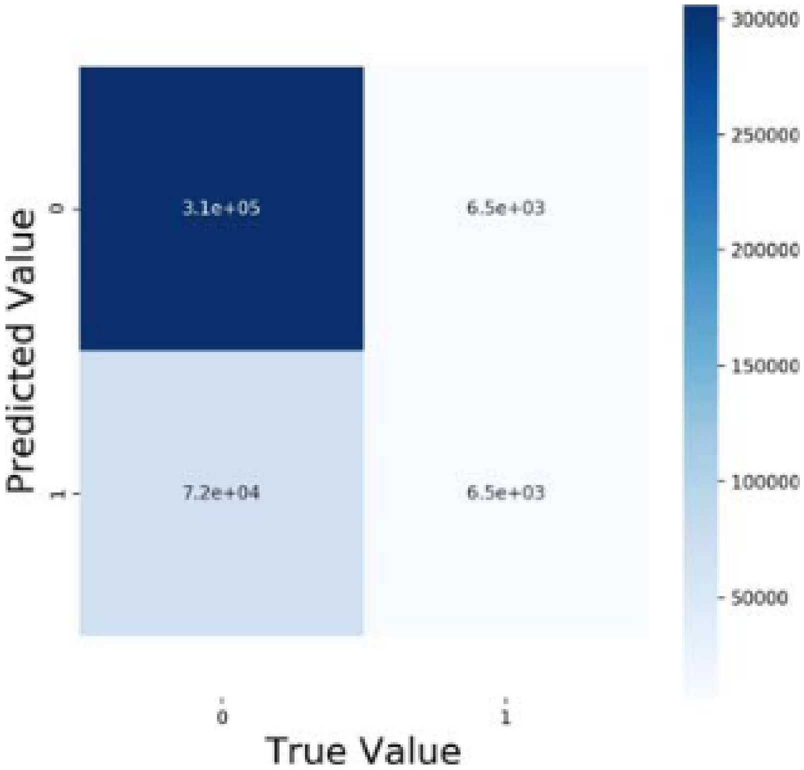


Figure 4.3: Confusion Matrix of Random Forest
[Madaan et al., 2021]

By looking at the above confusion matrices and classification reports for both the models, we can say that the Random Forest algorithm is a much better option over Decision Trees for loan prediction on the given dataset.

6. [Anand et al., 2022] have worked on Prediction of Loan Behaviour with Machine Learning Models for Secure Banking using a variety of classification methods (15 Classification algorithms) including Multiple Logistic Regression, Decision Tree, K-Nearest Neighbors, SVM, Random Forest, and other forms of Ensemble Boosting approaches and then show the results of top 5 algorithms which showed optimistic results. They used seven metrics for the evaluation: Confusion Matrix, Accuracy, F1- Score, Recall, Precision, ROC area and Feature Importance.

   Confusion Matrix :

   The tables below , displays the computed confusion matrix of the top five algorithms as shown in figure 3.3 :

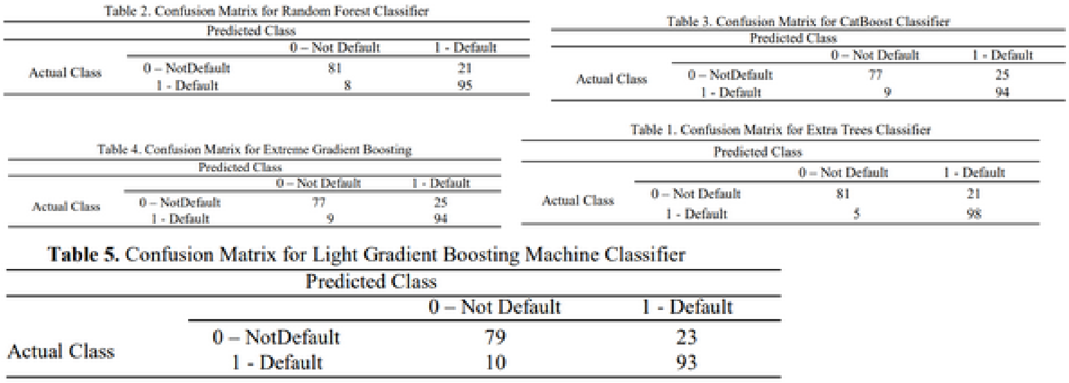


Table 2. Confusion Matrix for Random Forest Classifier

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – NotDefault	81	21
	1 - Default	8	95

Table 3. Confusion Matrix for CatBoost Classifier

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – NotDefault	77	25
	1 - Default	9	94

Table 4. Confusion Matrix for Extreme Gradient Boosting

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – NotDefault	77	25
	1 - Default	9	94

Table 1. Confusion Matrix for Extra Trees Classifier

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – Not Default	81	21
	1 - Default	5	98

Table 5. Confusion Matrix for Light Gradient Boosting Machine Classifier

		Predicted Class	
		0 – Not Default	1 - Default
Actual Class	0 – NotDefault	79	23
	1 - Default	10	93

Figure 4.4: Confusion Matrix of the top five algorithms

Accuracy :

The used formula to define it is :

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Algorithm	Accuracy
Extra Trees Classifier Forest	86.17
Random Forest Classifier Tree	85.55
CatBoost Classifier	84.92
Light Gradient Boosting	84.49
Extreme Gradient Boosting	83.87

Table 4.8: Accuarcy of top 5 algorithms in descending order

Recall :

Algorithm	Recall
Extra Trees Classifier Forest	88.20
Random Forest Classifier Tree	87.35
CatBoost Classifier	85.59
Light Gradient Boosting	84.68
Extreme Gradient Boosting	82.91

Table 4.9: Recall of top 5 algorithms in descending order

Precision :

Algorithm	Recall
Extra Trees Classifier Forest	85.06
Random Forest Classifier Tree	84.27
CatBoost Classifier	83.93
Light Gradient Boosting	83.46
Extreme Gradient Boosting	82.57

Table 4.10: Precision of top 5 algorithms in descending order

F1-Score :

Algorithm	Recall
Extra Trees Classifier Forest	85.97
Random Forest Classifier Tree	85.03
CatBoost Classifier	84.57
Light Gradient Boosting	83.83
Extreme Gradient Boosting	82.07

Table 4.11: F1-Score of top 5 algorithms in descending order

Feature Importance :

The top five algorithms have the most significant characteristics plotted out. The plot's conclusion is heavily influenced by the customer's job history and loan income.

ROC :

The top five algorithms are more accurate because they have greater AUCs.

As a conclusion this article uses predictive modeling to detect problematic clients among a large number of loan applicants, resulting in a more effective basis for loan credit approval and found that the most crucial parameters are customer's employment or job experience in years and debt income.

7. [Wu, 2022] In this paper they apply Random Forest and XGBoost algorithms to train the prediction model and compare their performance in prediction accuracy. The framework of this experiment is shown in the picture below:

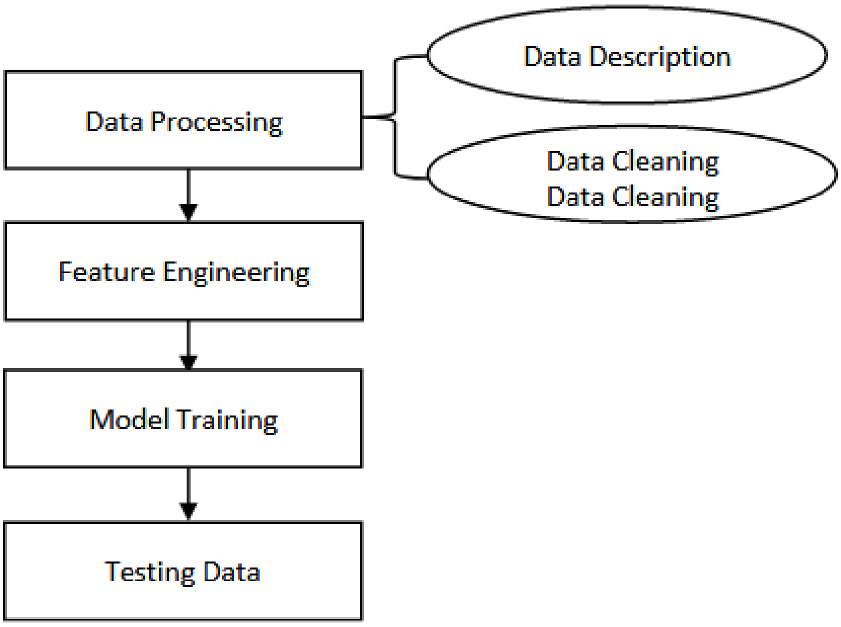


Figure 4.5: Framework of the experiment
[Wu, 2022]

Data Processing :

The data used is provided by Imperial College London includes 105,471 records and 771 columns, containing customers' ids, 778 features. and the loss of the record. They cleaned the NAs in the dataset. And filled those NAs with the average of the columns. **Feature Engineering :**

Feature selection aims to drop those redundant columns which contain little useful information and reduce the number of input features when developing an effective model.

At first, they used the variance threshold method to filter out those columns whose variance equals 0, because those columns do not contain useful information for classification. Then, they applied the Variance Inflation Factor method to reduce multicollinearity. Multicollinearity inflates unnecessarily the standard errors of the coefficients, and increased standard errors indicates that the coefficients of some features might be close to 0, and that will make some features insignificant when they should be significant. A useful way to measure multicollinearity is Variance Inflation Factor (VIF). If no features are correlated, the VIF will be 1. If the VIF is greater than 10. In this paper, They remove all those features whose VIF is greater than 10. After filtering only 419 columns are left. Model Training Testing Data To train the model they apply Random Forest and XGBoost algorithms and compare their performance in prediction accuracy.

Random Forest : The prediction accuracy of the Random Forest model is **0.90657**

XGBoost : The prediction accuracy of the XGBoost model is **0.9063**

Conclusions Discussion The result indicates that Random Forest and XGBoost show little difference in the accuracy of their predictions, and both get high accuracy in the loan default cases.

8. [Sayjadah et al., 2018] have worked on Predicting the Credit Defaulters using Machine Learning Techniques

their goal in this study was to ensure healthy lending practices on financial institutions developing a model that could accurately identify borrowers who were at the highest risk of defaulting on their loans

They began by collecting historical Credit data from the UCI repository, which they then processed and cleaned to ensure its quality. They further used sampling techniques to balance the data and create a representative dataset for further analysis.

Next, the team applied various machine learning algorithms to both the original and feature-selected datasets to compare their performance.

To evaluate the effectiveness of the models, the team used several metrics such as classification accuracy (Accuracy), recall (Recall), and precision (Precision), as presented in the tables above.

Model	Accuracy	Precision	Recall
Extra Tree Classifier	0.94	0.99	0.90
Decision Tree	0.86	0.98	0.73
Random Forest	0.95	1.00	0.89
Gradient Boosting)	0.9	0.99	0.81

Table 4.12: Accuracy of algorithms before feature selection

Model	Accuracy	Precision	Recall
Extra Tree Classifier	0.95	0.98	0.91
Decision Tree	0.89	1.00	0.77
Random Forest	0.97	1.00	0.94
Gradient Boosting)	0.91	1.00	0.80

Table 4.13: Performance of algorithms after feature selection

The study results indicate that feature selection has improved the performance of machine learning algorithms in predicting defaulters Furthermore, the RandomForest algorithm has outperformed the other models in identifying defaulters.

9. [Zhou et al., 2019] have studied default prediction from high-dimensional data based on machine learning

   The paper proposes the utilization of machine learning algorithms to effectively manage and control the default risk associated with P2P lending with the ultimate goal of enhancing the platform's risk control capabilities and solving the challenge of predictions based on high-dimensional and imbalanced data. This research proposes a heterogeneous ensemble learning-based default prediction model to estimate the probability of customers defaulting. To achieve this, the researchers integrated three individual classifier models - Gradient Boosting Decision Trees (GBDT), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). They used a learning-based feature ranking technique to process credit data and conducted hyperparameter optimization on each classifier to improve the model's accuracy.also they applied a linearly weighted integration strategy that is simple yet highly efficient for predicting customer default probability.

   Additionally, the researchers tested several well-known machine learning models as baselines to compare with the proposed model. These included neural network (NN), logistic regression(LR), random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN) and AdaBoost (adaptive boosting). These algorithms were further validated using an imbalanced, high-dimensional, sparse P2P lending credit dataset and evaluated using four different Evaluation metrics namely The area under the curve (AUC), specificity (SPE), sensitivity (SEN) and F1 Score. Table 3.13 provides a summary of the different results obtained from these algorithms.

Models	AUC Score	SEN	SPE	F1-Score
Ensemble model	0.7185	0.9596	0.1589	0.8615
GBDT	0.7002	0.9996	0.0033	0.8523
XGBoost	0.7162	0.8087	0.4702	0.8085
LightGBM	0.7029	0.9981	0.0166	0.8548
NN	0.6420	0.5133	0.6754	0.6184
LR	0.6265	0.7198	0.5331	0.7572
RF	0.6349	0.9863	0.0298	0.8520
SVM	0.6528	0.9989	0.0033	0.8520
KNN	0.5899	0.9977	0.0033	0.8514
AdaBoost	0.6755	0.6401	0.6192	0.7100

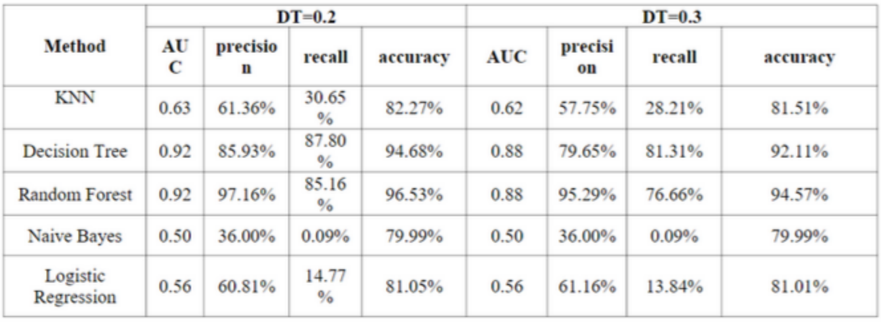
Table 4.14: Predicting Performance between default prediction model and benchmark models

The study's results indicate that the proposed model outperforms existing benchmark models and effectively addresses challenges related to high-dimensional and imbalanced data in default predictions.

10. [Wang et al., 2020] (Worked on A Comparative Assessment of Credit Risk Model Based on Machine Learning

The main objective of this paper is to compare the performance of five widely-used classifiers in machine learning for credit scoring: Naive Bayesian Model, Logistic Regression Analysis, Random Forest, Decision Tree, and K-Nearest Neighbor Classifier.

In the conducted comparison, the researchers utilized a dataset from a commercial bank. To safeguard the privacy of individuals, the data was desensitized, and the feature names were replaced with characters. After preprocessing the data, they applied to the five algorithms . The performance was measured using accuracy, Precision ,AUC (area under curve)and recall.



Method	DT=0.2				DT=0.3			
	AUC	precision	recall	accuracy	AUC	precision	recall	accuracy
KNN	0.63	61.36%	30.65%	82.27%	0.62	57.75%	28.21%	81.51%
Decision Tree	0.92	85.93%	87.80%	94.68%	0.88	79.65%	81.31%	92.11%
Random Forest	0.92	97.16%	85.16%	96.53%	0.88	95.29%	76.66%	94.57%
Naive Bayes	0.50	36.00%	0.09%	79.99%	0.50	36.00%	0.09%	79.99%
Logistic Regression	0.56	60.81%	14.77%	81.05%	0.56	61.16%	13.84%	81.01%

Figure 4.6: Models Performance
[Wang et al., 2020]

The experimental results indicate that Random Forest outperforms the other classifiers in terms of precision, recall, AUC (area under curve), and accuracy.

11. [Namvar et al., 2018] worked on Credit risk prediction in an imbalanced social lending environment This study aimed to address the issue of imbalanced dataset in creditworthiness evaluations. To achieve this, they went through five main steps :

collecting data : they used data from publicly available datasets released by Lending Club from 2016 to 2017
Feature Engineering : They cleaned the data and applied data transformation,leaky data removal and correlation analysis.
Imbalanced Learning Approaches: three categories of resampling approach were used

- **The under-sampling approach :** random under-sampling (RUS), and instance hardness threshold(IHT) algorithms.
- **Over-sampling approach :** Random over-sampling (ROS), synthetic minority over-sampling technique (SMOTE)), and adaptive synthetic sampling (ADASYN).
- **Hybrid approach :** Tomek links (SMOTE-TOMEK) and SMOTE+ edited nearest neighbor (SMOTE-ENN).

Classification Models : three algorithms were selected :

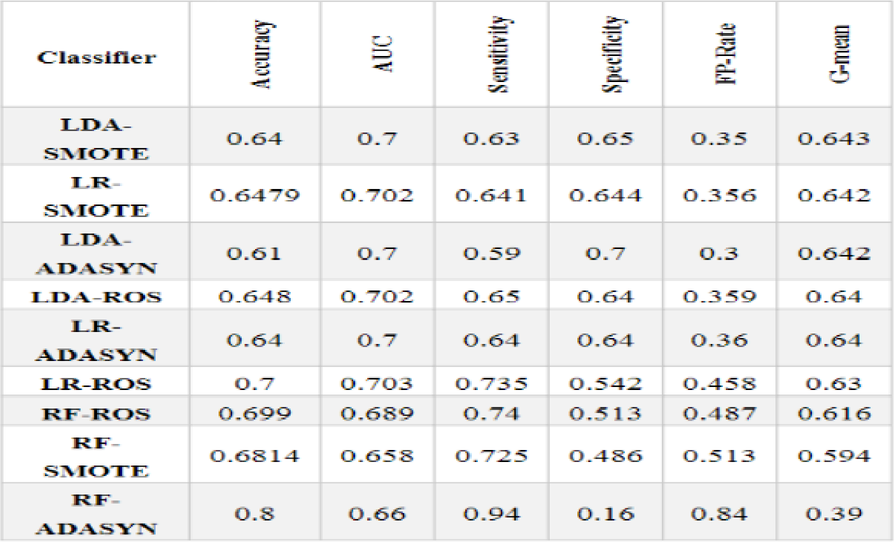
- Random Forest
- Linear Discriminate Analysis
- Logistic Regression

Performance Measurement : In order To evaluate the performance of the algorithms, they used various metrics , such as Accuracy, AUC, Sensitivity, Specificity, and FP-Rate, However, their primary focus was on the G-mean metric .

The study tested the credit risk prediction capabilities of the three classifiers using different resampling methods. The classifiers and resampling techniques were evaluated in groups, and the most effective pair from each group was selected . Finally, the best-performing pairs were compared with each other and with a non-sampling approach as shown in the tables :

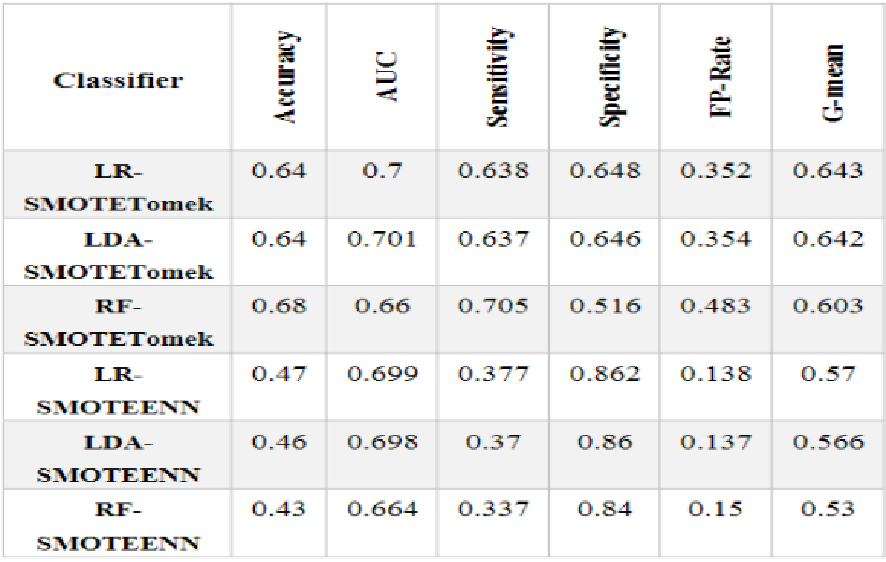
Classifier	Accuracy	AUC	Sensitivity	Specificity	FP-rate	G-mean
RF-RUS	0.692	0.69	0.717	0.582	0.42	0.65
LR-RUS	0.693	0.71	0.723	0.558	0.442	0.635
LDA-RUS	0.676	0.7034	0.695	0.589	0.42	0.64
LR-IHT	0.71	0.7	0.76	0.51	0.49	0.62
LDA-IHT	0.713	0.7	0.759	0.505	0.494	0.619
RF-IHT	0.75	0.688	0.83	0.4	0.61	0.51

Table 4.15: Classification Result (Under-Sampling approach)



Classifier	Accuracy	AUC	Sensitivity	Specificity	FP-Rate	G-mean
LDA-SMOTE	0.64	0.7	0.63	0.65	0.35	0.643
LR-SMOTE	0.6479	0.702	0.641	0.644	0.356	0.642
LDA-ADASYN	0.61	0.7	0.59	0.7	0.3	0.642
LDA-ROS	0.648	0.702	0.65	0.64	0.359	0.64
LR-ADASYN	0.64	0.7	0.64	0.64	0.36	0.64
LR-ROS	0.7	0.703	0.735	0.542	0.458	0.63
RF-ROS	0.699	0.689	0.74	0.513	0.487	0.616
RF-SMOTE	0.6814	0.658	0.725	0.486	0.513	0.594
RF-ADASYN	0.8	0.66	0.94	0.16	0.84	0.39

Figure 4.7: Classification Result (Over-Sampling approach). [Namvar et al., 2018]



Classifier	Accuracy	AUC	Sensitivity	Specificity	FP-Rate	G-mean
LR-SMOTETomek	0.64	0.7	0.638	0.648	0.352	0.643
LDA-SMOTETomek	0.64	0.701	0.637	0.646	0.354	0.642
RF-SMOTETomek	0.68	0.66	0.705	0.516	0.483	0.603
LR-SMOTEENN	0.47	0.699	0.377	0.862	0.138	0.57
LDA-SMOTEENN	0.46	0.698	0.37	0.86	0.137	0.566
RF-SMOTEENN	0.43	0.664	0.337	0.84	0.15	0.53

Figure 4.8: Classification Result (Hybrid approach). [Namvar et al., 2018]

Classifier	Accuracy	AUC	Sensitivity	Specificity	FP-rate	G-mean
RF-RUS	0.692	0.69	0.717	0.582	0.42	0.65
LDA-SMOTE	0.64	0.7	0.63	0.65	0.35	0.643
LR-Tomek	0.64	0.7	0.638	0.648	0.352	0.643
Logistic Regression	0.8173	0.703	0.988	0.048	0.95	0.218
Random Forest	0.8176	0.696	0.996	0.015	0.98	0.12

Table 4.16: Classification Result (Final Comparison)

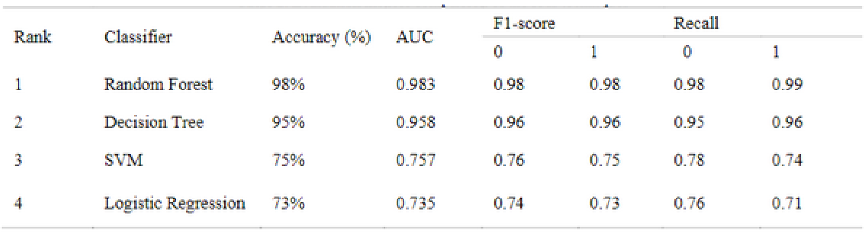
Results of the experiments proved the effectiveness of sampling techniques on the performance on prediction models and showed that the combination of random forest and random under-sampling may be an efficient approach for computing risk scores of loan applicants in lending markets.

12. [Zhu et al., 2019] Worked on A study on predicting loan default based on the random forest algorithm In this study they used the Random Forest algorithm to construct a loan default prediction model with the aim to enhance the loan evaluation and promote the growth and healthy development of the lending process.

    The data set used in this paper was collected from Lending Club for the first quarter of 2019, To address the problem of imbalance class in the dataset they used the SMOTE method and then they performed several operations including data cleaning and dimensionality reduction.

    They tested several well-known machine learning models as baselines to compare with the RF model. In particular, Decision Tree, SVM and Logistic Regression

    In order To evaluate the performance of these algorithms, they mainly focused on accuracy and AUC. Table 4.14 shows us evaluation metrics of four techniques :



Rank	Classifier	Accuracy (%)	AUC	F1-score		Recall	
				0	1	0	1
1	Random Forest	98%	0.983	0.98	0.98	0.98	0.99
2	Decision Tree	95%	0.958	0.96	0.96	0.95	0.96
3	SVM	75%	0.757	0.76	0.75	0.78	0.74
4	Logistic Regression	73%	0.735	0.74	0.73	0.76	0.71

Figure 4.9: Evaluation metrics comparison of the four techniques [Zhu et al., 2019]

The results indicate that the random forest algorithm outperforms the other three algorithms in terms of loan default prediction and demonstrates a strong generalization capability.

4.3 Comparative Table

In this section, after reading and evaluation each one of the In this section, after reading and evaluating each of the 12 articles, we created a comparison table showing the differences between them. The following table shows the results:

Reference	Features Selection	Models Used	Best Model	Result
[Zhang et al., 2020]	Feature Importance	- Random Forest - Decision Tree - Logistic Regression	- RF - accuracy : 0.86	After feature importance: loan-to-credit ratio, NumberOfTime30-59 DaysPastDueNotWors and NumberOfTime60 89Days PastDueNotWorse year are the variables to pay more attention when processing the loan application
[Ahmed and Rajaleximi, 2019]	/	Credit scoring to manage credit risks	Criterias to be considered in measuring risks such as Credit history, Capacity to repay,..etc	
[Kovvuri and Cheripelli, 2020]	/	LR, DT and RF with original dataset overbalanced data, underbalanced data, both and balanced with ROSE.	- LR - RF looking at accuracy, sensitivity and specificity	Chosen RF because sensitivity is slightly more compared to LR
[Madaan et al., 2021]	/	- Random Forest - Decision Tree	- Random Forest with 80% accuracy	RF much better option over DT on the given dataset

Table 4.17: Comparative table part 01

Reference	Features Selection	Models Used	Best Model	Result
[Anand et al., 2022]	Feature Importance	(15 Classification algorithms) including Multiple LR, DT, KNN, SVM, RF, and other forms of Ensemble Boosting	Extra Trees Classifier	Accuracy : Extra Trees : 86.17 RF : 85.55 most crucial parameters are customer's employment or job experience in years and debt income.
[Wu, 2022]	Low Variance	- Random forest - XGBoost	- Random Forest	- Accuracy : RF : 0.90657 XGBoost : 0.9063
[Sayjadah et al., 2018]	/	-Decision tree - Extra Trees Classifier - Random forest - Gradient Boosting	- Random Forest	With feature selection: Extra Trees: 0.95 DT: 0.89 MSE: 0.04 RF: 0.97 Gradient Boosting: 0.91
[Zhou et al., 2019]	/	- Gradient Boosting Decision Trees - Extreme Gradient Boosting - Light Gradient Boosting Machine.	- Ensemble model and XGBoost	- Accuracy : - Ensemble model : 0.7185 - XGBoost : 0.7162

Table 4.18: Comparative table part 02

Reference	Features Selection	Models Used	Best Model	Result
[Wang et al., 2020]	/	, - Linear Regression - Decision Tree - Random Forest - Naive Bayesian - KNN	Decision Tree and Random Forest	, Random Forest outperforms the other classifiers in terms of precision, recall, AUC
[Trivedi, 2020]	Information Gain Chi-Squared Gain Ratio	- Naïve Bayes - SVM - Random Forest - Decision Tree	Random Forest	Combine Random Forest with Information Gain Feature Selection gives more accurate model.
[Namvar et al., 2018]	/	- Random forest - Logistic Regression - Linear discriminate analysis (LDA)	Random Forest	Accuracy: 0.8176
[Zhu et al., 2019]	/	- Random Forest - Decision Tree - SVM - Logistic Regression	- Random Forest	Accyracy : RF : 98% TD : 95%

Table 4.19: Comparative table part 03

4.4 Synthesis

The effectiveness of machine learning has been proved in various complex problems by using different approaches to learn from data and extract valuable patterns and information.

Many articles have been written and published regarding the subject of loan approval prediction based on machine learning approaches.

However, after studying twelve articles and papers, we learned about the most performant models used in this topic and we extracted the folowing points :

- **Most used models :** They used 16 different models in the 12 articles and papers. The most used models are: Random Forest (used in 10 articles from 12) and Decision tree (Used in 8 articles from 12) to a lesser extent, they used Linear discriminate analysis (LDA) only in one article and SVM 3 articles.
- **Best models :** The Best models are: Random Forest (9 times from 10) and Decision tree,XGBoost,Extra Trees Classifier only once.
- **Pre-treatment techniques :** the most used techniques to analyze and clean all undesirable data are :dealing with Missing Values,data Normalization,feature selection, encoding categorical variables, data balancing..etc
- **Best evaluation methods :** They used different evaluation methods in the 12 articles and papers, such as accuarcy,recall, precision, F1-score, AUC, etc..

4.5 Conclusion

In this chapter, we have examined and studied the most recent approaches proposed by latest research. These approaches revolve around the utilization of machine learning techniques to deal with loan default or credit risk. The majority of the examined works demonstrated the ability to attain fairly satisfactory outcomes. It is important to note that using up-to-date and large datasets is crucial to obtain good result and it is required to improve the performance of the model.

Part IV

General Conclusion

Chapter 5

General Conclusion

Loan approval prediction is a very interesting field of study. It offers immense value to financial institutions, borrowers, and regulators by enabling informed decision-making and efficient resource allocation. With advancements in machine learning and data analytics, researchers are continually refining models to accurately forecast loan approval outcomes.

The advancement and progress made in artificial intelligence,specifically in Machine Learning, enables us to deal with more complex problems and achieve good results, which is very important for overcoming the different challenges in life. Artificial intelligence provides many advantages and benefit in exploring and simplifying complex tasks.

Our work fits into this same line of research, where we propose a comparative study on different Machine Learning algorithms with our subject on credit risk management and how to predict loan defaults.

After reading these multiple articles we realized that there is a very crucial problem on this field, and if banks and financial institutions in general did not focus on credit risk management and take it in proper consideration it will lead to serious damage and lot of financial losses.

Due to the restricted access to confidential data from most financial institutions, especially banks, researchers face challenges in obtaining comprehensive and valid datasets for free even for scientific purposes.

As a result, researchers have been limited to utilize datasets available on platforms like Kaggle such as "Give me Some credit" dataset which was used in a Kaggle competition (we did not use it in our study because of the lack of features it has only 12 feature), or a very small dataset with max 700 samples, or the one we used which is "Lending Club" dataset.

However, Despite these constraints, researchers have still managed to develop impressive models and provide valuable solutions to the existing problem.

We presented a simple literature of the loan approval prediction techniques, features and some surveys covering all the aspects of this problem.

This MASTER degree research will be considered as the first step toward our objective ‘building a machine learning loan approval prediction model’.

The following points outline the major perspectives to improve our work:

- We will use an approach based on Machine learning algorithms. We are going to choose between the best Machine Learning models concluded from the several articles and the comparison table. Those models are (XGBOOST model, Random forest classifier and Decision Tree) and other models such as (KNN, Logistic Regression, Naive Bayes, SVM), then we select the best model.

- After reading several articles we realized that feature selection is a very crucial step to achieve an accurate model, and because of that we focused on selecting the most relevant features in the building of our model.

- In the end, the evaluation metrics that we are going to use in order to see which one of the models is the best are : Accuracy, precision, recall, F1-Score and ROC-AUC.

Bibliography

[Abiola et al., 2014] Abiola, I., Olausi, A. S., et al. (2014). The impact of credit risk management on the commercial banks performance in nigeria. *International Journal of Management and sustainability*, 3(5):295–306.

[Addo et al., 2018] Addo, P. M., Guegan, D., and Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, 6(2):38.

[Ahmed and Rajaleximi, 2019] Ahmed, M. I. and Rajaleximi, P. R. (2019). An empirical study on credit scoring and credit scorecard for financialinstitutions. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 8(7):2278–1323.

[Alpaydin, 2020] Alpaydin, E. (2020). *Introduction to machine learning.* MIT press.

[Amershi et al., 2019] Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. (2019). Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300.

[Anand et al., 2022] Anand, M., Velu, A., and Whig, P. (2022). Prediction of loan behaviour with machine learning models for secure banking. *Journal of Computer Science and Engineering (JCSE)*, 3(1):1–13.

[Aslam et al., 2019] Aslam, U., Tariq Aziz, H. I., Sohail, A., and Batcha, N. K. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience*, 16(8):3483–3488.

[Bandyopadhyay, 2016] Bandyopadhyay, A. (2016). *Managing portfolio credit risk in banks.* Cambridge University Press.

[Bao et al., 2019] Bao, W., Lianju, N., and Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128:301–315.

[Bessis, 2015] Bessis, J. (2015). *Risk management in banking (4th ed.). John Wiley Sons, Incorporated.* MIT press.

[Chen and Li, 2010] Chen, F.-L. and Li, F.-C. (2010). Combination of feature selection approaches with svm in credit scoring. *Expert systems with applications*, 37(7):4902–4909.

[Chen et al., 2020] Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., Wang, X., Bian, H., Zhang, S., Pradhan, B., et al. (2020). Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. *Science of The Total Environment*, 701:134979.

[Chiu et al., 2021] Chiu, Y.-T., Zhu, Y.-Q., and Corbett, J. (2021). In the hearts and minds of employees: A model of pre-adoptive appraisal toward artificial intelligence in organizations. *International Journal of Information Management*, 60:102379.

[Dridi, 2021] Dridi, S. (2021). Supervised learning-a systematic literature review.

[Fu et al., 2017] Fu, Q., Li, B., Hou, Y., Bi, X., and Zhang, X. (2017). Effects of land use and climate change on ecosystem services in central asia's arid regions: a case study in altay prefecture, china. *Science of the Total Environment*, 607:633–646.

[Garg, 2022] Garg (2022). How feature selection techniques for machine learning are important?

[Goh and Lee, 2019] Goh, R. Y. and Lee, L. S. (2019). Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, 2019.

[Himberg, 2021] Himberg, T. (2021). Loan default prediction with machine learning.

[Holzinger et al., 2018] Holzinger, A., Kieseberg, P., Weippl, E., and Tjoa, A. M. (2018). Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable ai. In *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*, pages 1–8. Springer.

[Janiesch et al., 2021] Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3):685–695.

[Jiang and Zhu, 2022] Jiang, M. and Zhu, Z. (2022). The role of artificial intelligence algorithms in marine scientific research. *Frontiers in Marine Science*, 9:920994.

[Kaplan and Haenlein, 2019] Kaplan, A. and Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, 62(1):15–25.

[Konovalova et al., 2016] Konovalova, N., Kristovska, I., and Kudinska, M. (2016). Credit risk management in commercial banks. *Polish Journal of Management Studies*, 13.

[Kotsiantis, 2011] Kotsiantis, S. (2011). Feature selection for machine learning classification problems: a recent overview. *Artificial Intelligence Review*, 42(1):157–176.

[Kovvuri and Cheripelli, 2020] Kovvuri, R. S. and Cheripelli, R. (2020). Credit risk valuation using an efficient machine learning algorithm. In *Advances in Decision Sciences, Image Processing, Security and Computer Vision: International Conference on Emerging Trends in Engineering (ICETE), Vol. 2*, pages 648–657. Springer.

[Krenn et al., 2022] Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., et al. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769.

[Kumar and Minz, 2014] Kumar, V. and Minz, S. (2014). Feature selection: A literature review. *Smart Comput. Rev.*, 4:211–229.

[Kumari et al., 2018] Kumari, K., Yadav, S., et al. (2018). Linear regression analysis study. *Journal of the practice of Cardiovascular Sciences*, 4(1):33.

[Li et al., 2023] Li, Y., Zhang, Y., Timofte, R., Van Gool, L., Yu, L., Li, Y., Li, X., Jiang, T., Wu, Q., Han, M., et al. (2023). Ntire 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1959.

[Liu et al., 2018] Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., and Lee, I. (2018). Artificial intelligence in the 21st century. *Ieee Access*, 6:34403–34421.

[Madaan et al., 2021] Madaan, M., Kumar, A., Keshri, C., Jain, R., and Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In *IOP Conference Series: Materials Science and Engineering*, volume 1022, page 012042. IOP Publishing.

[Namvar et al., 2018] Namvar, A., Siami, M., Rabhi, F., and Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. *arXiv preprint arXiv:1805.00801*.

[Pan, 2016] Pan, Y. (2016). Heading toward artificial intelligence 2.0. *Engineering*, 2(4):409–413.

[Prasad et al., 2020] Prasad, P. S., Pathak, R., Gunjan, V. K., and Ramana Rao, H. (2020). Deep learning based representation for face recognition. In *ICCCE 2019: Proceedings of the 2nd International Conference on Communications and Cyber Physical Engineering*, pages 419–424. Springer.

[Ravikumar, 2019] Ravikumar, T. (2019). Digital financial inclusion: A payoff of financial technology and digital finance uprising in india. *International Journal of Scientific & Technology Research*, 8(11):3434–3438.

[Rupali and Amit, 2017] Rupali, M. and Amit, P. (2017). A review paper on general concepts of artificial intelligence and machine learning. *International Advanced Research Journal in Science, Engineering and Technology*, 4(4):79–82.

[Sayjadah et al., 2018] Sayjadah, Y., Hashem, I. A. T., Alotaibi, F., and Kasmiran, K. A. (2018). Credit card default prediction using machine learning techniques. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, pages 1–4. IEEE.

[Schilling et al., 2019] Schilling, K. G., Nath, V., Hansen, C., Parvathaneni, P., Blaber, J., Gao, Y., Neher, P., Aydogan, D. B., Shi, Y., Ocampo-Pineda, M., et al. (2019). Limits to anatomical accuracy of diffusion tractography using modern approaches. *Neuroimage*, 185:1–11.

[Siadati et al., 2018] Siadati, S., Tarokh, M. J., and Noorossana, R. (2018). Improving sampling using fuzzy lhs in healthcare supply chain. *Industrial Engineering & Management Systems*, 17(2):294–301.

[Trivedi, 2020] Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, 63:101413.

[Uthayakumar et al., 2020] Uthayakumar, J., Vengattaraman, T., and Dhavachelvan, P. (2020). Swarm intelligence based classification rule induction (cri) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis. *Journal of King Saud University-Computer and Information Sciences*, 32(6):647–657.

[Vishal, 2022] Vishal (2022). Feature selection — correlation and p-value.

[Wang et al., 2020] Wang, Y., Zhang, Y., Lu, Y., and Yu, X. (2020). A comparative assessment of credit risk model based on machine learning——a case study of bank loan data. *Procedia Computer Science*, 174:141–149.

[Wu et al., 2010] Wu, C., Guo, Y., Zhang, X., and Xia, H. (2010). Study of personal credit risk assessment based on support vector machine ensemble. *International Journal of Innovative Computing, Information and Control*, 6(5):2353–2360.

[Wu, 2022] Wu, W. (2022). Machine learning approaches to predict loan default. *Intelligent Information Management*, 14(5):157–164.

[Zhang et al., 2020] Zhang, H., Bi, Y., Jiang, W., Luo, C., Cao, S., Guo, P., and Zhang, J. (2020). Application of random forest classifier in loan default forecast. In *Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, July 17–20, 2020, Proceedings, Part III 6*, pages 410–420. Springer.

[Zhou et al., 2019] Zhou, J., Li, W., Wang, J., Ding, S., and Xia, C. (2019). Default prediction in p2p lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and its Applications*, 534:122370.

[Zhu et al., 2019] Zhu, D., Zhang, Z., Cui, P., and Zhu, W. (2019). Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1399–1407.