الجمهورية الجزائرية الديمقراطية الشعبية

People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي

Ministry of Higher Education and Scientific Research

المدرسة الوطنية العليا للتكنولوجيا

National Higher School of Technology

Department : Industrial Engineering and Maintenance

Final Year Project to Obtain the Diploma of Engineering

Field :

Industrial Engineering

Specialty :

Industrial Engineering

Subject :

Development of a Predictive Tool of Startup Success

Realized by :

MAHIEDDINE Maroua

Members of the Jury

BOUDHAR Hamza (MCB)	President
GHOMARI Leila (MCA)	Supervisor
SI AHMED Boualem (MAA)	Examinator

Algiers,the 25/06/2023.

Academic year 2022/2023

الجمهورية الجزائرية الديمقراطية الشعبية

People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي

Ministry of Higher Education and Scientific Research

المدرسة الوطنية العليا للتكنولوجيا

National Higher School of Technology

Department : Industrial Engineering and Maintenance

Final Year Project to Obtain the Diploma of Engineering

Field :

Industrial Engineering

Specialty :

Industrial Engineering

Subject :

Development of a Predictive Tool of Startup Success

Realized by :

MAHIEDDINE Maroua

Members of the Jury

BOUDHAR Hamza (MCB)	President
GHOMARI Leila (MCA)	Supervisor
SI AHMED Boualem (MAA)	Examinator

Algiers,the 25/06/2023.

Academic year 2022/2023

Acknoledgements

First and foremost, praise and gratitude to Allah Almighty for providing me with the patience, bravery, determination, and motivation to complete this work.

I would like to express my deepest gratitude to my supervisor, Dr. GHOMARI Leila, for their invaluable guidance, support, and expertise throughout the entire process of conducting this project. Their insightful feedback, unwavering patience, and constant encouragement have been instrumental in shaping the direction and quality of this thesis.

I express my sincere gratitude to BRENCO members, Mr. BROURI, and Mr. BOUDJEMAA, for providing me with the needed support.

To my loving parents and dear sisters, thank you for enduring my moments of stress and for being a constant source of encouragement. I am blessed to have you by my side.

I would also like to acknowledge and extend my gratitude to my dedicated teachers, especially Mr. BOUDHAR, for their commitment, and guidance throughout my academic journey. Their support has been instrumental in my academic growth and development.

Finally, I want to express my heartfelt gratitude to all my friends who have provided emotional support and encouragement during the ups and downs of this journey. Your presence and friendship have made this experience more enjoyable and meaningful.

I am deeply grateful to all of you, thank you.

- Maroua -

Table of Contents

Table des figures

Liste des tableaux

Glossary

ADASYN	*Adaptive Synthetic Sampling*
ANN	*Artificial Neural Networks*
ANPT	*The National Agency for the Promotion and Development of Technological Parks*
AUC-ROC	*Area Under the Receiver Operating Characteristic curve*
CA	*Classification Accuracy*
ERC	*Entrepreneurial Research Consortium*
IDE	*Innovation Driven Entreprise*
IPO	*Initial Public Offering*
KNN	*K-Nearest Neighbor*
LightGBM	*Light Gradient Boosting Machine*
LGBM	*Light Gradient Boosting Machine*
ML	*Machine learning*
PCA	*Principal Component Analysis*
RF	*Random Forest*
SMEs	*Small and Medium size Entreprises*
SMOTE	*Synthetic Minority Over-sampling Technique*
SVM	*Support Vector Machine*
XGBoost	*Extreme Gradient Boosting*

General Introduction

Startups have become a popular avenue for individuals and teams to pursue their entrepreneurial dreams and make a significant impact on the economy, innovation, and job creation. Governments, companies, and educational institutions are increasingly promoting and encouraging the growth of startups, recognizing their potential for driving economic growth and fostering innovation.

This year, the Algerian Ministry of Higher Education and Scientific Research introduced a new procedure for students in the final stage. The procedure, under the title "Un diplôme... une startup", (outlined in the Ministerial Bylaw No. 1275 of September 27th, 2022), aims to transform students from job seekers into wealth-creating entrepreneurs and job creators. Within this procedure, eligible students can benefit from the "innovative project" and "startup" labels, aligning their academic pursuits with real-world entrepreneurship and offering them a pathway to contribute to economic growth and create employment opportunities in Algeria[3]. The significant turnout of students to this procedure, with brilliant ideas and creative projects has caught our attention, raising concerns about how unclear and uncertain the future of these projects is. We could not help but wonder, will all of them succeed ?

After making some research, we discovered statistics revealing that approximately 90% of startups experience complete failure, with 10% failing within the first year and 70% failing between the second and fifth year [4]. These findings further emphasize the challenges and uncertainties involved in building a startup. These challenges include factors such as financial resources, team dynamics, competencies, market demand, and competition.

Traditionally, decision-makers have relied on experience, intuition, and trial and error to navigate the challenges associated with startups. However, with the advent of machine learning and artificial intelligence, there is a growing interest in leveraging these technologies to analyze data and predict the success of startups. In 2013, the term "Unicorn" was first introduced to describe a startup that has achieved a valuation of over $1 billion, referring to their rarity [5]. Variants of the term, such as Decacorn and Hectocorn, have been introduced for startups valued over $10 billion and $100 billion, respectively [6]. The unprecedented growth of unicorns has made predicting the success of startups a prominent topic within the startup ecosystem. Investors, venture capitalists, and entrepreneurs alike are actively seeking ways to identify the next unicorn and replicate similar levels of success. Consequently, in the last decade, an increasing number of research projects have emerged, focusing on predicting the outcomes of startups, using machine learning and deep learning technologies.

Startup success prediction using machine learning involves the utilization of diverse machine learning techniques to analyze and predict the probability of a startup achieving success. This prediction process involves analyzing historical data, including factors such as financial indicators, market, team composition, and external factors, identifying patterns and correlations, and generating predictive models that can estimate the likelihood of success for startups in different contexts and industries.

Predicting startup success addresses the inherent difficulties and uncertainties that investors and entrepreneurs face in the dynamic startup ecosystem. It helps stakeholders navigate the challenges of making decisions regarding investment, resource allocation, and strategy in a high-risk and unpredictable environment.

The main objective of this study is to develop a predictive model that can accurately classify startups as either successful or not, exploring both binary and multi-class classification approaches. Additionally, in this project, we aim to include startups that are still operating and striving to grow, yet have not yet demonstrated clear success, or failure. Furthermore, the study will involve the evaluation and comparison of various machine learning algorithms to determine their effectiveness in generating the predictive model. By testing and analyzing different algorithms, the research seeks to identify the most suitable approach for this specific problem and for the used database. The ultimate goal of this thesis is to make a contribution to the advancement of the startup ecosystem in Algeria by motivating and promoting the field of startup success prediction and contribute to the understanding of the process of identifying success drivers. Furthermore, we aspire to create a comprehensive guide that outlines the best practices, methodologies, and considerations for building startup success predictors in Algeria. By sharing our findings and methodologies, we aim to facilitate the development of similar predictive models tailored to other regions and domains, ultimately contributing to a broader understanding of the process of identifying success drivers in startup ecosystems.

The subsequent sections of this thesis are organized as follows, aiming to provide a comprehensive analysis of the topic at hand. Chapter I, Theoretical Background, delves into startup concepts, the startup ecosystem, and the fundamentals of machine learning. It also includes a literature review of the startup success prediction problem, highlighting the gaps in previous research in the field. Chapter II, Methodology, outlines the research process encompassing data preparation, feature engineering, modeling techniques, and the selection of evaluation metrics. This chapter describes the careful selection of machine learning algorithms and discusses the chosen evaluation metrics. Chapter III, Implementation and Results, focuses on the practical implementation of the study and presents the obtained results. Finally, Chapter IV, General Conclusion, summarizes the key findings, highlights their implications, and proposes potential avenues for future research.

Chapter I

Theoretical Background

I.1 Introduction

In this chapter, we lay the theoretical foundation for our exploration of predicting startup success using machine learning algorithms. We begin by examining the startup world, delving into startup definitions and the lifecycle of startups, which helps us understand their unique nature. We then shift our focus to the concept of a startup ecosystem, with a specific emphasis on the Algerian startup ecosystem. By exploring this ecosystem, we gain valuable insights into the environment in which Algerian startups operate.

Next, we provide an overview of machine learning, encompassing different types of algorithms and their applications. Within the realm of machine learning, we specifically explore unsupervised learning algorithms, which have the ability to learn patterns in data without explicit guidance. Additionally, we delve into the evaluation metrics used to assess the performance of these algorithms.

Furthermore, we conduct a literature review to examine prior research on predicting startup success. We explore existing studies and frameworks that have attempted to forecast the success of startups using various methodologies. This literature review helps us identify the current state of knowledge in the field and highlights any limitations or gaps in prior research.

By establishing this comprehensive theoretical background, we develop a solid understanding of the fundamental principles that underpin our exploration of predicting startup success through machine learning.

I.2 Startup world

I.2.1 Startup definition

Definition 1

"A startup is a human institution designed to create a new product or service under conditions of extreme uncertainty." - Eric Ries [7]

Definition 2

"A startup is a temporary organization designed to search for a repeatable and scalable business model." -Steve Blank [8]

Definition 3

"A startup is a company that is designed to grow fast. Being newly founded does not in itself make a company a startup. Nor is it necessary for a startup to work on technology, or take venture funding, or have some sort of "exit." The only essential thing is growth." - Paul Graham [9]

Definition 4

"A startup is a company working to solve a problem where the solution is not obvious and success is not guaranteed." -Neil Blumenthal [10]

These definitions highlight different aspects of a startup's operation, providing varied perspectives on what characterizes a startup. They emphasize elements such as extreme uncertainty, the search for a scalable business model, rapid growth, and the focus on solving problems with uncertain outcomes. Together, these definitions offer a multifaceted view of startups, encompassing their dynamic nature and their drive to navigate challenges and achieve success in an ever-changing landscape.

I.2.2 Startup funding lifecycle

The evolution of a startup from an idea to exit is a continuous process. It is often difficult to precisely identify exactly where you are in the startup lifecycle because it involves many factors. The length of each startup stage will vary greatly depending on business execution, your industry or sector, and your fundraising abilities. Stages of funding are one way to identify in which phase the startup is.

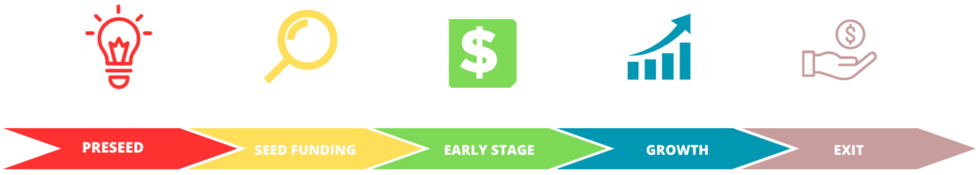


FIGURE I.1 – Funding stages of a startup

Pre-seed : In the pre-seed stage, founders rely on their personal savings or bank loans to finance their startup. They may also participate in competitions to secure additional subsidies if they win.

Seed funding : Seed funding involves raising capital from family and friends, often referred to as "friends and family" funding. Crowdfunding platforms can also be used to gather donations or offer compensation or equity in return.

Early stage : This stage is characterized by the initial development and market validation of the startup. Funding can come from angel investors or early-stage venture capital firms.

Growth : As the startup progresses and expands, it may seek funding from venture capital funds and traditional sources such as bank loans. Venture capital funds provide larger investments but require close monitoring and detailed reporting.

Exit : When a startup reaches a mature phase, investors look to recover their investments and generate profits. This can be achieved through exit strategies like merger and acquisition (M&A), where another company acquires the startup, or going public through an IPO. M&A can be a result of either a successful venture or a resolution for a less successful one. IPOs allow the company to offer its shares to the public, raising additional capital and creating liquidity for early investors. These exit strategies enable investors to realize returns and potentially support new startup ventures.

I.2.3 Initial Public Offering - IPO

Initial Public Offering is the initial sale of a privately held company's stock to the public. It represents a significant milestone in the company's lifecycle, as it transitions from private ownership to being publicly traded. Through an IPO, the company gains access to additional financing, which can support its continued growth and expansion. Moreover, insiders and early investors have the opportunity to eventually sell their shares to the public, providing liquidity and potentially realizing profits. Overall, an IPO represents a critical event that allows a company to tap into the public markets, attracting capital and providing liquidity opportunities for both the company and its stakeholders. [11].

I.2.4 Merger and Acquisition - M&A

Mergers and acquisitions (M&A) involves the consolidation of companies/startups through mergers or acquisitions. Mergers occur when two companies join forces to create a stronger entity, often under a new name, while acquisitions involves one company acquiring another. M&A activities aim to achieve objectives such as market expansion, diversification, and gaining competitive advantage. Strategic management plays a crucial role in planning, due diligence, negotiation, integration, and post-M&A activities. Successful M&A requires careful planning, analysis, and effective integration strategies for long-term value creation and organizational synergy [12].

I.2.5 Startup & SME

Small and Medium-sized Enterprises (SMEs) typically refer to established businesses with a moderate scale of operations and a steady growth trajectory. They often have a defined market presence, established customer base, and a relatively stable business model. SMEs are characterized by their ability to generate consistent revenue and maintain long-term sustainability. They may focus on incremental, rather than radical innovation and process improvements [1].

On the other hand, startups are usually newly founded ventures with a high degree of innovation and a focus on rapid growth. Startups are characterized by their pursuit of disruptive ideas, novel business models, and high-risk tolerance. They aim to create and capture new market opportunities, often leveraging technology and innovation to drive their competitive advantage. Startups typically prioritize scalability, market disruption, and attracting external funding to fuel their growth trajectory [13].

While both SMEs and startups contribute to the economy and foster innovation, they differ in terms of their stage of development, growth objectives, risk profile, and approach to innovation. SMEs often emphasize stability, gradual growth, and incremental innovation, while startups prioritize rapid growth, market disruption, and breakthrough innovations. Figure I.2 depicts the revenue, cash flow, and job creation trends over time for startups and SMEs (startups are referred to as IDE : Innovation Driven Entreprises). For SMEs, the graph is a straight line indicating a relatively stable and consistent revenue growth pattern. In contrast, the revenue trend for startups is more complex. Initially, startups may experience a decline or slower growth during their early stages as they navigate challenges and establish their market presence. However, as they gain traction and refine their business models, startups often exhibit a rapid escalation in revenue.

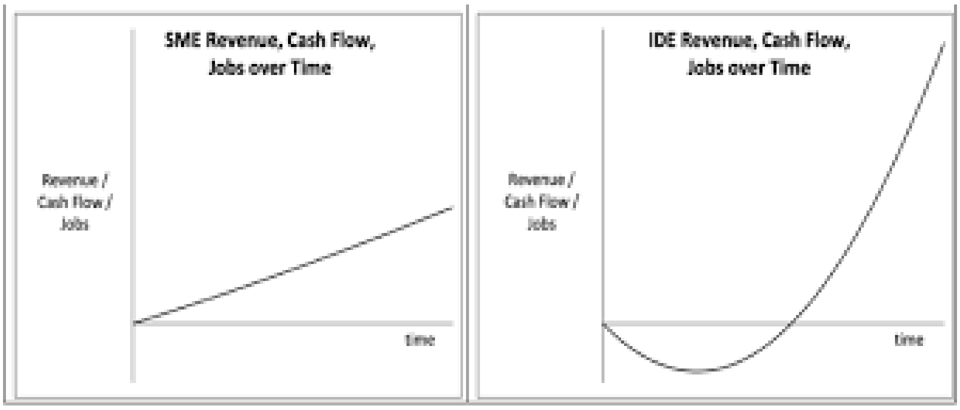


FIGURE I.2 – Startup VS SME [1]

Understanding these distinctions is crucial for policymakers, investors, and entrepreneurs seeking to navigate the diverse landscape of small enterprises within the innovation economy.

I.2.6 Startup ecosystem

In the context of a startup ecosystem, various entities, such as entrepreneurs, investors, incubators, accelerators, government agencies, educational institutions, and support organizations, come together to create a dynamic and supportive business environment. Each of these actors brings their unique strengths and resources to the ecosystem, contributing to its overall vitality.

Entrepreneurs form the core of the startup ecosystem. They are the individuals who conceive innovative ideas and take the risk of turning them into viable businesses. These entrepreneurs often require financial resources, human capital, expertise, and infrastructure to bring their ideas to fruition. This is where other actors in the ecosystem play a crucial role.

Investors, such as venture capitalists and angel investors, provide the necessary funding for startups to develop and scale their operations. They not only inject capital but also bring valuable industry knowledge, networks, and mentorship to help startups thrive. By investing in promising ventures, they fuel the growth and expansion of the startup ecosystem.

Incubators and accelerators are organizations that provide support, guidance, and resources to early-stage startups. They offer mentorship, workspace, access to networks, and various services to help startups refine their business models, develop prototypes, and acquire customers. Incubators focus on nurturing startups in their initial stages, while accelerators aim to expedite their growth and provide them with intensive support.

Government agencies play a vital role in creating a favorable regulatory and policy environment for startups. They can offer financial incentives, tax benefits, grants, and programs to encourage entrepreneurship and innovation. Additionally, they facilitate access to resources, infrastructure, and international markets, enabling startups to expand their reach and compete globally.

Educational institutions contribute to the startup ecosystem by fostering a culture of entrepreneurship and innovation. They provide entrepreneurship education, research facilities, and collaboration opportunities that enable students and researchers to develop their ideas and launch startups. Universities often collaborate with industry partners and support organizations to bridge the gap between academia and the business world.

Support organizations, such as industry associations, co-working spaces, and networking platforms, play a critical role in connecting entrepreneurs and facilitating knowledge exchange. They organize events, workshops, and conferences that allow startups to showcase their ideas, network with potential collaborators, and learn from experienced professionals.

By bringing together these diverse actors and resources, the startup ecosystem creates a collaborative and interconnected environment. Startups benefit from the collective expertise, networks, and shared resources, leading to increased innovation, faster growth, and improved competitiveness. The ecosystem also attracts talent, investment, and business opportunities, further enhancing its vibrancy and sustainability.

a startup ecosystem is a dynamic network of actors who collaborate and contribute their strengths to create a supportive business environment. By fostering cooperation, sharing resources, and leveraging diverse expertise, the ecosystem empowers startups to achieve their goals, deliver value to customers, and drive economic growth.

I.3 Algerian Startup ecosystem

The Algerian startup ecosystem is gradually evolving, supported by both public and private initiatives. The role of the Algerian state is instrumental in creating an environment conducive to the emergence and growth of startups. The establishment of the Ministry Delegate for the Knowledge Economy, Startups and Microenterprises highlights the government's commitment to fostering innovation and entrepreneurship. Additionally, the creation of a dedicated fund for supporting and developing the startup ecosystem further demonstrates the state's focus on providing financial resources. Public funds are also made available to finance startups, enabling them to access the necessary capital for their growth.

The ecosystem of support and financing for startups in Algeria plays a crucial role in their development and integration into the professional landscape. Public structures are established to provide support and guidance to startup founders. The National Agency for the Promotion and Development of Technological Parks (ANPT) facilitates the creation of startups through its technology parks, which offer expertise, assistance, and personalized coaching to innovative projects in the field of information and communication technologies.

University incubators and "Maisons d'entrepreneuriat", contribute to the ecosystem by nurturing innovative projects closely tied to the academic community. These incubators assist student entrepreneurs in refining their ideas and validating the feasibility of their projects. They also promote entrepreneurial culture within the university environment through various activities such as conferences and seminars. Additionally, With the new procedure put in place by the Ministry of Higher Education and Scientific Research, which encourages students to bring their innovative ideas to life, many graduating students have enrolled to complete their end-of-study projects as part of the "Un diplôme, une Startup" program, with innovative ideas in various fields. This program serves as a catalyst for entrepreneurship among students, allowing them to transform their academic projects into viable startups. The program offers a unique opportunity for students to receive support and guidance from experienced mentors and professionals.

Private initiatives are equally significant in bolstering the Algerian startup ecosystem. Private incubators provide crucial support and assistance to startups during their early stages, helping them navigate the challenges of starting a business. Accelerators like Sylabs and The Pivot offer coaching, training, and workspace to startups, fostering their growth and facilitating access to funding opportunities.

While the Algerian startup ecosystem is progressing, it still faces challenges in terms of scale and impact. The number of startups contributing significantly to the national economy remains relatively low. However, efforts are underway to improve the ecosystem and create a more favorable environment for startup creation and growth. With the combined support of public initiatives, private actors, and increased investment, the Algerian startup ecosystem holds the potential to become a vibrant hub of innovation and entrepreneurship in the future.

I.4 Machine Learning

Machine learning is a subfield of artificial intelligence that deals with the development of algorithms and statistical models that enable computer systems to learn from and improve their performance on a task without being explicitly programmed.

Definition 1

"The field of study that gives computers the ability to learn without explicitly being programmed." - Arthur Samuel, 1950 [14]

Definition 2

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.” Tom Mitchell, 1997 [15]

Definition 3

"Machine learning is the set of methods that allow a computer to learn how to perform a task from data, without being explicitly programmed." - Pedro Domingos [16]

These definitions capture the essence of what machine learning is, the ability of computers to learn from data and improve their performance without human intervention. This field has revolutionized problem-solving and decision-making processes, by extracting valuable insights from vast amounts of data, which enhances efficiency, and empowers organizations to make more informed decisions.

I.4.1 Types of machine learning algorithms

Machine learning algorithms can be broadly classified into different types based on the amount of human effort involved in their coordination and the type of data they utilize. There are four major types of machine learning : supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

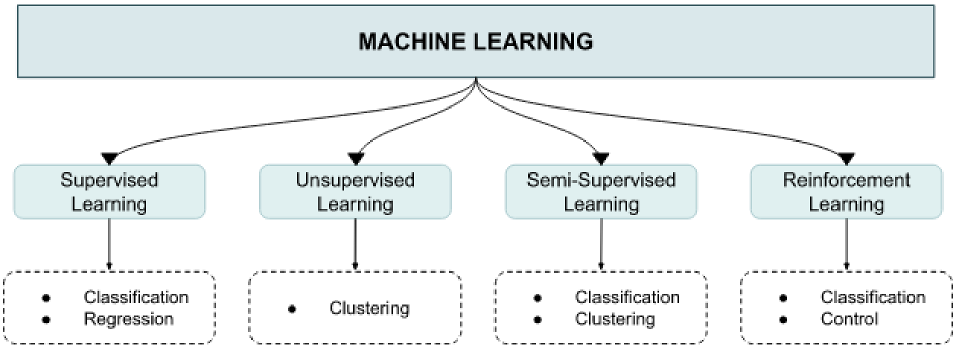


FIGURE I.3 – Types of Machine Learning

Supervised Learning

Supervised learning involves training a machine learning model on a labeled dataset. In this type of learning, the algorithm is given input data and their corresponding outputs, and it tries to learn the mapping function between them. The goal is to create a model that can predict the output of new input data.

Unsupervised Learning

In unsupervised learning, the algorithm is not provided with any labeled data. Instead, it has to identify patterns or structures in the input data on its own. Clustering is a popular example of unsupervised learning, where the algorithm groups similar data points together based on their similarities.

Semi-supervised Learning

Semi-supervised learning is a combination of supervised and unsupervised learning. In this type of learning, the algorithm is provided with a small amount of labeled data and a large amount of unlabeled data. The algorithm first learns from the labeled data and then uses this knowledge to make predictions on the unlabeled data.

Reinforcement Learning

Reinforcement learning is a type of machine learning that involves an agent interacting with an environment to learn how to perform a task. The agent receives feedback in the form of rewards or punishments based on its actions, and it tries to learn a policy that maximizes the cumulative reward over time.

I.4.2 Supervised Learning

Startup success prediction is a classification problem that falls within the realm of supervised learning. In supervised learning, a predictive model is developed using labeled data, where each data point is associated with a known outcome or target variable. The process of supervised learning, as described in Figure I.4, encompasses various stages from raw data to a fully trained classifier.

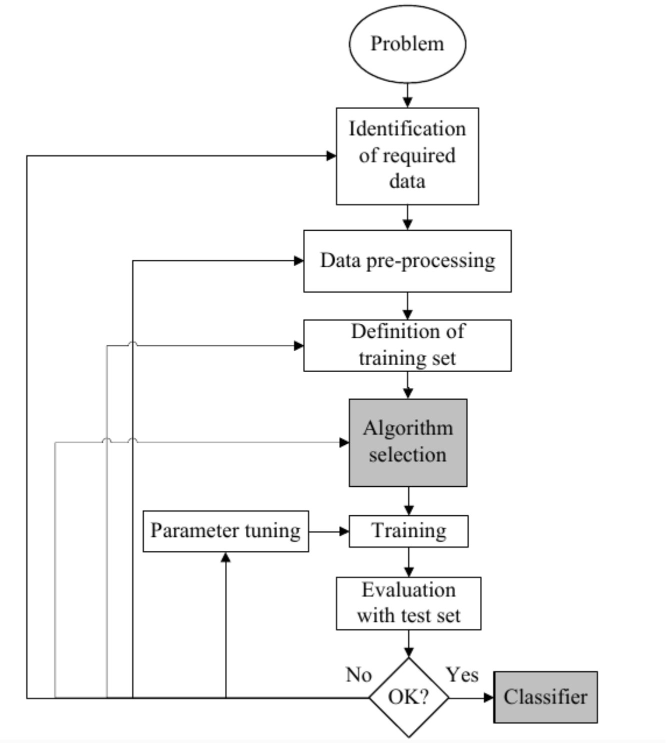


FIGURE I.4 – The process of Supervised ML [2]

First, the problem at hand needs to be clearly defined. This involves understanding what we are aiming to accomplish, such as predicting an outcome or classifying data into categories.

Next, the relevant data for training the model needs to be identified or collected. This data consists of labeled examples where the desired output is known for each datapoint.

Once the data is ready, preprocessing is performed. This step involves cleaning the data by handling missing values, removing outliers, and transforming it into a suitable format for the learning algorithm. Additional preprocessing techniques may include feature scaling, normalization, or feature engineering to improve the model's performance.

The data is then divided into two sets : the training set and the testing set. The training set is used to train the model, while the testing set is used to assess its generalization and predictive capabilities.

Next, an appropriate learning algorithm is selected based on the problem type, data characteristics, and desired outcome. Some algorithms will be presented later in this chapter.

After selecting the algorithm, parameter tuning is performed. Many algorithms have adjustable parameters that affect their performance. The optimal parameter values are selected through techniques like grid search or randomized search to optimize the model's performance.

Once the algorithm and parameters are chosen, the model is trained using the training data. It learns the underlying patterns and relationships between the input features and the corresponding output labels by adjusting its internal parameters.

Finally, the trained model is evaluated using the testing set. Its predictive performance is assessed by comparing its predictions on the testing data to the true labels. Evaluation metrics like accuracy, precision, recall, F1 score, and area under the curve are commonly used to measure the model's performance.

Through these steps, supervised learning enables the development of models that can make predictions or classifications on new data based on patterns learned from labeled examples [17].

I.4.3 Algorithms

Machine learning offers a wide array of algorithms that can be utilized to develop predictive models. Understanding the fundamentals of these algorithms and their underlying principles will provide a solid foundation for selecting the most suitable approach for predicting startup success.

Decision Trees

Decision Trees are a type of tree-based classifiers that classify instances by sorting them based on feature values. A decision tree is built by recursively partitioning the data into subsets based on the values of one of the input features. The goal is to create a tree-like model where each internal node represents a decision based on a feature value, and each leaf node represents the predicted output. Instances are classified starting at the root node, based on their feature values. [18]

Random Forest

Random forest is an ensemble classification learning algorithm that builds multiple decision trees and combines their predictions to make a final one. The algorithm works by selecting a random subsets of features and a subset of the training data, to build each tree. To make a prediction, the algorithm aggregates the predictions of all the trees in the forest and outputs the class that has the most votes [19].

Extreme Gradient Boosting

XGBoost is designed to create highly accurate predictive models by combining multiple weak predictive models, such as decision trees, in an additive manner. It utilizes a gradient-based optimization approach to iteratively refine the model's predictions, making it particularly effective in handling complex, structured datasets.

Light Gradient Boosting Machine

LightGBM is a fast and efficient gradient boosting framework that is optimized for performance and memory use. It shares similar principles with XGBoost but introduces some unique features to enhance training speed and reduce memory consumption. LightGBM uses a histogram-based approach for binning continuous features, which enables faster training and allows for parallel computation. It is commonly used in scenarios where large-scale datasets or time constraints necessitate quick model training.

Logistic Regression

Logistic regression is a binary classification algorithm, which models the probability of an instance belonging to a certain class. During prediction, the algorithm estimates the probability of a new instance belonging to one class and makes a binary decision based on a decision threshold [20].

K-Nearest Neighbor

K-nearest neighbors (kNN) is one of the simplest machine learning algorithms. Building the model only consists of storing the training dataset. To make a prediction for a new data point, KNN finds the k data points in the training set that are closest to the new data point, and assigns the new data point to the most common class among those k neighbors.

Support Vector Machine

SVMs find a hyperplane in a high-dimensional space that best separates data points of different classes. In binary classification, the hyperplane that maximizes the margin, or distance between the closest data points of each class, is found. SVMs can also handle non-linearly separable data by mapping the data to a higher-dimensional space using a kernel function. SVMs have been widely used in various fields, as they are very powerful for solving complex classification problems [2].

Neural Network

Neural networks is an algorithm inspired by the structure and function of the human brain. It consists of interconnected nodes called artificial neurons, that process and analyze data to identify patterns and relationships. The nodes are organized in layers, with the input layer receiving the data, one or more hidden layers processing it, and an output layer producing the results [21].

These machine learning algorithms have a wide range of application domains, each with their own set of advantages and drawbacks. To provide a comparative analysis, we have selected specific application domains of these algorithms, along with their respective advantages and drawbacks, summarized in Table I.1.

TABLE I.1 – Comparison between ML algorithms

Algorithm	Application domain	Advantages	Drawbacks %
Decision Trees	Medical diagnosis [22]	Ability to handle both categorical and numerical data, ease of use and interpretation [19]	Increased memory consumption [23] [2]
Random Forest	environmental science [24]	Ability to handle high-dimensional data and provide robust predictions [20]	Computationally expensive
XGBoost	Financial fraud detection [25]	Feature Importance and Interpretability	Complexity and Parameter Tuning
LightGBM	Sentiment Analysis [26]	Fast Training Speed	Limited Handling of Missing Data
Logistic Regression	Medecine [27]	Easy to implement and train	Sensitive to the outliers in dataset
kNN	Image classification [28]	Ease of implementation	Large storage requirements [2]
SVM	Bioinformatique	Effectiveness in handling high-dimensional data [2, 29]	Poor interpretability [29]
Neural Network	Speech Recognition	Ability to work with incomplete knowledge [30]	Poor interpretability [2, 30]

I.4.4 Evaluation metrics

The selection and evaluation of machine learning (ML) algorithms are crucial steps in predicting startup success. Various evaluation metrics are employed to assess the performance and effectiveness of these algorithms. These metrics provide insights into how well the models are capturing patterns and making accurate predictions.

One commonly used metric is accuracy, which measures the overall correctness of the predictions made by the ML algorithm. It calculates the ratio of correctly classified instances to the total number of instances in the dataset. While accuracy is a straightforward measure, it may not always be sufficient, especially when dealing with imbalanced datasets where the distribution of classes is uneven.

Precision and recall are two complementary metrics that are often used together to evaluate ML algorithms. Precision measures the proportion of true positive predictions out of all positive predictions, focusing on the accuracy of positive predictions. On the other hand, recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive instances in the dataset, emphasizing the algorithm's ability to identify positive instances correctly. These metrics are particularly useful when the cost of false positives or false negatives is different.

F1 score is another widely used metric that combines precision and recall into a single value, providing a balanced measure of the algorithm's performance. It calculates the harmonic mean of precision and recall, emphasizing both metrics equally. The F1 score is especially valuable when there is an imbalance between positive and negative instances in the dataset.

Area Under the Receiver Operating Characteristic curve (AUC-ROC) is a metric commonly used for binary classification problems. It measures the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) for different classification thresholds. The AUC-ROC provides an aggregate measure of the algorithm's performance across various threshold values, offering a robust evaluation of its ability to discriminate between positive and negative instances.

Selecting the appropriate evaluation metrics depends on the specific problem, the nature of the data, and the desired outcome. It is important to consider the strengths and limitations of each metric and choose the ones that align with the objectives of the study.

I.5 Startup Success Prediction Litterature Review

This section aims to analyze and synthesize the knowledge and findings from previous studies, providing a solid foundation for the subsequent chapters of this thesis.

I.5.1 Predicting startup success

The application of machine learning techniques in predicting startup outcomes has attracted significant attention in the literature, due to its potential to benefit various stakeholders involved in the startup ecosystem. Firstly, investors can utilize the findings to effectively screen and evaluate potential investment options, enabling them to make more reliable and timely decisions. Secondly, startups themselves can leverage the research outcomes to identify areas of their business that require additional attention in order to enhance their chances of success. Additionally, policymakers can use research findings to create a conducive ecosystem that supports the growth and success of local startups.

In order to construct the startup success predictive model, researchers started by defining startup success. Many researchers define success as profitability and high Return on investment (ROI) through going public (Initial Public Offering, IPO, the process by which a startup becomes publicly traded by offering its shares to the public), or through a process of getting merged or acquired (M&A, the process by which the startup gets combined with or acquired by another company) [31, 32, 33, 34, 35, 36, 37, 38]. However, some researchers have criticized this definition, asserting that financial metrics may not be highly relevant, particularly in the early stages of startups, as it takes time for them to become profitable [39, 40]. Instead, as a proxy for success, researchers have turned to funding events. Consequently, some authors expanded the definition of success to include the amount of follow-on funding raised, provided it exceeded the previous funding round, as it indicates the growth potential of startups over time [34].

To gather data for their analysis, researchers employed various data sources, with Crunchbase being the most commonly used. Crunchbase offers comprehensive data on companies, including startups. However, this historical dataset has limitations, such as missing information, changes in reporting formats, and a lack of consideration for the temporal aspect during data recording [31]. Mattermark is another database that provides insights on startups markets and investors. This dataset is biased towards successful startups, and additional surveys had to be conducted in order to make it consistent enough to be used in research [41].

To investigate the reasons behind startup failure, researchers have made use of Autopsy, a public database containing postmortem reports of failed startups. Autopsy serves as a valuable resource for entrepreneurs to learn from the mistakes of failed startups by providing insights into the factors contributing to their failure [42, 41].

Researchers also employed data collection methods to ensure representative samples and unified reporting. The Entrepreneurial Research Consortium (ERC), for example, is an international research project that examines random and representative samples of entrepreneurs throughout the startup process [43]. The Danish Business Authority, under the Danish Ministry of Business, generates and collects data on startups, offering another valuable source [44].

Other sources include Dealroom, Deadpool, Indiegogo, Kickstarter, LinkedIn and the web in general. Many researchers combined data from multiple sources in order to form a complete database of a representative sample.

The data that was leveraged by researchers to evaluate startups in their early stages included [43, 44] :

— General information about the founders : such as gender and age
— Human capital : degrees, work experience, and skills of the founder and team

- Motivation : How motivated and what's keeping them motivated
- Process : whether or not they have a business plan, whether or not they receive information and guidance.
- Financial environment : Startup capital
- Network : Industry experience
- Intended organization : Industry type, market, ambition to grow

In addition to the aforementioned factors, research studies made on startups in a more advanced stage [36, 37], used additional factors, such as :

- Average time between funding
- Number of funding rounds
- Number of investors
- Total funding amount

Researchers have explored a variety of machine learning algorithms to predict startup success. These algorithms utilize patterns and relationships within startup data to make accurate predictions. Some of the algorithms commonly employed include Decision Trees [35, 36, 2], Random Forest [20, 19, 43, 31, 32, 39, 35, 36, 38], Logistic Regression [43, 31, 32, 39, 35, 38, 45], K-Nearest Neighbors [38, 2], Neural Networks [31, 39, 35, 45, 2], and Support Vector Machines [39, 36, 41, 2]. It is worth noting that this is not an exhaustive list, as researchers have also explored other algorithms and ensemble methods, which have shown improved accuracy when compared to individual models.

Several studies have identified several factors that significantly impact a startup's probability of success. Among these factors, the age of the startup, funding amounts, total number of funding rounds, and the startup's location have been found to be influential [34, 37]. The reputation and number of investors associated with a startup have also been identified as important predictors of success. Furthermore, incorporating a startup's web presence in the analysis has been shown to enhance the quality of predictions, as it reflects the startup's visibility and perception among its target audience [40].

I.5.2 Limitations and gaps in prior research

Prior research has made significant contributions within the field of predicting startup success, however, there are several gaps and limitations that still need to be addressed in order to advance the field and improve the accuracy of prediction models.

One notable gap in prior research is the lack of industry-specific models. Existing studies often work on generic models that overlook the unique characteristics and dynamics of specific industries and sectors. By not considering industry-specific factors, such as market trends, regulatory environments, and competition, the applicability and accuracy of the prediction models may be limited.

Additionally, there is a gap in the consideration of country-specific factors. Startup ecosystems vary significantly across countries due to variations in policies, business environments, and even cultural norms. However, existing models often overlook these nuances and fail to capture the specific characteristics of individual countries. Developing country-specific models can provide more accurate forecasts and tailored insights into the startup landscape within each country.

Another limitation in prior research is the reliance on data sources with limited value and inconsistent information. The availability of comprehensive and reliable data is crucial for training and validating prediction models effectively. However, many studies have suffered from the use of data sources that lack depth and fail to provide a comprehensive view of the variables. This limitation can introduce biases and inaccuracies into the models, ultimately impacting their accuracy and reliability.

Moreover, there is a lack of consistent evaluation criteria in the field. Different studies employ varying metrics and definitions to measure startup success, making it challenging to compare and generalize findings across different studies. Establishing standardized and meaningful evaluation criteria for success and failure is essential to ensure consistency and enhance the practicality of prediction models.

In conclusion, prior research in the field of predicting startup success has made significant strides. However, several gaps and limitations still need to be addressed. Developing industry-specific models, considering country-specific factors, improving data quality and availability, and establishing consistent evaluation criteria are key areas that warrant further exploration. By addressing these gaps, future research can contribute to the advancement of the field and enhance the accuracy and applicability of prediction models in driving economic growth and innovation.

I.6 Conclusion

This chapter provides a solid theoretical foundation for our study on startup success prediction. We first introduced the startup world, defining startups and exploring their lifecycle, as well as key exit events such as Initial Public Offerings (IPOs) and Merger and Acquisition (M&A) activities. We also discussed the difference between startups and Small and Medium Enterprises (SMEs) and examined the startup ecosystem, emphasizing its significance in fostering entrepreneurial growth.

To contextualize our study, we dedicated a section to the Algerian startup ecosystem, shedding light on its unique characteristics and highlighting its potential for growth and innovation, in order to understand the specific challenges and opportunities within this ecosystem.

Furthermore, we introduced machine learning. We explored different types of machine learning algorithms, with a particular focus on supervised learning methods. Through a detailed examination of algorithms and evaluation metrics, we gained insights into the foundations of startup success prediction using machine learning techniques.

Lastly, we conducted a literature review on startup success prediction, examining the existing research in the field. We identified the current state of knowledge, highlighted the methodologies employed, and recognized the limitations and gaps that provide the motivation for our study. By building upon the existing body of work, we aim to contribute valuable insights and develop a robust predictive model for startup success.

Chapter II

Methodology

II.1 Introduction

This chapter provides a detailed account of the processes and techniques employed in this study to develop a predictive model for startup success. This chapter focuses on the practical aspects, including data analysis and model development, to achieve the research objectives. Figure II.1 illustrates a BPMN process that includes the steps we will follow.

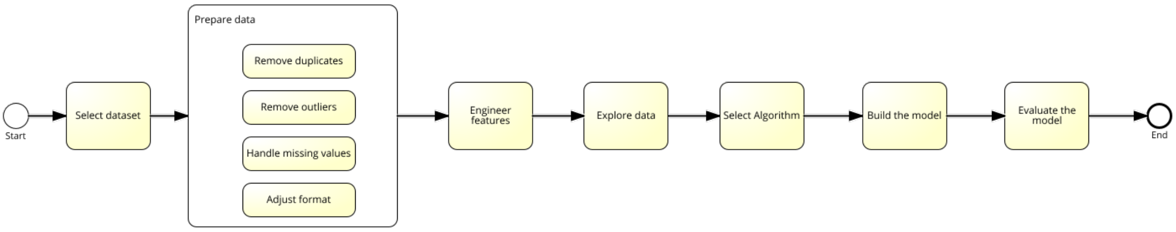


FIGURE II.1 – Methodology overview

To begin, the chapter introduces the concept of startup success and defines the criteria used to determine success in this study. This definition lays the groundwork for subsequent data analysis and model development, ensuring clarity and consistency throughout the research.

Understanding and preparing the dataset is the next step in the methodology. This section provides an overview of the dataset used, highlighting its composition, structure, and relevant variables. Furthermore, data preparation techniques, such as handling missing values, addressing outliers, and performing necessary data transformations, are outlined to ensure data quality and reliability.

By following these steps, we aims to develop a comprehensive and reliable predictive model for startup success. The subsequent sections of this chapter will delve into the specific steps taken for data analysis, model development, and evaluation metrics, providing comprehensive details of the methodology employed to achieve the research objectives.

II.2 Success definition

In the context of this thesis, it is crucial to begin our study by establishing a precise definition of success. This foundational step holds great significance as success can be subjectively interpreted, often leading to conflicting definitions. In addition to traditional measures of success such as M&A and IPOs, we adopt a broader perspective that includes fundraising capability. This expanded definition allows us to evaluate startups that are still in operation and actively raising funds. We made this decision due to the challenging nature of sustaining a startup, considering the high mortality rate among them.

It is worth noting that this study does not aim to assess the financial health of startups. Instead, the primary objective of this thesis is to develop a predictive system that accurately forecasts the outcomes of startups. For the purpose of this prediction, we consider startups that remain operational and capable of raising funds as successful.

II.3 Overview of the dataset

In order to carry out our project, we initially approached governmental offices, specifically the Ministry Delegate for the Knowledge Economy, Startups and Microenterprises, with the aim of obtaining access to authentic data regarding Algerian startups. Unfortunately, our efforts yielded no results due to the scarcity and confidentiality of the data maintained by these offices. We then turned to the incubator BRENCO for assistance. Although they did not possess a readily available database for us to explore, they provided valuable support throughout the project by imparting their domain knowledge and guiding us in making informed choices at every step.

Subsequently, we reached out to well-known platforms that specialize in startup-related information, including Crunchbase, but our requests for access to their databases went unanswered.

Therefore, we proceeded to search for freely available datasets that could be utilized. Our search led us to Kaggle, where we discovered a database extracted from the Crunchbase database. This particular dataset comprises 66,368 rows and 14 columns, offering comprehensive information about the featured companies. The dataset includes general details such as company name, website, country, region, founding date, and category. Furthermore, it provides valuable financial insights, including the total funding amount received, the number of funding rounds conducted, and the first and last funding dates. Notably, the dataset also incorporates the status of each startup, adding further depth to the available information for analysis. Table II.1 provides an overview of the dataset.

	Column	Non-Null values	Datatype	Unique values
0	permalink	66368	object	66368
1	name	66367	object	66102
2	homepage url	61310	object	61191
3	category list	63220	object	27296
4	funding total usd	66368	object	18895
5	status	66368	object	4
6	country code	59410	object	137
7	state code	57821	object	311
8	region	58338	object	1092
9	city	58340	object	5111
10	funding rounds	66368	int64	19
11	founded at	51147	object	3978
12	first funding at	66344	object	4817
13	last funding at	66368	object	4518

TABLE II.1 – Summary of the dataset

After examining the table it is evident that there are specific adjustments required to enhance the dataset. Primarily, one crucial modification involves changing the datatype of the "funding total usd" column from object to a numerical datatype, such as float. This adjustment is necessary to represent and utilize the financial information contained within the column. Additionally, the last three columns in the table represent dates but are currently stored as object values. To ensure their proper utilization, we need to convert these columns to the appropriate date datatype.

The dataset includes a column indicating the status of each startup, which will serve as the target variable for our startup success prediction model. This particular column consists of four distinct values : "acquired," "closed," "ipo," and "operating." To provide a visual representation of the distribution of startups across these categories, Figure II.3 illustrates the number of startups within each class.

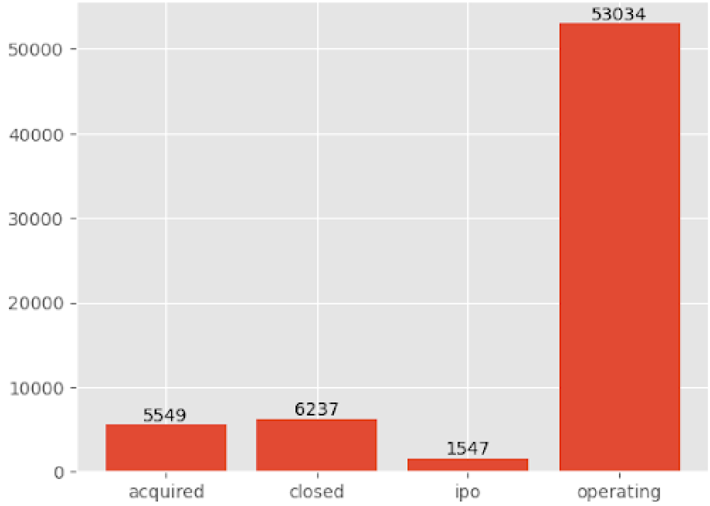


FIGURE II.2 – Number of startups by status

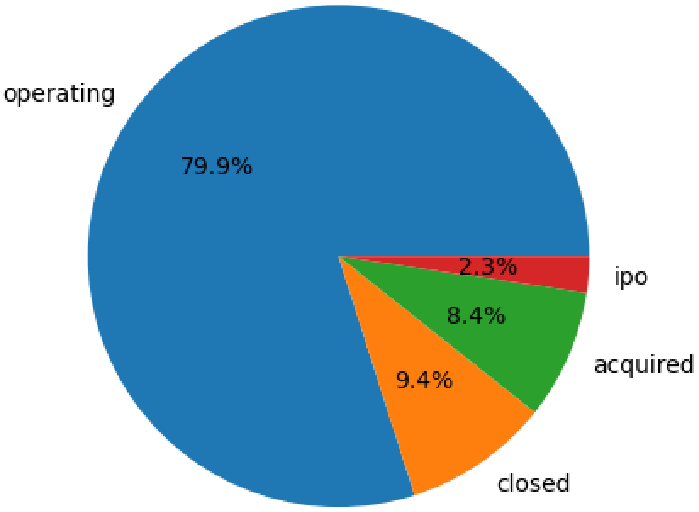


FIGURE II.3 – Number of startups by status - Pie chart

Figure II.4 illustrates the distribution of missing values within the dataset. The visualization reveals that some columns have no missing values, indicating complete data, while others contain missing values. Notably, the "foundation year" column shows a significant number of missing values. Additionally, it is observed that some data points within the dataset have missing values in the four columns dedicated to geographical location, namely "country code," "state code," "region," and "city." This implies that certain entries lack information regarding their specific geographic location.

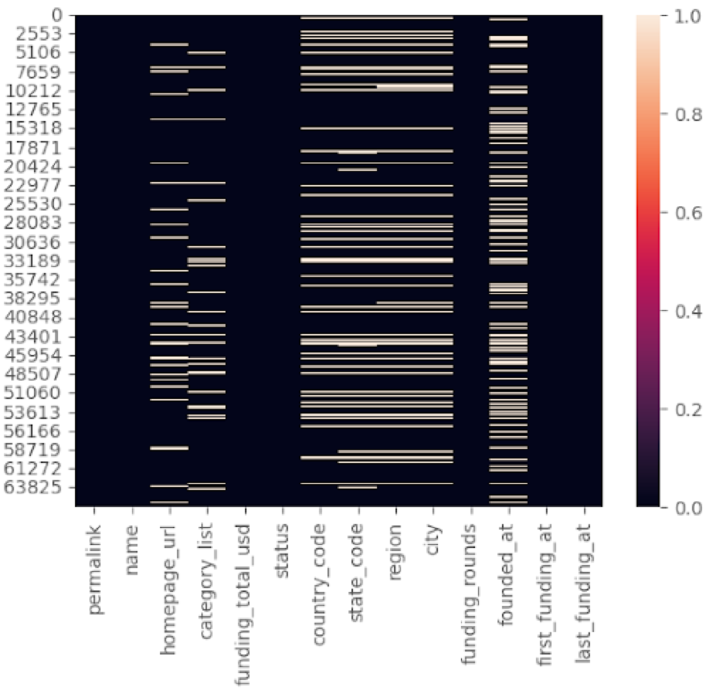
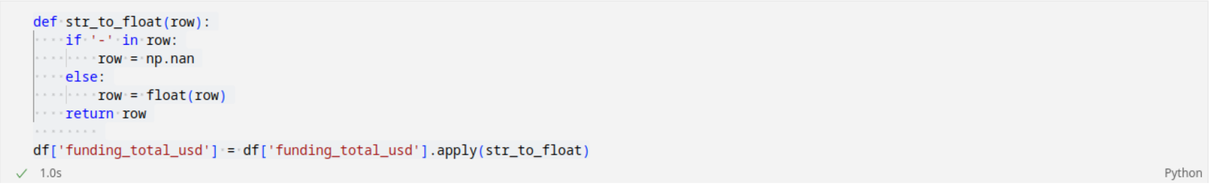


FIGURE II.4 – Missing values distribution

II.4 Data Preparation

In order to prepare our data for use, we start by adjusting the format of the columns, from object to the appropriate format. Figure II.5 represents the code used to convert the "funding total usd" column to float format. This particular column initially contained missing values represented by "-", which were subsequently removed, and replaced with NaN using the NumPy library, in order to allow the conversion process.



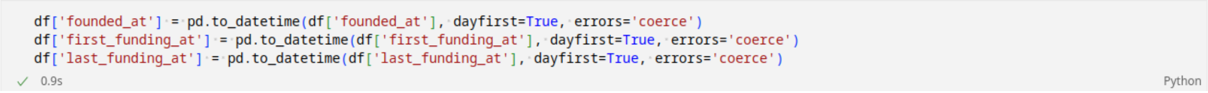
```
def str_to_float(row):
    if '-' in row:
        row = np.nan
    else:
        row = float(row)
    return row

df['funding_total_usd'] = df['funding_total_usd'].apply(str_to_float)
```

✓ 1.0s Python

FIGURE II.5 – Conversion to float format

The three columns representing the dates of startup foundation, the first funding round, and the last funding round required conversion to date format. Figure II.6 shows the code used for this conversion using the Pandas library. The code applies the pd.to_datetime() function to the three columns, transforming their data type into date format. This conversion enables accurate date-based operations and analysis within the dataset.



```
df['founded_at'] = pd.to_datetime(df['founded_at'], dayfirst=True, errors='coerce')
df['first_funding_at'] = pd.to_datetime(df['first_funding_at'], dayfirst=True, errors='coerce')
df['last_funding_at'] = pd.to_datetime(df['last_funding_at'], dayfirst=True, errors='coerce')
```

FIGURE II.6 – Conversion to date format

Figure II.7 illustrates the number of startups founded over the years. However, it appears that there are some inaccuracies in the recorded dates within the database.

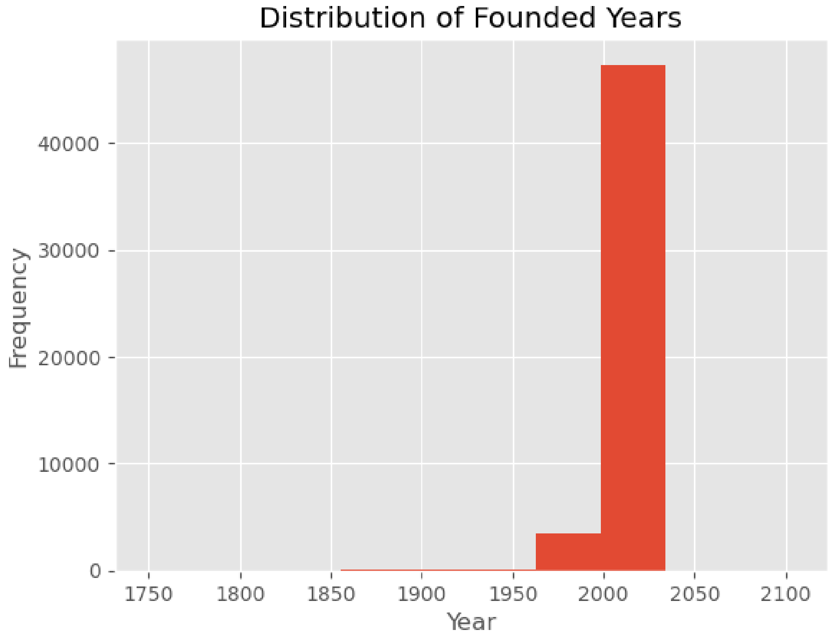


FIGURE II.7 – Distribution of founded years

After printing the oldest and newest startups :

Oldest startup : 1749.0
Newest startup : 2105.0

To ensure that the analysis focuses on a relevant timeframe, these startups will be excluded. Only startups created between 1990 and 2018 will be taken into consideration.

We used boxplots to get a visual summary of funding rounds and funding total distributions, showing key statistics such as the median, quartiles, and range, as well as outliers in the data.

Figure II.8 represents a box plot of the funding_total_usd column. Because there is a large number of datapoints, which are highly concentrated within a narrow range, the box is not visible. Nevertheless, the plot clearly reveals the presence of outliers, represented by data points that are significantly distant from the rest of the dataset.

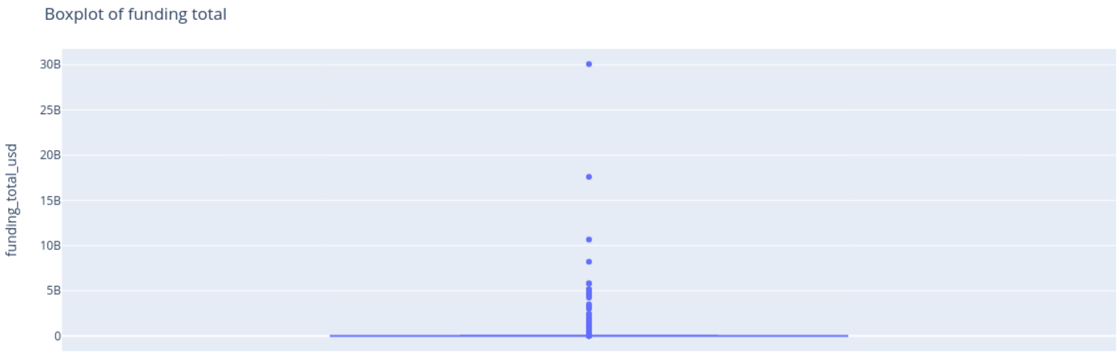


FIGURE II.8 – Boxplot of funding total

Outliers can introduce a high degree of variability and distort the interpretation of statistical measures, leading to misleading conclusions about the funding distribution. By removing these extreme values, the analysis will be more accurate and representative of the majority of the data points.

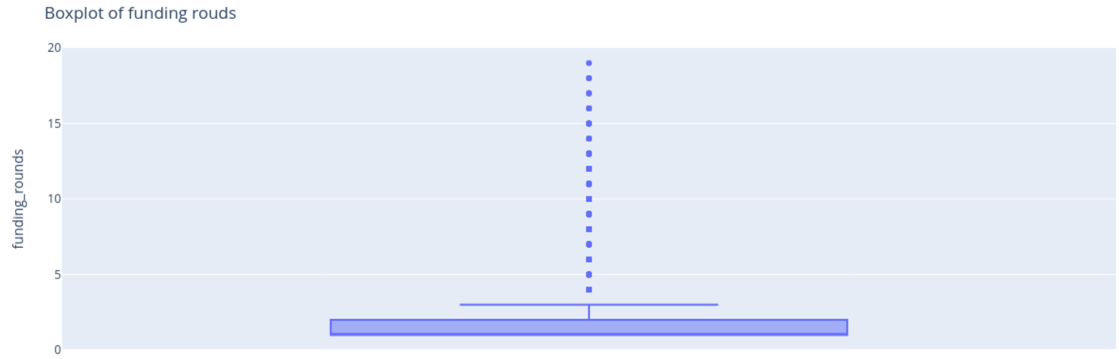


FIGURE II.9 – Boxplot of funding rounds

Handling missing values We will handle the missing values in the dataset by applying specific techniques to the columns necessary for training the model, while dropping the irrelevant columns.

Firstly, for the funding_total_usd column, we will group the startups based on the number of funding rounds they have undergone. The missing values in this column will be imputed using the median value within each funding round group (Appendix I).

Next, for the founded_at column, missing values will be filled with the corresponding value from the first_funding_at column. This assumption helps to ensure consistency between the startup's founding date and the date of their first funding.

Lastly, for the category_list and country_code columns, any missing values will be replaced with the label "other." This approach allows us to retain the existing data while indicating the absence of a specific category.

By implementing these techniques, we can address missing values effectively and maintain a comprehensive dataset for model training. At the end of this phase, we were left with a dataset of 42 858 instances.

II.5 Feature Engineering

In the feature engineering phase, we will apply transformations and create new features based on the existing variables. This process aims to optimize the predictive power of the dataset by leveraging all the available information and facilitating more effective model training. By extracting valuable insights and patterns from the data, feature engineering enhances the overall performance and accuracy of the predictive models.

We began by addressing the "category_list" column, where certain startups had recorded multiple categories separated by a vertical bar "|". To simplify this, we introduced a new column called "main_category" and decided to retain only the first category mentioned, discarding the additional ones.

The "category_list" column initially contained over 20 000 unique values. However, after extracting the first category and creating the "main_category" column, we were left with around 800 unique values. We also observed that certain categories referred to the same concept but were recorded differently. To streamline the analysis further, we grouped similar categories into broader categories. As a result, we managed to reduce the number of unique values in the "main_category" column to 604.

To enhance the analysis and focus on significant categories, we established a threshold of 200 startups. For categories with more than 200 startups, we created new dummy variables to represent each individual category. For the remaining categories that fell below this threshold, we consolidated them into an "other category" dummy variable. This approach allowed us to better capture the important categories while still accounting for the less common ones.

Following the same methodology, we applied a similar threshold of 50 startups to create dummy variables for countries. This enabled us to represent countries with a substantial presence as individual dummy variables, while grouping less prominent countries into an "other country" category.

To fully utilize the information contained in the datetime columns, specifically the "founded_at," "first_funding_at," and "last_funding_at" columns, we have created three essential features :

— "time_to_first_funding" : This feature calculates the duration between the startup's foundation and the date it received its first funding round. It provides valuable insights into how long it took for the startup to secure its initial funding, indicating its early-stage financial stability and attractiveness to investors.

— "time_between_first_last_funding" : This feature calculates the duration between the first and last funding rounds of the startup.

— "funding_lifecycle" : This feature calculates the duration between the startup's foundation and its last funding round.

We can further enhance our analysis by creating an interaction feature using the feature we created "time_between_first_last_funding", and the existing feature "funding_rounds" :

— "funding_frequency" : This feature represents the average frequency of funding rounds for each startup. It is calculated by dividing the number of funding rounds by the time duration between the first and last funding rounds. This feature provides insights into the regularity of securing funding.

By incorporating these features, we gain valuable temporal information that can be invaluable for evaluating a startup's financial health, growth potential, and investor attractiveness.

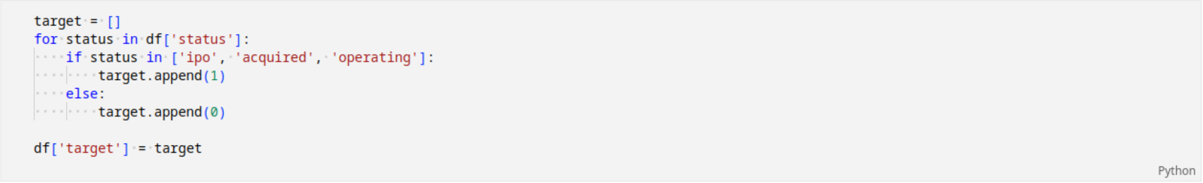
II.6 Target variable

We tested two approaches regarding the target variable, in order to choose the best one. The first one is by making a binary classification (success or failure), as well as a multi class approach by separating the 'operating' class.

II.6.1 Binary classification

We added a new feature, which we built based on the status column. In this feature, we grouped the three statuses "acquired", "ipo", "operating" as a success and assigned the value 1. And for the the status closed we assigned the value 0. The code snippet is shown in figure II.10.

Unfortunately, we did not have access to sufficient information about the funding amounts in each round, as well as the period of time between funding rounds, which limited our ability to establish a more informed, and precise definition of success. Having insights into the funding progression would have allowed us to better evaluate the startup's growth.



```python
target = []
for status in df['status']:
    if status in ['ipo', 'acquired', 'operating']:
        target.append(1)
    else:
        target.append(0)

df['target'] = target
```

FIGURE II.10 – Creating a new target variable

We can now visualize the percentage of success recorded in our dataset. Figure II.11 illustrates that 92.1% of startups in our model are considered a success.

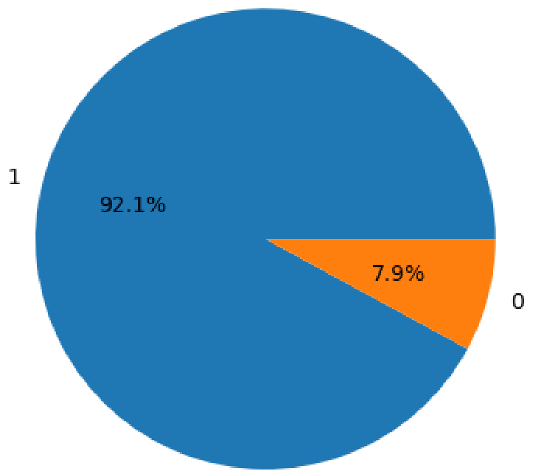
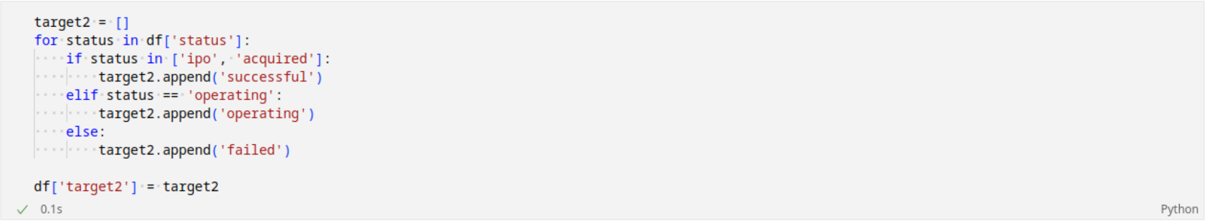


FIGURE II.11 – Success Failure Pie chart

It's worth noting that this sample is not meant to be representative of the overall startup population, and should not be taken as a reference while measuring startup success rate, because generally speaking, successful startups are more likely to get recorded and to report information about their success factors.

II.6.2 Multi-class classification

This time, we grouped startups that demonstrated success through going public or through getting merged or acquired, into a successful class, startups that are still operating in a separate class. The remaining startups, marked with the status 'closed,' were labeled as failed. Figure II.12 illustrates the code snippet used to create a new column called 'target2' to store these classifications.



```python
target2 = []
for status in df['status']:
    if status in ['ipo', 'acquired']:
        target2.append('successful')
    elif status == 'operating':
        target2.append('operating')
    else:
        target2.append('failed')

df['target2'] = target2
```

FIGURE II.12 – Creating a multiclass target variable

Following this grouping process, we generated a pie chart, in figure II.13, to visualize the distribution of the different classes . It reveals that one class, specifically the ’operating’ status, remains dominant among the data points.

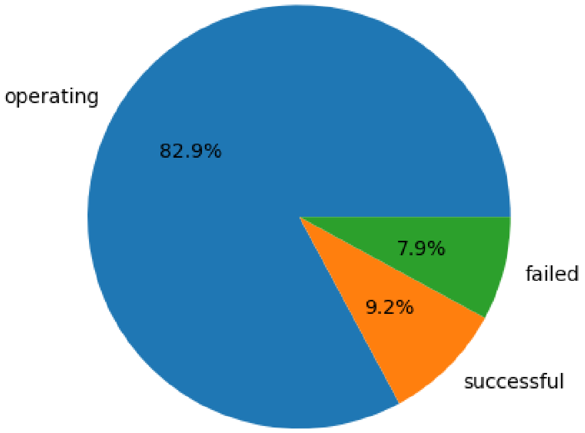
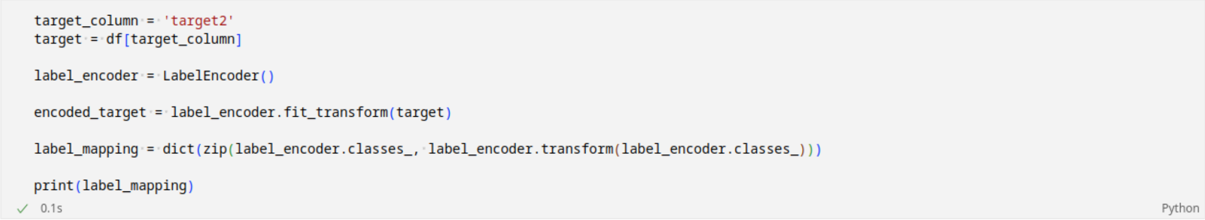


FIGURE II.13 – Pie chart

In the case of multi-class classification, label encoding is often necessary, because the machine learning algorithms we used, require the target variable to be in numerical format. Label encoding converts categorical labels into numeric labels, assigning a unique integer to each category. This encoding allows the algorithm to work with the target variable as numbers. Figure II.14 presents the code snippet that demonstrates how label encoding is applied in this scenario, using Label Encoder from scikit-learn.



```
target_column = 'target2'
target = df[target_column]

label_encoder = LabelEncoder()

encoded_target = label_encoder.fit_transform(target)

label_mapping = dict(zip(label_encoder.classes_, label_encoder.transform(label_encoder.classes_)))

print(label_mapping)
```

FIGURE II.14 – Encoding

II.7 Modeling

As mentioned in the previous section, we are working on 2 approaches, multi-class and binary classification. We have selected four machine learning algorithms to assess and determine the most reliable option, namely Random Forest, XGBoost, LGBM, and Neural networks. This methodology allows us to thoroughly evaluate the performance of different algorithms and make informed decisions based on their effectiveness. These algorithms have been chosen based on their proven effectiveness and suitability for addressing the specific challenges of the startup success prediction problem.

II.7.1 Algorithm choice

Random Forest, XGBoost, and LGBM are tree-based classifiers known for their interpretability. They provide insights into the decision-making process by highlighting important features and the impact they have on predictions. This interpretability aspect is valuable in the context of startup success prediction, as it allows us to understand the key factors influencing success and make informed decisions based on the model's outputs.

Neural Networks, on the other hand, offer a different advantage. They are powerful models capable of capturing complex relationships and patterns in the data.

II.7.2 Handling class imbalance

As observed in Figure II.11 and Figure II.13, we encountered significant class imbalance in both cases. This imbalance has the potential to introduce bias towards the majority class, which could compromise the reliability of the predictions. To mitigate this issue, we explored three different approaches :

Weight Balancing : Weight balancing assigns higher weights to minority classes and lower weights to majority classes during training to address class imbalance. It helps the model focus on learning patterns from the minority class, reduces bias towards the majority class, and improves overall classification performance.

Synthetic Minority Over-sampling Technique (SMOTE) : SMOTE creates synthetic samples by interpolating between existing minority class samples. It helps balance the class distribution by increasing the representation of the minority class. SMOTE is a widely used technique but does not consider classification difficulty or sample density.

Adaptive Synthetic Sampling (ADASYN) : ADASYN is an advanced technique that adaptively generates synthetic samples for the minority class based on the distribution of existing samples. It increases the density of the minority class, making decision boundaries less biased towards the majority class. ADASYN is useful for severe class imbalance scenarios.

II.7.3 Normalization and dimentionality reduction

After splitting the data into training and test sets, with a 30% allocation for testing, we proceeded with data normalization. The goal of performing data normalization is to standardize the features and bring them to a similar scale. This process ensures that no particular feature dominates the learning process due to differences in their magnitudes.

Next, we applied dimensionality reduction using Principal Component Analysis (PCA). PCA is a technique used to reduce the dimensionality of a dataset while retaining most of its relevant information. It achieves this by transforming the original features into a new set of uncorrelated variables called principal components. These principal components capture the maximum variance in the data, allowing us to represent the data in a lower-dimensional space.

By performing PCA, we aim to reduce the complexity and redundancy of the feature space, eliminating any unnecessary noise or irrelevant information. This can lead to improved computational efficiency, better interpretability, and potentially enhanced model performance.

II.7.4 Hyperparameters optimization

When it came to optimizing the hyperparameters, we initially employed a grid search, and then decided to switch to randomized search to optimize our model's hyperparameters due to the significant time saving. Grid search exhaustively explores all possible combinations of hyperparameters within the defined hyperparameter grid, resulting in a computationally intensive process. In contrast, randomized search randomly samples a specified number of hyperparameter configurations from the hyperparameter grid, which allows a faster exploration.

In addition, we implemented k-fold cross-validation with k=5. This technique involves dividing the dataset into k equal parts and using k-1 for training while reserving one part for validation. By repeating this process k times, k-fold cross-validation helps us assess the generalization performance of our models and mitigate any potential overfitting issues.

II.7.5 Evaluation metrics

Selecting appropriate evaluation metrics is essential as they provide different perspectives on the performance of a model. The choice of metrics should depend on the specific objectives of the study and can vary based on the stakeholders involved and their goals.

If the goal is to identify as many successful startups as possible, the recall metric would be more appropriate. Recall measures the proportion of truly successful startups that are correctly identified by the model. Maximizing recall ensures that the model captures a high percentage of successful startups, minimizing the risk of missing out on potential investment opportunities. This metric is particularly important when the focus is on identifying promising startups and avoiding false negatives.

On the other hand, if the objective is to minimize false positives and avoid investing in unsuccessful startups, precision becomes a more relevant metric. Precision measures the accuracy of positive predictions, specifically the proportion of correctly predicted successful startups out of all the predicted successful startups. High precision ensures that the model has a low rate of false positives, reducing the risk of investing in startups that are likely to fail. This metric is crucial for stakeholders who aim to make prudent investment decisions and want to avoid wasting resources on unsuccessful startups.

Additionally, the F1 score combines precision and recall into a single metric, providing a balanced measure of a model's performance. It considers both false positives and false negatives and is valuable when there is a need to strike a balance between precision and recall. The F1 score is particularly useful when the cost of misclassifications is significant and needs to be minimized. By calculating the harmonic mean of precision and recall, the F1 score gives equal weight to both metrics, providing a comprehensive evaluation of the model's performance in predicting startup success.

In the case of a general startup success prediction model, where the goal is to assist a wide range of stakeholders and objectives, it is appropriate to consider all these evaluation metrics. Since the model aims to provide a holistic assessment of startup success, it should be evaluated based on accuracy, precision, recall, and F1 score. This comprehensive evaluation approach ensures that the model performs well in correctly identifying successful startups, minimizing false positives and false negatives, and striking an appropriate balance between precision and recall.

II.8 Conclusion

In conclusion, this chapter outlined the methodology employed in developing a predictive model for startup success. The dataset went through several transformations to ensure its suitability for analysis, including handling missing values and outliers, as well as performing necessary data transformations.

Four machine learning algorithms, namely Random Forest (RF), XGBoost, LightGBM, and Artificial Neural Networks (ANN), were selected to test their performance in predicting startup success. These algorithms were chosen based on their effectiveness and suitability for the task at hand.

Hyperparameter optimization technique, Randomized Search, was selected to tune the selected algorithms and optimize their performance. This process involves tuning the model's parameters to achieve the best possible results.

To evaluate the performance of the predictive model, several evaluation metrics were chosen, namely, accuracy, recall, precision, and F1 score. These metrics provide a comprehensive assessment of the model's predictive capabilities, considering both the overall accuracy and its ability to correctly identify successful startups.

By following this methodology, which encompassed dataset transformations, algorithm selection, hyperparameter optimization, and evaluation metric choice, a predictive model for startup success was developed. The subsequent chapters will delve into the specific results and findings derived from implementing this methodology.

Chapter III

Implementation and results

III.1 Introduction

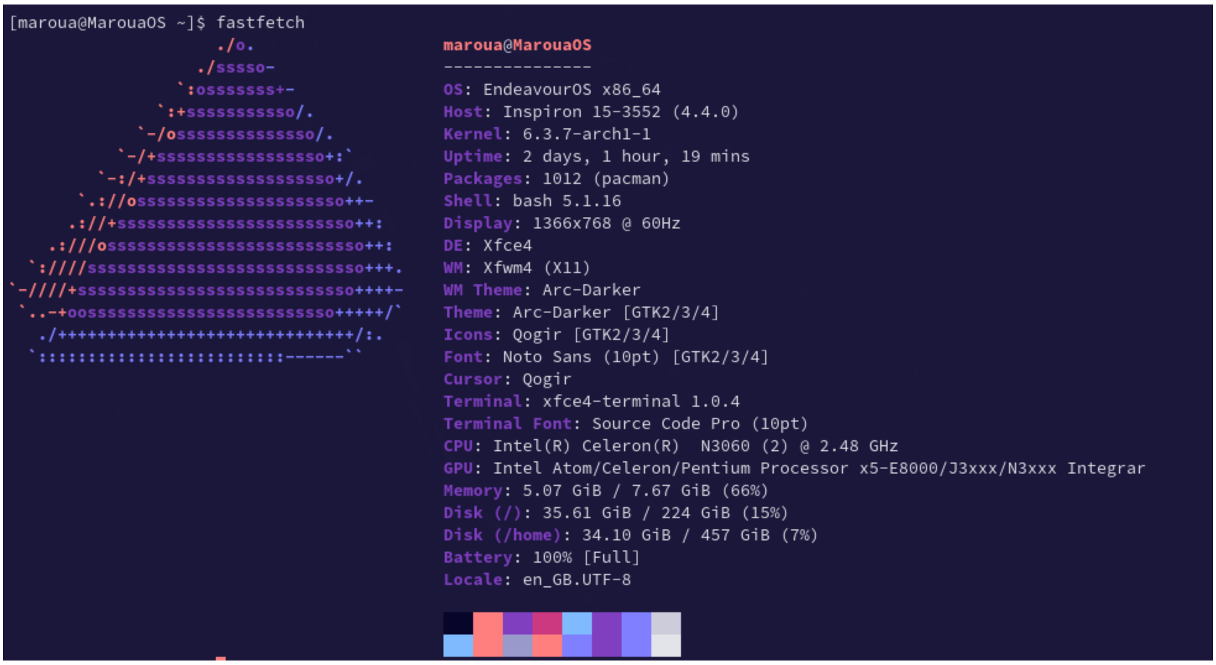
This chapter presents the implementation and the results obtained from the conducted experiments. This chapter provides an overview of the hardware and software specifications of the machine used for the experiments, the tools used, and the performance of the trained models. Additionally, we employ a validation test for our model in order to ensure its reliability and generalizability.

The chapter also examines the concept of feature importance and significance, shedding light on the factors that significantly influence the models' predictions. Feature importance analysis is conducted using various approaches, including LightGBM model feature importance, XGBoost model feature importance, and statistical approaches.

The primary objective of this chapter is to present the implementation details, evaluate the performance of the trained models, and provide insights into the significance of different features. By analyzing the results, it aims to draw meaningful conclusions and contribute to a comprehensive understanding of the implemented models' capabilities and the importance of various features in the prediction process.|

III.2 Hardware and Software Specifications

The experiments and analysis presented in this thesis were conducted on a machine with the following specifications :



```
[maroua@MarouaOS ~]$ fastfetch
                     ./o.                   maroua@MarouaOS
                   ./sssso-                 ---------------
                 `:osssssss+-               OS: EndeavourOS x86_64
               `:+sssssssssso/.             Host: Inspiron 15-3552 (4.4.0)
             `-/ossssssssssssso/.           Kernel: 6.3.7-arch1-1
           `-/+sssssssssssssssso+:`         Uptime: 2 days, 1 hour, 19 mins
         `-:/+sssssssssssssssssso+/.        Packages: 1012 (pacman)
       `.://osssssssssssssssssssso++-       Shell: bash 5.1.16
      .://+ssssssssssssssssssssssso++:      Display: 1366x768 @ 60Hz
    .:///ossssssssssssssssssssssssso++:     DE: Xfce4
  `:////ssssssssssssssssssssssssssso+++.    WM: Xfwm4 (X11)
`-////+ssssssssssssssssssssssssssso++++-    WM Theme: Arc-Darker
 `..-+oosssssssssssssssssssssssso+++++/`    Theme: Arc-Darker [GTK2/3/4]
   ./++++++++++++++++++++++++++++++/:.      Icons: Qogir [GTK2/3/4]
  `:::::::::::::::::::::::::------``        Font: Noto Sans (10pt) [GTK2/3/4]
                                            Cursor: Qogir
                                            Terminal: xfce4-terminal 1.0.4
                                            Terminal Font: Source Code Pro (10pt)
                                            CPU: Intel(R) Celeron(R)  N3060 (2) @ 2.48 GHz
                                            GPU: Intel Atom/Celeron/Pentium Processor x5-E8000/J3xxx/N3xxx Integrar
                                            Memory: 5.07 GiB / 7.67 GiB (66%)
                                            Disk (/): 35.61 GiB / 224 GiB (15%)
                                            Disk (/home): 34.10 GiB / 457 GiB (7%)
                                            Battery: 100% [Full]
                                            Locale: en_GB.UTF-8
```

FIGURE III.1 – Machine Specifications - Terminal Output

Figure III.1 above displays the machine specifications obtained from running FastFetch on the terminal. The machine used for this thesis implementation was an Inspiron 15-3552 with an Intel(R) Celeron(R) N3060 (2) CPU running at 2.48 GHz. It had 7.67 GiB of memory and featured Intel Atom/Celeron/Pentium Processor x5-E8000/J3xxx/N3xxx Integrated Graphics.

The operating system was EndeavourOS x86_64 with Linux kernel version 6.3.7-arch1-1.

These specifications provide essential insights into the machine's capabilities and configuration, ensuring the reproducibility of the results obtained during the implementation.

III.3 Tools used

VSCodium

VSCodium is a free and open-source code editor based on Visual Studio Code (VS Code) without the proprietary Microsoft branding. It provides a similar feature set and user experience as VS Code, including support for various programming languages and extensive extensions [46].

Jupyter

Jupyter is a project and community whose goal is to ”develop open-source software, open-standards, and services for interactive computing across dozens of programming languages”. It was spun off from IPython in 2014 by Fernando Pérez [47].

Pycharm

PyCharm is an integrated development environment used in computer program- ming, specifically for the Python language. It is developed by the Czech company JetBrains [48].

III.4 Trained Models' Performance

Considering that only one algorithm will be selected as the best among the four (Random Forest, XGBoost, LGBM, and Neural Networks), it is important to evaluate their performance based on specific criteria such as accuracy, precision, recall, F1 score, and computational efficiency.

Among these four algorithms, Random Forest and Artificial Neural Networks proved to be computationally intensive with long runtimes. As a result, they were eliminated from further consideration based on their computational efficiency.

The remaining algorithms, XGBoost and LGBM, demonstrated superior computational efficiency compared to Random Forest and Neural Networks. Thus, they were identified as the most suitable choices for further analysis and model development.

To present the performance of the tested models, Table III.1 displays their classification accuracy (CA) and F1-score (F1). This table allows a comparison of the models' performance based on the specified evaluation metrics.

TABLE III.1 – CA & F1-score results of the models

Classification	Class imbalance technique	LGBM		XGBoost	
		CA	F1	CA	F1
Binary	Weight balancing	74.96%	85.14%	74,40%	84,71%
	SMOTE	89.11%	94.19%	87.89%	93.51%
	ADASYN	89.22%	94.27%	88.22%	93.70%
Multi-class	Weight balancing	55.20%	61.63%	51.55%	58.36%
	SMOTE	77.38%	74.97%	74.64%	73.59%
	ADASYN	78.77%	75.12%	76.32%	74.06%

One notable observation is the comparatively lower performance of the multi-class models when compared to the binary classification models. In particular, the multi-class models with weight balancing, which exhibited the poorest accuracy and F1 score among all the models evaluated.

On the other hand, the binary classification models demonstrated strong performance, with the ADASYN LGBM model standing out as the best performer. This model achieved an impressive accuracy of 89.22% and an F1 score of 94.27%. However, it is worth mentioning that all of the binary classification models showed satisfactory performances.

To further validate the models' performance, it is important to test them on a new dataset. This additional testing will help assess the models' generalization capabilities and provide a more robust evaluation of their effectiveness.

III.5 Validation Test

We discovered a dataset online that we intended to utilize for validating our model. An overview of the data is presented in Table III.2.

Upon examining the table, it becomes apparent that the data requires cleaning and several transformations before we can effectively utilize it for model validation. These steps may involve handling missing values, standardizing formats, and ensuring consistency in data types.

The dataset we obtained online originally comprised 1154 startup records, all originating from the United States.

TABLE III.2 – Validation Dataset Overview

No.	Column Name	Non-null Count	Data Type
0	state_code	1154	int64
1	zip_code	1154	int64
2	age_first_funding_year	1154	float64
3	age_last_funding_year	1154	float64
4	age_first_milestone_year	959	float64
5	age_last_milestone_year	959	float64
6	relationships	1154	int64
7	funding_rounds	1154	int64
8	funding_total_usd	1154	int64
9	milestones	1154	int64
10	is_CA	1154	int64
11	is_NY	1154	int64
12	is_MA	1154	int64
13	is_TX	1154	int64
14	is_otherstate	1154	int64
15	is_software	1154	int64
16	is_web	1154	int64
17	is_mobile	1154	int64
18	is_enterprise	1154	int64
19	is_advertising	1154	int64
20	is_gamesvideo	1154	int64
21	is_ecommerce	1154	int64
22	is_biotech	1154	int64
23	is_consulting	1154	int64
24	is_othercategory	1154	int64
25	has_VC	1154	int64
26	has_angel	1154	int64
27	has_roundA	1154	int64
28	has_roundB	1154	int64
29	has_roundC	1154	int64
30	has_roundD	1154	int64
31	avg_participants	1154	float64
32	is_top500	1154	int64
33	status	923	object

We tested all of our binary classification models on this dataset. Results are shown in Table III.2

TABLE III.3 – Validation test results

Model	Class imbalance technique	CA	F1
LightGBM	Weight balancing	63,10%	75,37%
	SMOTE	61,21%	74,02%
	ADASYN	60,78%	73,76%
XGBoost	Weight balancing	64,57%	88,44%
	SMOTE	61.00%	73,72%
	ADASYN	60,88%	73,70%

All of the models demonstrated good performance with over 60% accuracy and over 70% for F1-score, which indicates that the models have good generalization ability on new unseen data. (reference)

III.6 Feature importance and significance

In this section, our primary aim is to identify the key factors that have the most significant impact on startup success within the context of our study. To achieve this objective, we will employ a statistical approach, measure feature importance in our machine learning models, and information gain measures. These methods provide us with valuable insights into the relationship between various features and the target variable, which, in this case, is the measure of startup success.

III.6.1 LightGBM Model Feature Importance

Feature importance refers to a metric that measures the relative importance or contribution of each feature in a machine learning model. It helps to understand which features have a significant impact on the model's predictions.

Figure III.2 is a barplot that illustrates the feature importance in our LightGBM model.

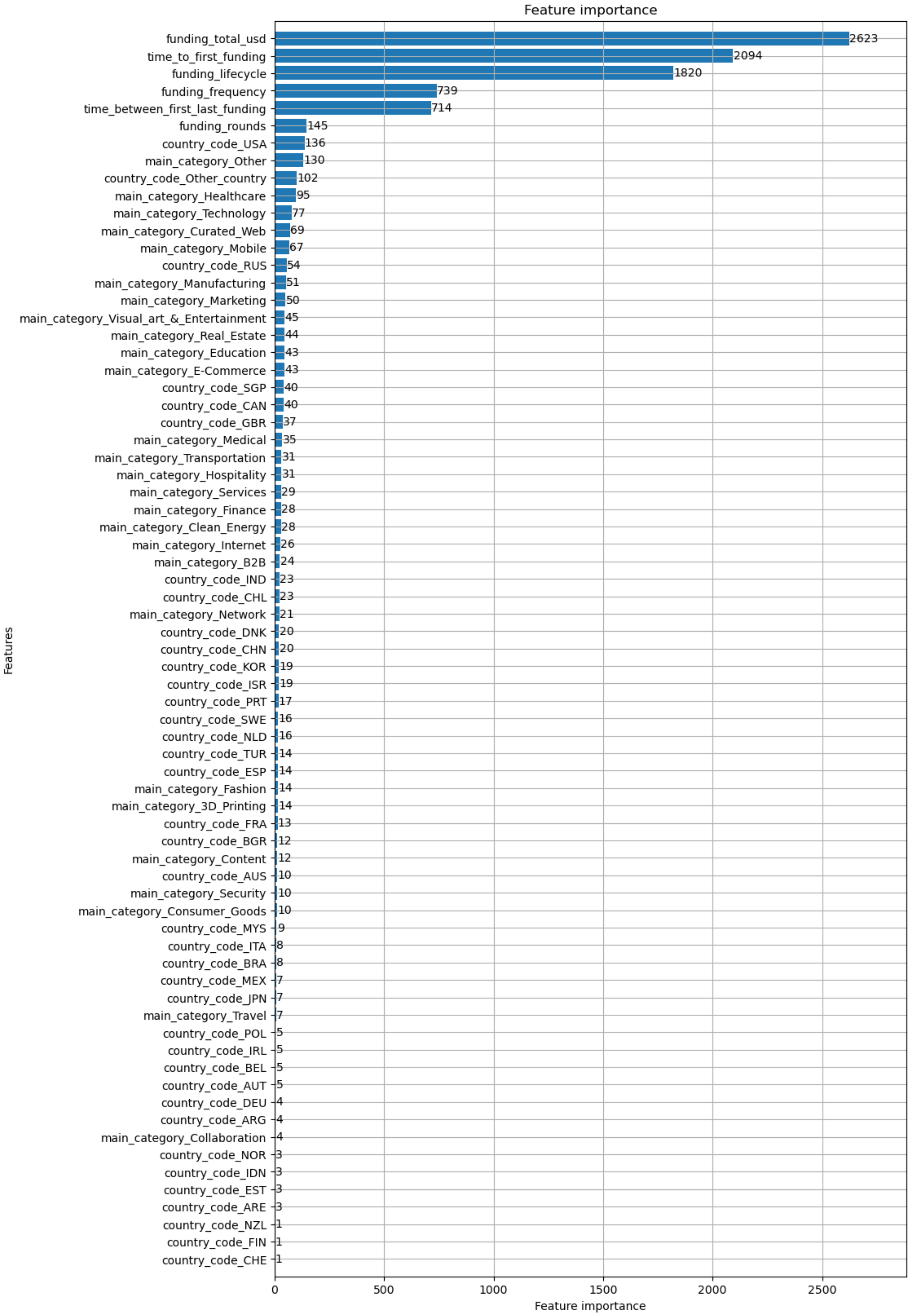


FIGURE III.2 – LightGBM Feature Importance

III.6.2 XGBoost Model Feature Importance

We plotted the feature importance in our XGBoost model too, plot illustrated in Figure III.3

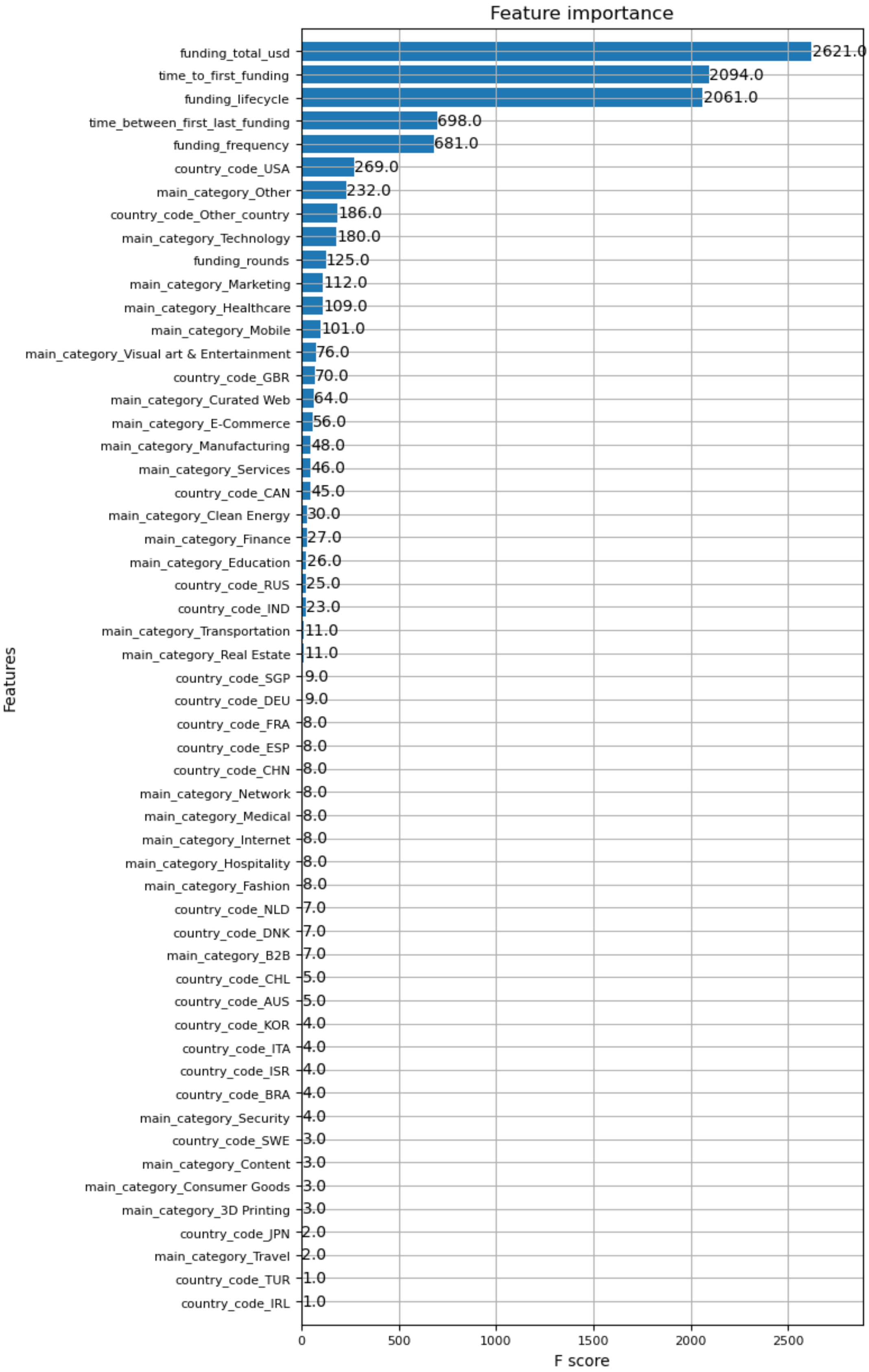
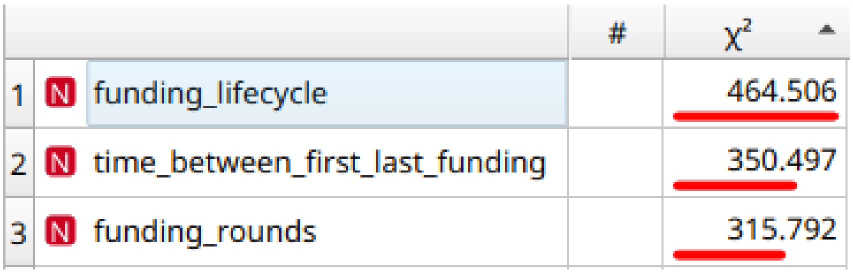


FIGURE III.3 – XGBoost Feature Importance

III.6.3 Statistical Approach - Chi-Square

In the context of feature ranking, the chi-square measures the dependency between each feature and the target variable. Higher chi-square values indicate stronger associations between the feature and the target, suggesting higher relevance.

Using the feature ranking widget from Orange (Appendix II), we were able to rank the features based on the chi-square measure. The ranking results are depicted in Figure III.4, where higher-ranked features have a greater chi-square measure and are considered more influential in the model.



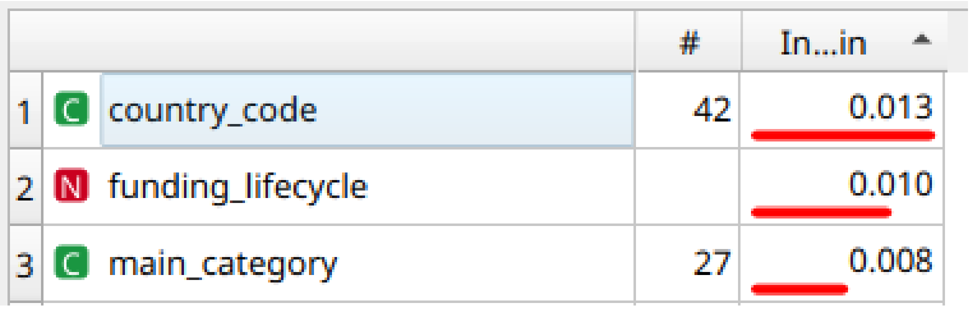
			#	χ^2
1	N	funding_lifecycle		464.506
2	N	time_between_first_last_funding		350.497
3	N	funding_rounds		315.792

FIGURE III.4 – Chi-Square ranking

III.6.4 Information Gain

Information gain quantifies the reduction in entropy achieved by splitting a node based on a particular feature. Features with higher information gain values provide more information about the target variable and are considered more important.

Using the feature ranking widget from Orange, we conducted a ranking of the features based on the information gain measure. The resulting rankings are displayed in Figure III.5, where features with higher ranks indicate a higher information gain and are considered more significant for the model.



			#	In...in
1	C	country_code	42	0.013
2	N	funding_lifecycle		0.010
3	C	main_category	27	0.008

FIGURE III.5 – Information gain ranking

III.7 User Interface Implementation

As a final step, we implemented our trained model into an application with a graphical user interface in order to be able to make predictions easily. Figure III.6 illustrates a use case diagram of the application's functionality.

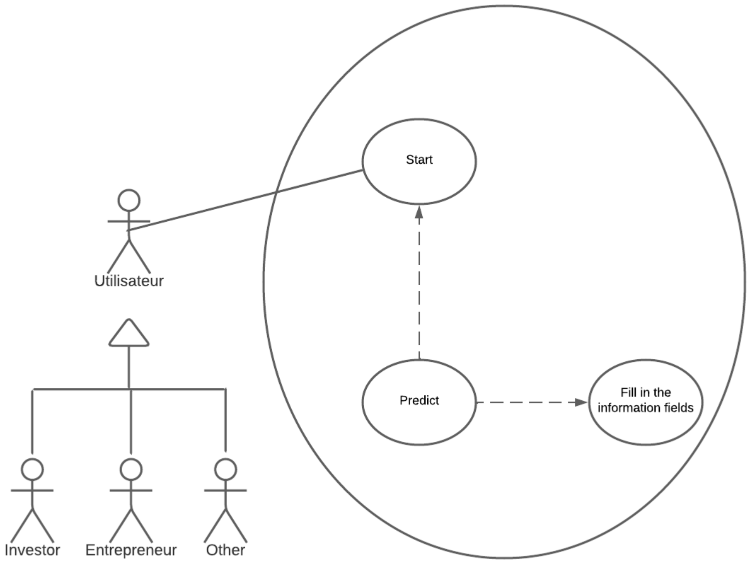


FIGURE III.6 – Use Case diagram of the application

Upon clicking the "Start" button in the welcome window, Figure III.7, a new window is displayed. This window contains input fields where we can enter the necessary information about the startup to make a prediction. The window with the input fields is depicted in Figure III.8.

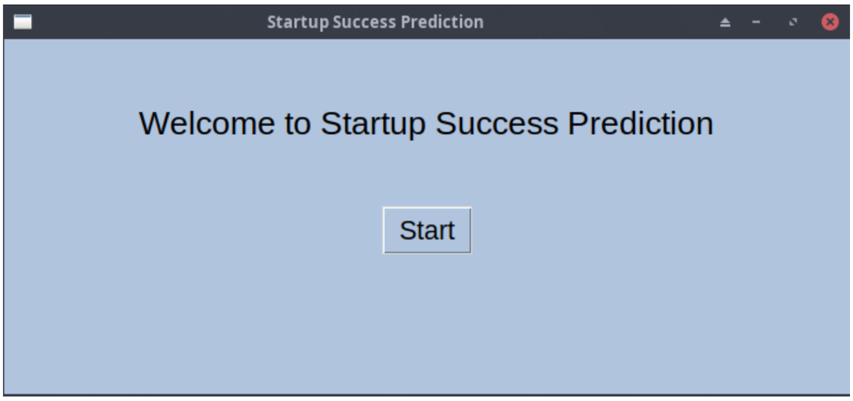

Startup Success Prediction
Welcome to Startup Success Prediction
Start

FIGURE III.7 – Welcome window

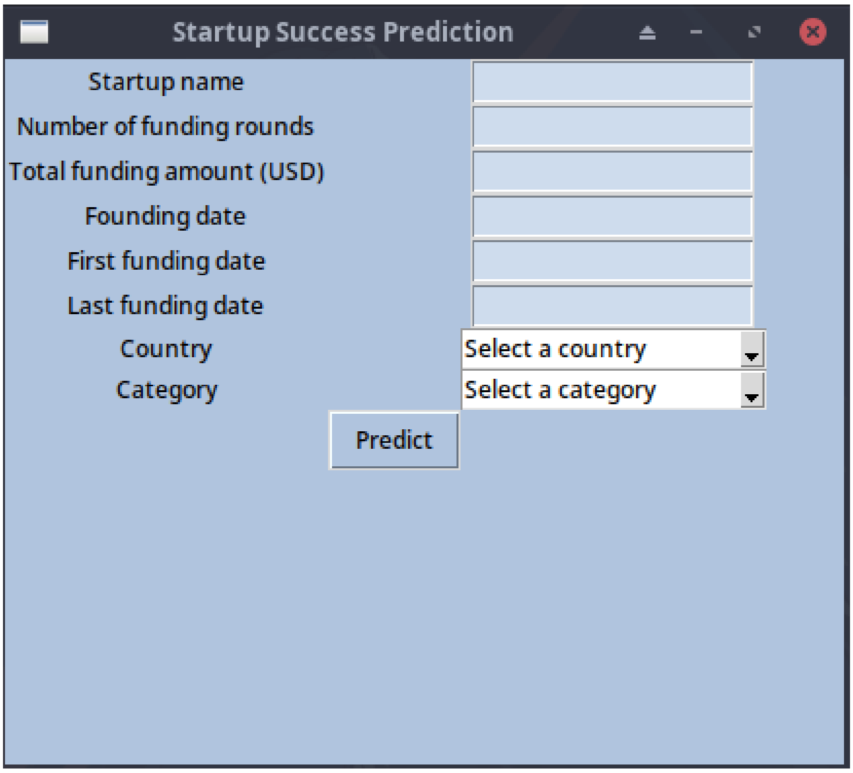


FIGURE III.8 – Main window

Once the required information fields are filled, the input data is processed and transformed into a format that can be utilized by our trained model. This ensures that the input aligns with the training data used by the model. Subsequently, a prediction is made based on the transformed input data.

III.8 Conclusion

In this chapter, we have presented the implementation details and results obtained from our experiments. We discussed the hardware and software specifications used, the tools employed, and evaluated the performance of the trained models. Additionally, we explored validation techniques to ensure the reliability and generalizability of the models.

One of the key aspects of our analysis was visualizing the feature importance to understand the factors that impact startup outcomes. We observed that the total amount of funding emerged as the most important feature in predicting success, as indicated by the highest score in both the XGBoost feature importance plot Figure III.3 and the LightGBM feature importance plot Figure III.2. This finding suggests that the level of funding received plays a crucial role in determining the success of startups.

Furthermore, we found that variables related to the timing and duration of funding were also significant predictors. The time to receive the first funding and the funding lifecycle demonstrated high importance in the models, as reflected by their high chi-square values. Additionally, the time between the first and the last funding, as well as the number of funding rounds, exhibited strong associations with the target variable according to the chi-square analysis.

In contrast, when considering information gain as the measure of feature importance, we observed that the country code variable emerged as highly significant to the model. This implies that the country in which a startup operates influences its chances of success, as different countries may have varying levels of support, resources, or market dynamics.

General Conclusion

The main objective of the present study was to generate a model to classify successful startups. By building a binary classification model to classify startups as successful or not-successful with an accuracy of 89,33% and an F1-score of 94,33% , it is assumed that the objective was achieved.

To achieve these results, we employed the LightGBM machine learning algorithm, known for its speed, interpretability, and effectiveness. The utilization of LightGBM allowed us to generate a model that not only produced good classification performance but also provided ease of interpretation and implementation.

Throughout our research, we utilized diverse techniques to preprocess the data and address challenges such as class imbalance. These techniques were instrumental in improving the overall performance and reliability of the model. By comprehensively outlining these processes, we aim to provide a valuable guide that can be followed by others interested in building similar startup classification models. By sharing our approach, we hope to contribute to the field and provide practical insights for future students, proffessionals, and decision-makers.

Throughout our study, we have also gained valuable insights into the factors that influence startup outcomes.

A key finding from our analysis is the significant role played by funding in predicting startup success. Specifically, the total amount of funding received by startups emerged as the most important feature in our models, indicating its crucial impact on determining the success of startups. Furthermore, we found that variables related to the timing and duration of funding were also influential predictors. The time to receive the first funding and the funding lifecycle demonstrated high importance in our models, suggesting their strong association with startup success. Additionally, the time between the first and the last funding, as well as the number of funding rounds, showed noteworthy associations with the target variable. In considering information gain as a measure of feature importance, we observed that the country code variable emerged as highly significant to the model. This finding implies that the country in which a startup operates plays a role in its chances of success, possibly due to varying levels of support, resources, or market dynamics across different countries.

Despite the insights gained from our study on startup success prediction, it is important to acknowledge its limitations. The data used in our analysis lacked detailed information, such as the time between funding rounds and specific recording dates. Furthermore, the restricted availability of comprehensive data from online platforms and organizations constrained our analysis. Access to better quality data could greatly enhance the performance of predictive models. Future studies addressing these limitations and utilizing higher-quality data have the potential to yield more accurate and reliable models for predicting startup success.

Future research should prioritize data collection efforts, particularly in the context of Algerian startups, to build startup success prediction models that are specific to the country's unique characteristics. This would involve gathering data that is rich, informative, and of high quality.

In addition, it is crucial to include more features that capture the complexities of the startup's business model. These could encompass variables such as revenue streams, customer segment, marketing channels, and industry-specific metrics. By incorporating these additional features, the models can provide a more comprehensive understanding of the factors driving startup success in Algeria.

An important aspect for future research is to explore advanced techniques for improving the model's performance and robustness while considering the Algerian context. This can involve investigating state-of-the-art machine learning algorithms, such as deep learning models or ensemble methods, to assess their effectiveness in predicting startup success. Additionally, leveraging advanced techniques like natural language processing (NLP) to analyze textual data associated with startups can significantly enhance the model's capabilities and overall robustness.

Furthermore, by focusing on data collection, quality enhancement, and the inclusion of relevant features, future research can develop models specifically tailored to the Algerian startup landscape. This comprehensive approach will enable policymakers, investors, and entrepreneurs to make more informed decisions based on reliable and insightful predictions.

Appendix

Appendix I : Data Visualization

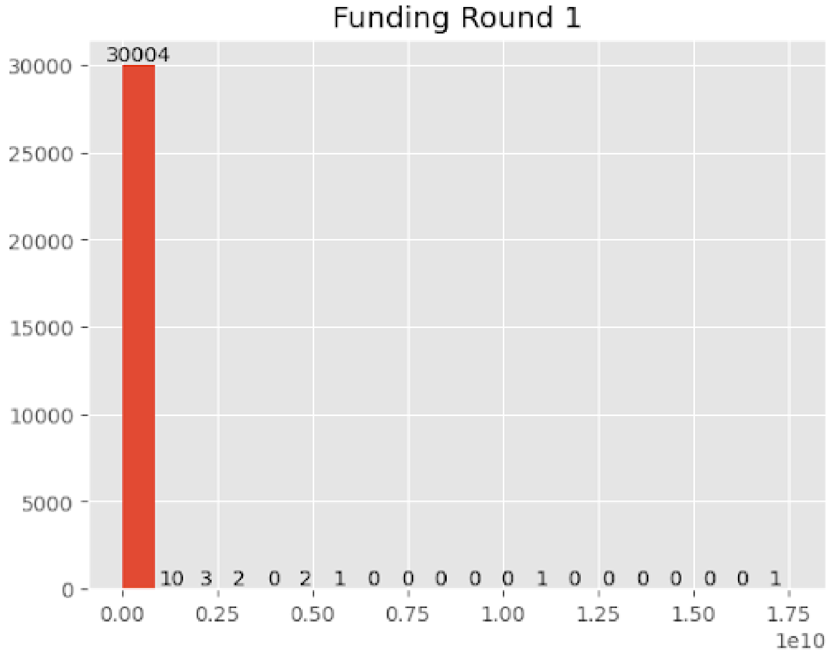


FIGURE III.9 – Distribution of funding total when N° of funding rounds is 1

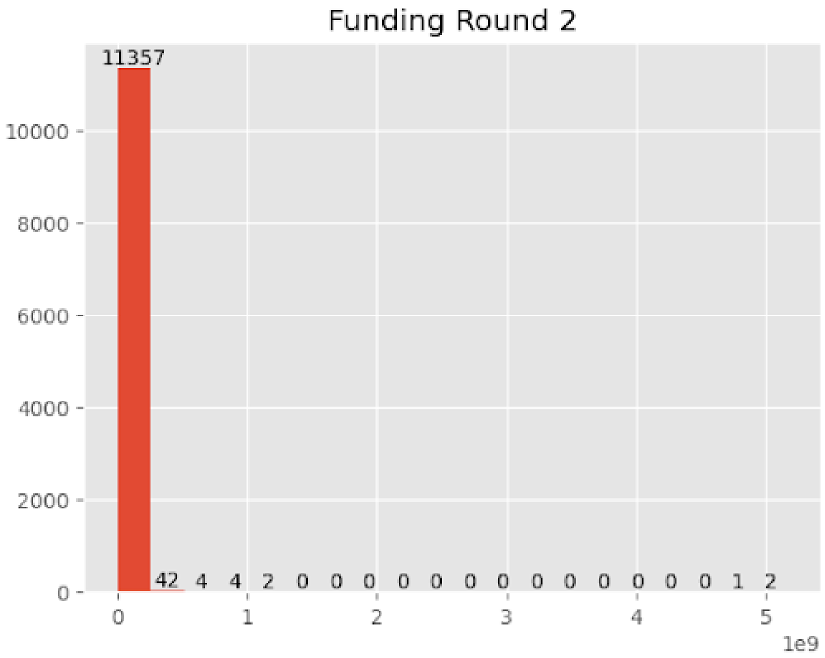


FIGURE III.10 – Distribution of funding total when N° of funding rounds is 2

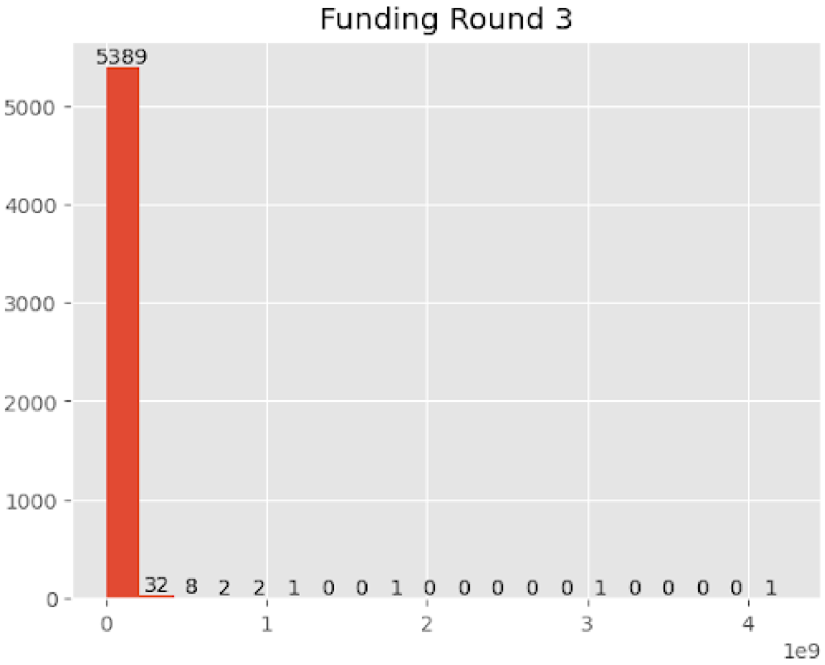


FIGURE III.11 – Distribution of funding total when N° of funding rounds is 3

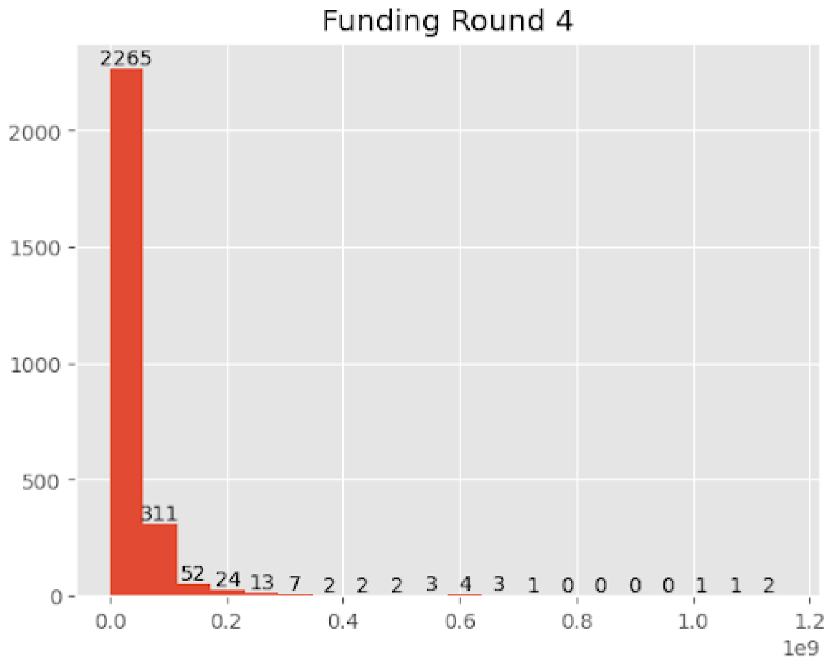


FIGURE III.12 – Distribution of funding total when N° of funding rounds is 4

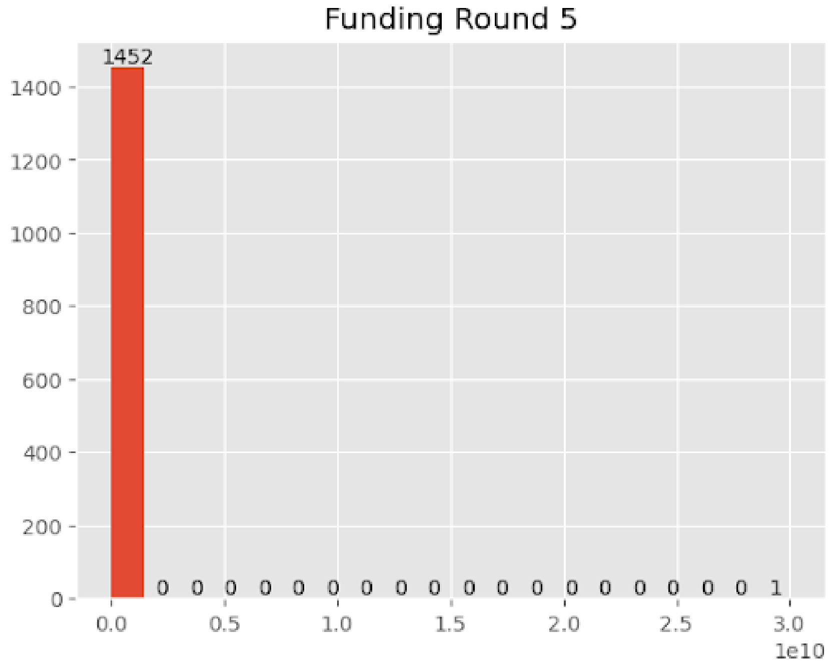


FIGURE III.13 – Distribution of funding total when N° of funding rounds is 5

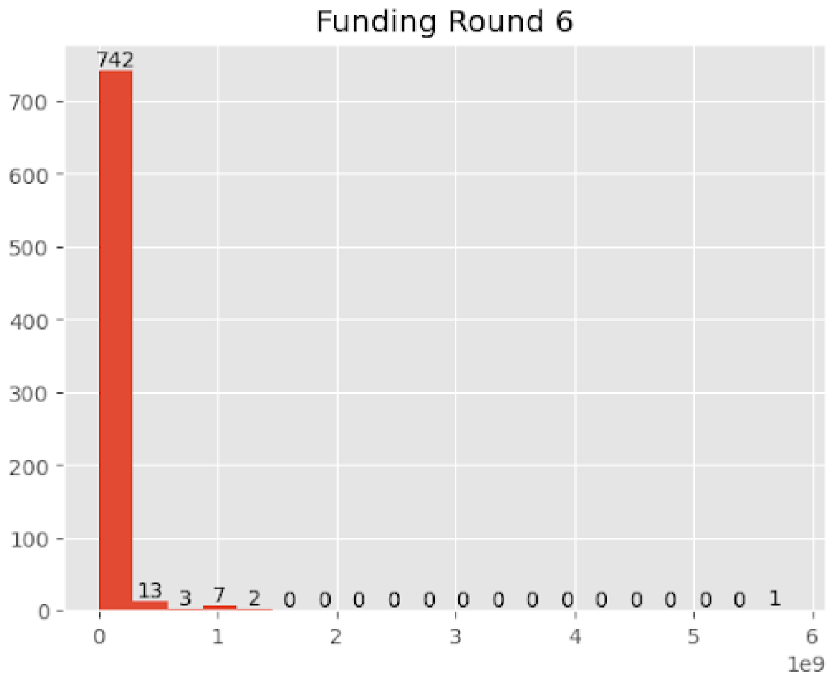


FIGURE III.14 – Distribution of funding total when N° of funding rounds is 6

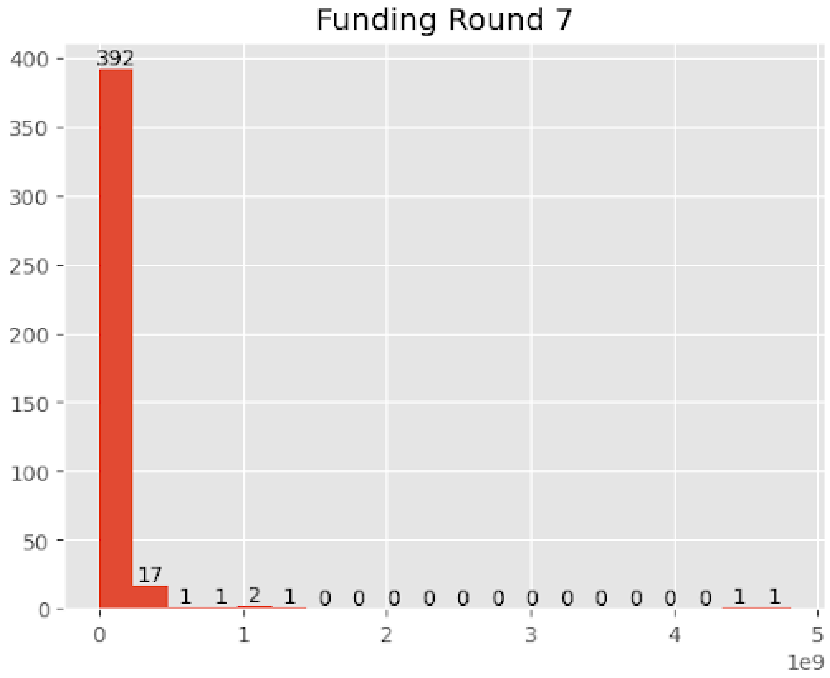


FIGURE III.15 – Distribution of funding total when N° of funding rounds is 7

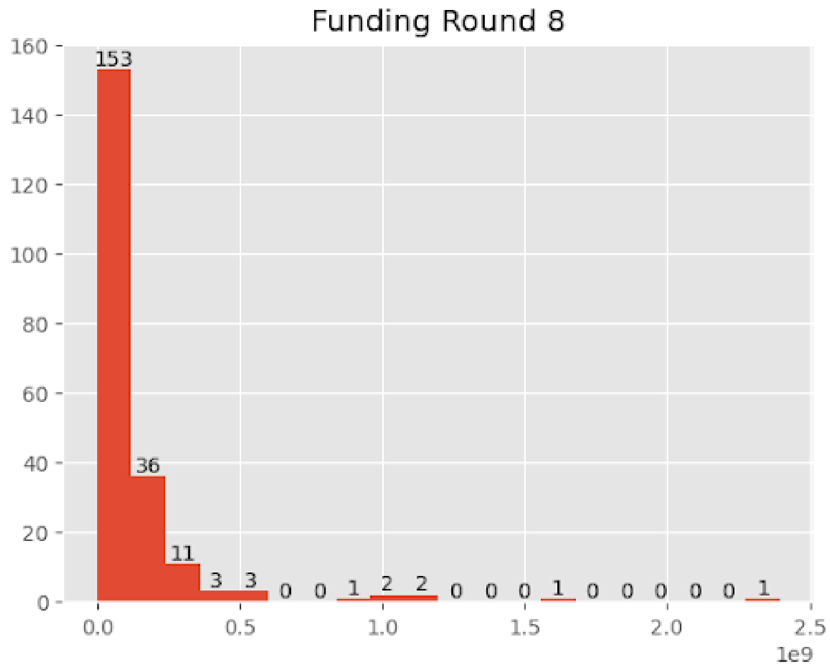


FIGURE III.16 – Distribution of funding total when N° of funding rounds is 8

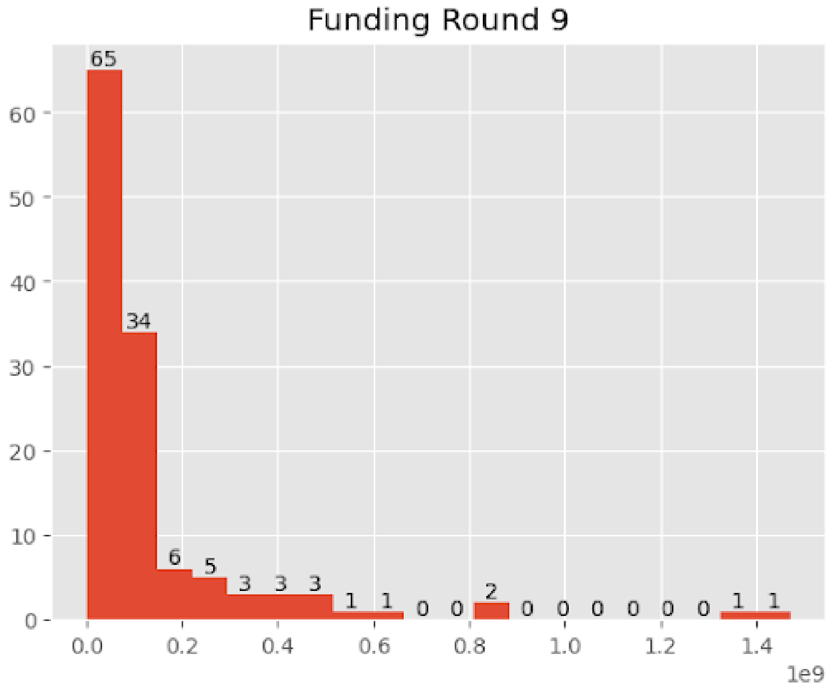


FIGURE III.17 – Distribution of funding total when N° of funding rounds is 9

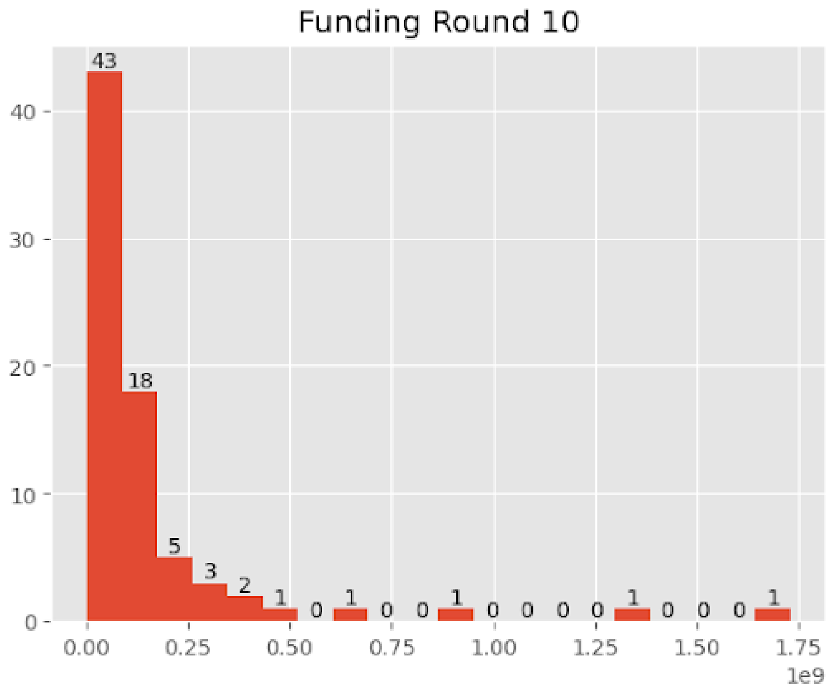


FIGURE III.18 – Distribution of funding total when N° of funding rounds is 10

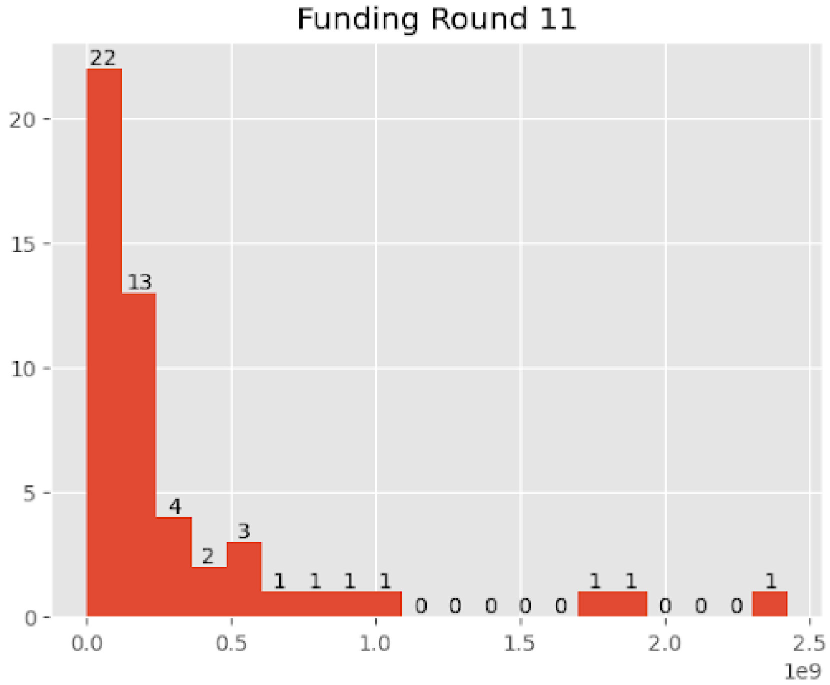


FIGURE III.19 – Distribution of funding total when N° of funding rounds is 11

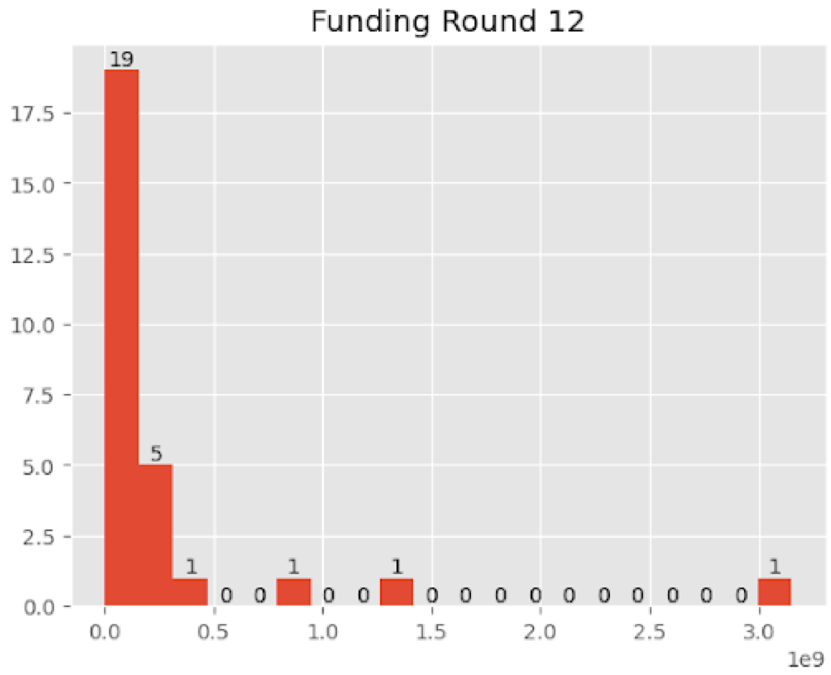


FIGURE III.20 – Distribution of funding total when N° of funding rounds is 12

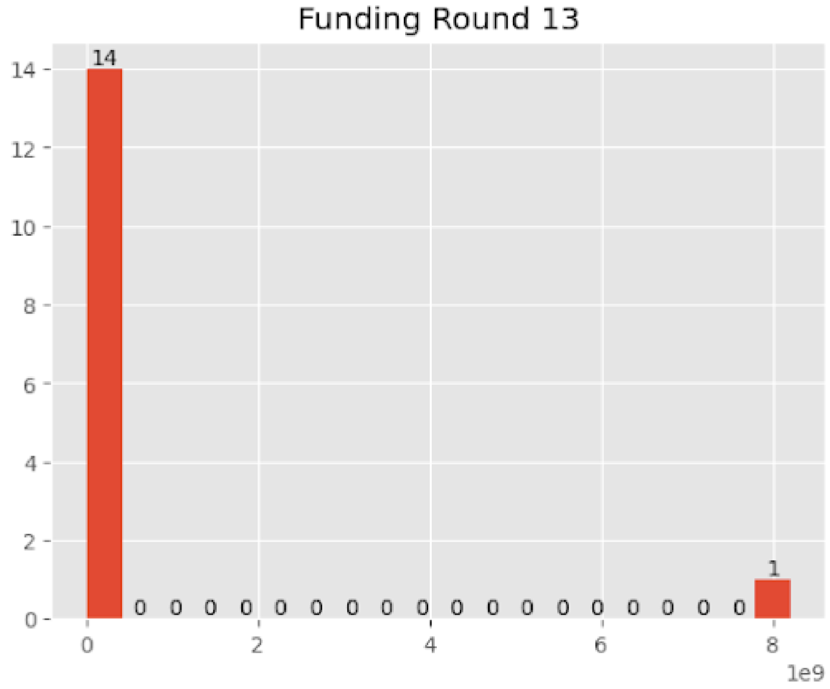


FIGURE III.21 – Distribution of funding total when N° of funding rounds is 13

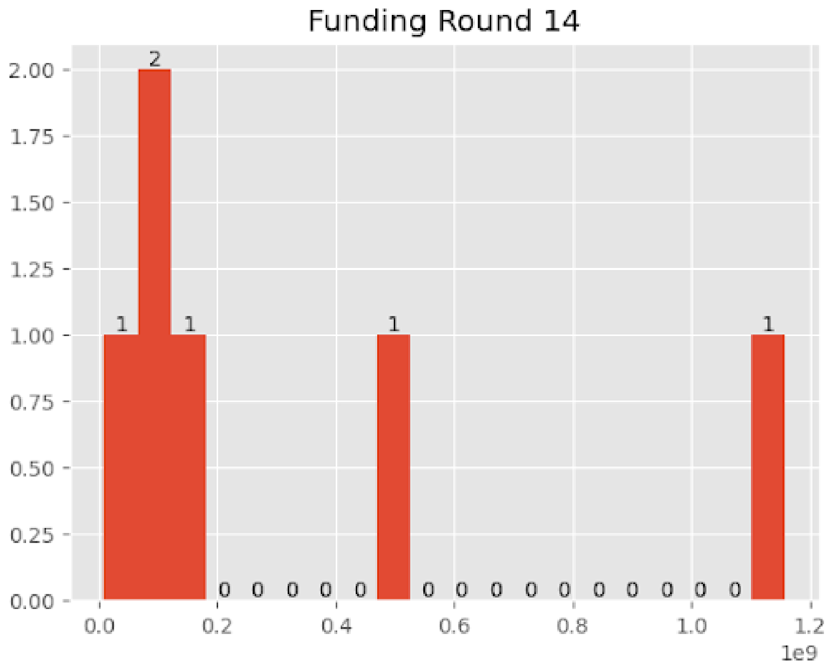


FIGURE III.22 – Distribution of funding total when N° of funding rounds is 14

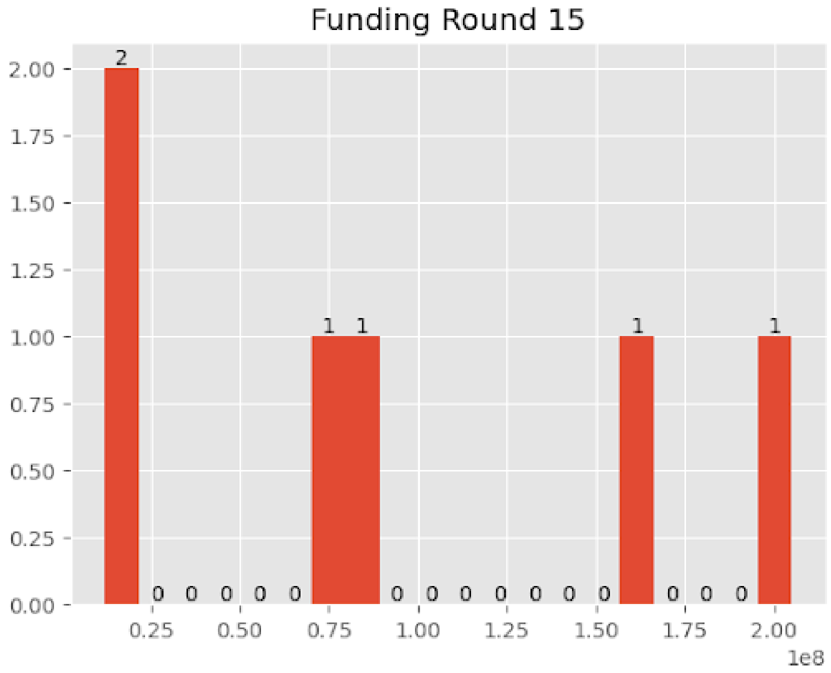


FIGURE III.23 – Distribution of funding total when N° of funding rounds is 15

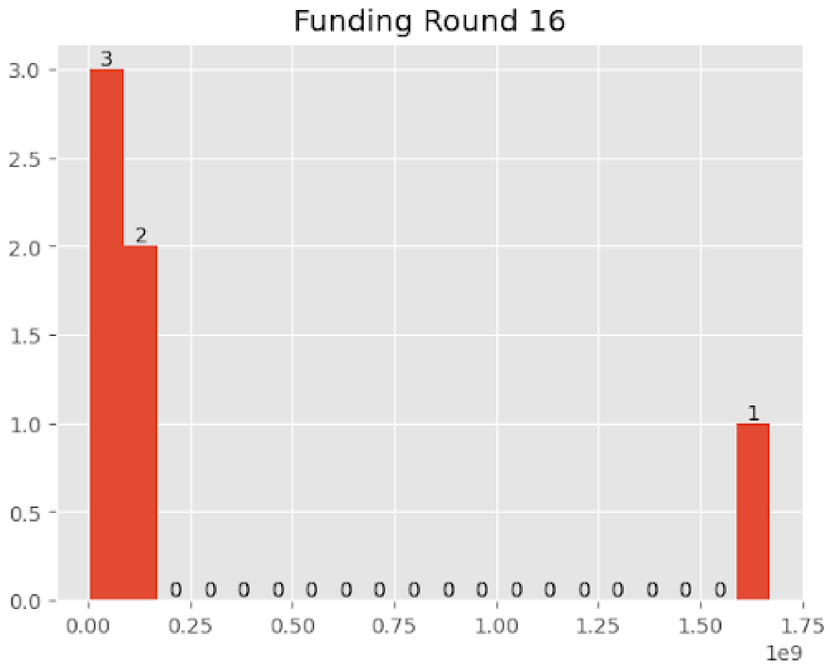


FIGURE III.24 – Distribution of funding total when N° of funding rounds is 16

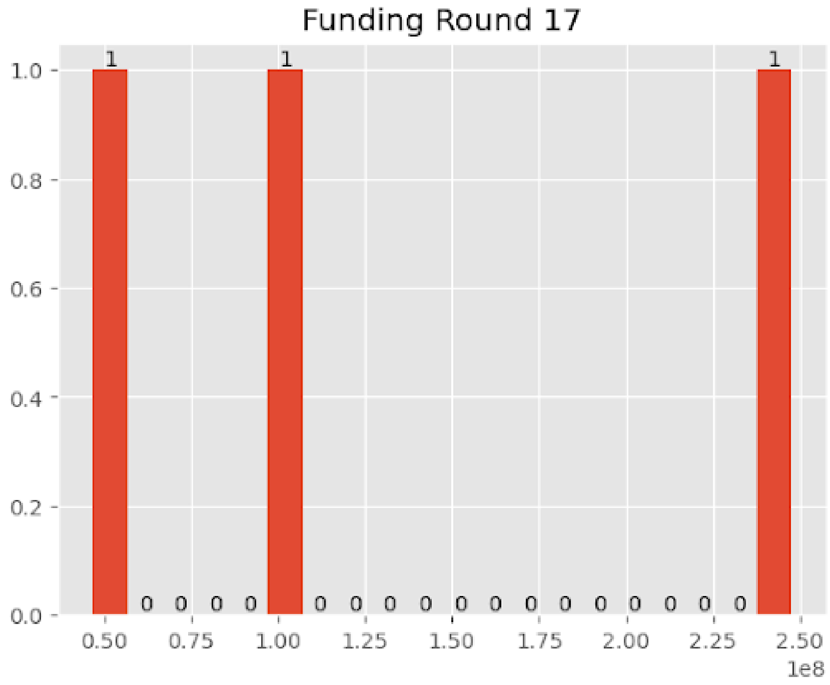


FIGURE III.25 – Distribution of funding total when N° of funding rounds is 17

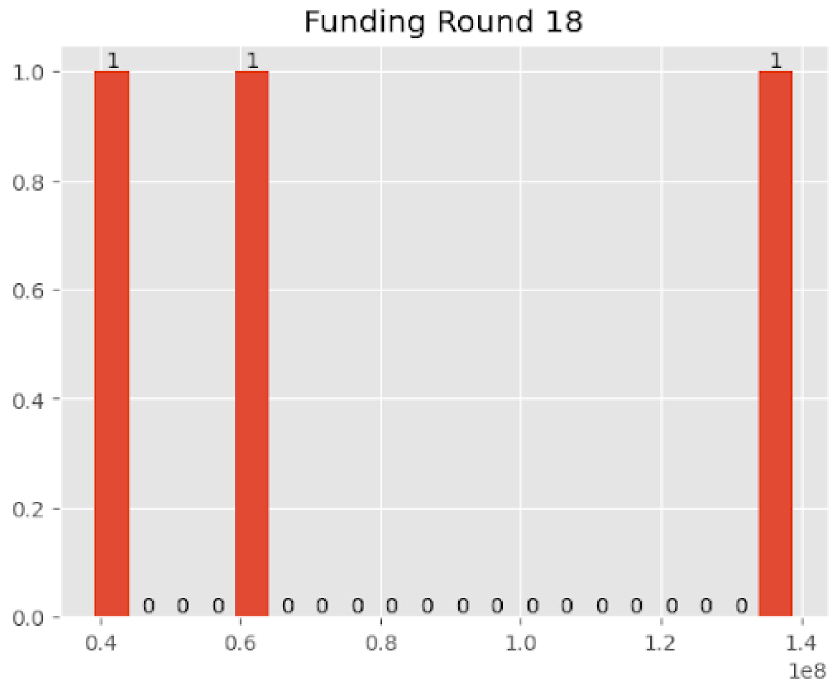


FIGURE III.26 – Distribution of funding total when N° of funding rounds is 18

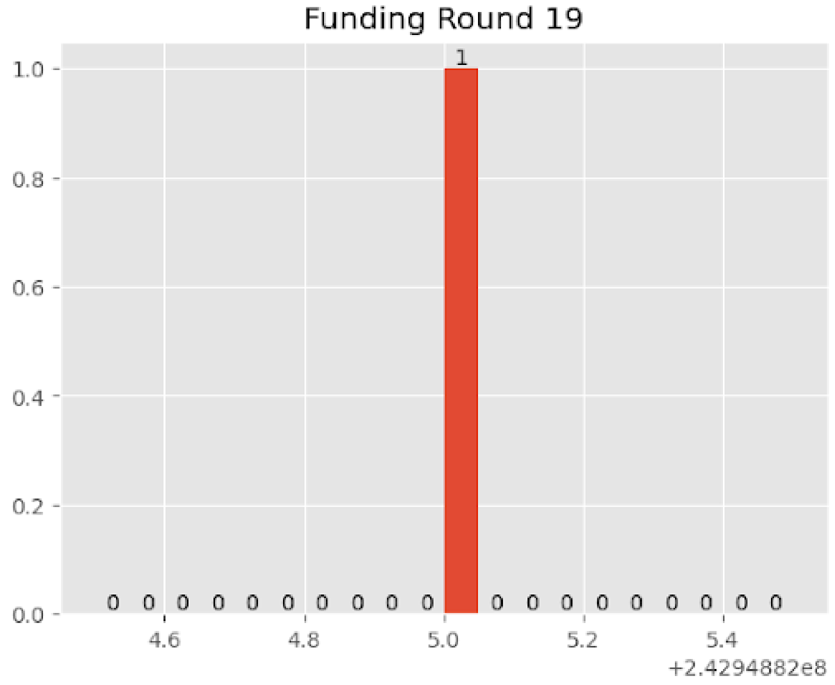


FIGURE III.27 – Distribution of funding total when N° of funding rounds is 19

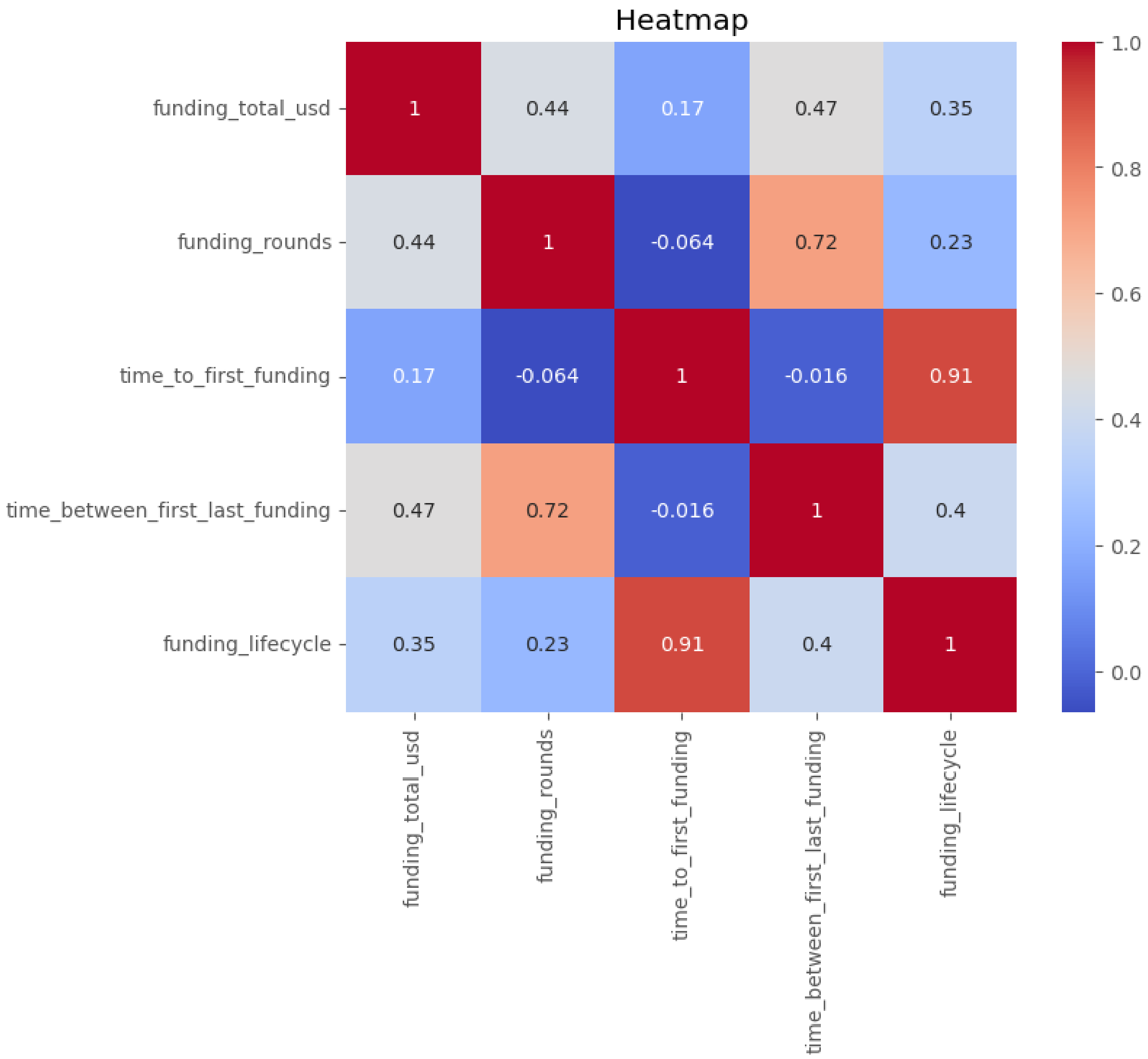


FIGURE III.28 – Heatmap

Appendix II : Used tools

Orange Orange is an open-source data analysis and visualization software with a user-friendly interface, offering tools for data exploration, preprocessing, and machine learning [49].

Signavio Signavio is a cloud-based BPM tool for modeling, analyzing, and optimizing business processes, with features for process documentation, collaboration, and performance monitoring [50].

Lucidchart Lucidchart is a user-friendly cloud-based diagramming tool that simplifies visual communication with its extensive library and collaborative features. It’s ideal for creating professional diagrams, flowcharts, and mind maps [51].

Bibliography

Bibliography

[1] Phil Budden, Fiona Murray, and Ogbogu Ukuku. Differentiating small enterprises in the innovation economy : Start-ups, new smes & other growth ventures, 2021.

[2] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. Supervised machine learning : A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1) :3–24, 2007.

[3] Ministry of Higher Education and Scientific Research of Algeria. Inciter les étudiants à concrétiser leurs idées innovantes. `https://www.mesrs.dz/index.php/fr/2022/12/06/inciter-les-etudiants-a-concretiser-leurs-idees-innovantes-2/`, December 2022.

[4] Startup Genome. The state of the global startup economy. `https://startupgenome.com/article/the-state-of-the-global-startup-economy`, 2022. Accessed on 28 February 2023.

[5] Aileen Lee. Welcome to the unicorn club, 2015 : Learning from billion-dollar companies, Jul 19 2015. Copyright - Copyright AOL Inc. Jul 19, 2015 ; Last updated - 2022-11-08.

[6] CB Insights. CB Insights - Unicorn Companies. `https://www.cbinsights.com/research-unicorn-companies`, Accessed 2023.

[7] Eric Ries. *The Lean Startup : How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. Crown Business, 2011.

[8] Steve Blank. Why companies are not startups. `https://steveblank.com/2014/03/04/why-companies-are-not-startups/`, March 2014. Accessed on March 18, 2023.

[9] Paul Graham. Startup = growth. `http://www.paulgraham.com/growth.html`, July 2012. Accessed on March 18, 2023.

[10] Muataz HF AlHazza, Islam Faisal Bourini, MH Zubaidah, and Norsyafira Bt Selamat. Success factor in new product development for startup companies using fuzzy logic approach. In *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pages 1–5. IEEE, 2019.

[11] Jia Liu and Dairui Li. The life cycle of initial public offering companies in china. *Journal of Applied Accounting Research*, 2014.

[12] Ayesha Alam, Sana Khan, and Fareeha Zafar. Strategic management : Managing mergers and acquisitions. *International Journal of BRIC Business Research (IJBBR)*, 3(1) :1–10, 2014.

[13] Aija Vonoga. Start-ups–an element for economic growth and innovativeness. *Latgale National Economy Research*, 1(10) :159–167, 2018.

[14] Arthur L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3) :210–229, 1959.

[15] Tom Mitchell. *Machine Learning*. 1997.

[16] Mahesh Batta and Others. Machine learning algorithms : A review. *ResearchGate*, 2020.

[17] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee, 2016.

[18] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning : From theory to algorithms*. Cambridge university press, 2014.

[19] Amar Krishna, Ankit Agrawal, and Alok Choudhary. Predicting the outcome of startups : less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 798–805. IEEE, 2016.

[20] Eliganti Ramalakshmi and Sindhuja Reddy Kamidi. Predictions for startups. *International Journal of Engineering & Technology*, 7(3.12), 2018.

[21] FY Osisanwo, JET Akinsola, O Awodele, JO Hinmikaiye, O Olakanmi, J Akinjobi, et al. Supervised machine learning algorithms : classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3) :128–138, 2017.

[22] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman. Decision trees : An overview and their use in medicine. *Journal of Medical Systems*, 26(5) :445–463, 2002.

[23] Wolfgang Fuhl, Gjergji Kasneci, Wolfgang Rosenstiel, and Enkeljda Kasneci. Training decision trees as replacement for convolution layers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3882–3889, 2020.

[24] Yu Zhan, Yuzhou Luo, Xunfei Deng, Kaishan Zhang, Minghua Zhang, Michael L. Grieneisen, and Baofeng Di. Satellite-based estimates of daily no2 exposure in china using hybrid random forest and spatiotemporal kriging model. *Environmental Science & Technology*, 52(7) :4180–4189, 2018.

[25] Hongwei Chen and Lun Chen. An application of xgboost algorithm for online transaction fraud detection based on improved sailfish optimizer. In *2022 4th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pages 294–299, 2022.

[26] Fatimah Alzamzami, Mohamad Hoda, and Abdulmotaleb El Saddik. Light gradient boosting machine for general sentiment classification on short texts : A comparative evaluation. *IEEE Access*, 8 :101840–101858, 2020.

[27] Stephenie C Lemon, Jason Roy, Melissa A Clark, Peter D Friedmann, and William Rakowski. Classification and regression tree analysis in public health : methodological review and comparison with logistic regression. *Annals of behavioral medicine*, 26 :172–181, 2003.

[28] Giuseppe Amato and Fabrizio Falchi. Knn based image classification relying on local feature similarity. In *Proceedings of the Third International Conference on SImilarity Search and APplications*, SISAP '10, page 101–108, New York, NY, USA, 2010. Association for Computing Machinery.

[29] Sasan Karamizadeh, Shahidan M Abdullah, Mehran Halimi, Jafar Shayan, and Mohammad javad Rajabi. Advantage and drawback of support vector machine functionality. In *2014 international conference on computer, communications, and control technology (I4CT)*, pages 63–65. IEEE, 2014.

[30] Maad M Mijwel. Artificial neural networks advantages and disadvantages. *Retrieved from LinkedIn https//www. linkedin. com/pulse/artificial-neuralnet Work*, 2018.

[31] Greg Ross, Sanjiv Das, Daniel Sciro, and Hussain Raza. Capitalvx : a machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*, 7 :94–114, 2021.

[32] Cemre Ünal and Ioana Ceasu. A machine learning approach towards startup success prediction. Technical report, IRTG 1792 Discussion Paper, 2019.

[33] Lele Cao, Vilhelm von Ehrenheim, Sebastian Krakowski, Xiaoxue Li, and Alexandra Lutz. Using deep learning to find the next unicorn : A practical synthesis. *arXiv preprint arXiv :2210.14195*, 2022.

[34] Yu Qian Ang, Andrew Chia, and Soroush Saghafian. *Using machine learning to demystify startups' funding, post-money valuation, and success*. Springer, 2022.

[35] Malhar Bangdiwala, Yashvi Mehta, Smrithi Agrawal, and Sunil Ghane. Predicting success rate of startups using machine learning algorithms. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6. IEEE, 2022.

[36] Javier Arroyo, Francesco Corea, Guillermo Jimenez-Diaz, and Juan A Recio-Garcia. Assessment of machine learning performance for decision support in venture capital investments. *Ieee Access*, 7 :124233–124243, 2019.

[37] Dafei Yin, Jing Li, and Gaosheng Wu. Solving the data sparsity problem in predicting the success of the startups with machine learning methods. *arXiv preprint arXiv :2112.07985*, 2021.

[38] Chenchen Pan, Yuan Gao, and Yuzi Luo. Machine learning prediction of companies' business success. *CS229 : Machine Learning, Fall 2018, Stanford University, CA*, 2018.

[39] Dominik Dellermann, Nikolaus Lipusch, Philipp Ebel, Karl Michael Popp, and Jan Marco Leimeister. Finding the unicorn : Predicting early stage startup success through a hybrid intelligence method. *arXiv preprint arXiv :2105.03360*, 2021.

[40] Boris Sharchilev, Michael Roizner, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov, and Maarten de Rijke. Web-based startup success prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 2283–2291, 2018.

[41] Markus Böhm, Jörg Weking, Frank Fortunat, Simon Müller, Isabell Welpe, and Helmut Krcmar. The business model dna : Towards an approach for predicting business model success. 2017.

[42] Marco Cantamessa, Valentina Gatteschi, Guido Perboli, and Mariangela Rosano. Startups' roads to failure. *Sustainability*, 10(7) :2346, 2018.

[43] Marco van Gelderen, Roy Thurik, and Niels Bosma. Success and risk factors in the pre-startup phase. *Small business economics*, 24 :365–380, 2005.

[44] Ulrich Kaiser and Johan M Kuhn. The value of publicly available, textual and non-textual information for startup performance prediction. *Journal of Business Venturing Insights*, 14 :e00179, 2020.

[45] Jen-Yin Yeh and Chi-Hua Chen. A machine learning approach to predict the success of crowdfunding fintech project. *Journal of Enterprise Information Management*, (ahead-of-print), 2020.

[46] VSCodium GitHub Repository. `https://github.com/VSCodium/vscodium`, Accessed 2023.

[47] Jupyter : Interactive computing. `https://jupyter.org/`. Accessed 2023.

[48] Pycharm : Python ide for professional developers. `https://www.jetbrains.com/pycharm/`. Accessed 2023.

[49] Orange : Open-source data analysis and visualization. `https://orange.biolab.si/`, Accessed 2023.

[50] Signavio. `https://www.signavio.com/`, Accessed 2023.

[51] Lucidchart. `https://www.lucidchart.com/`, 2023.

Abstract

English

Startups play a crucial role in driving economic growth, innovation, and job creation. However, the uncertain future of these projects raises concerns about their success rates. Statistics indicates that approximately 90% of startups fail, highlighting the challenges they face, such as financial resources, team dynamics, and market demand. To address these challenges, machine learning and artificial intelligence have gained interest in predicting startup success. This study aims to develop a predictive model using diverse machine learning techniques to classify startups as successful or not. It explores binary and multi-class classification approaches and evaluates various algorithms to determine their effectiveness. By contributing to the understanding of success drivers and providing insights for decision-makers, investors, and entrepreneurs, this research aims to advance the startup ecosystem in Algeria. Our analysis revealed that funding plays a significant role in determining success, with the amount of funding, its timing, and duration being highly influential factors. Additionally, the country in which a startup operates also influences its chances of success. These insights contribute to understanding success drivers and provide valuable guidance for decision-makers, investors, and entrepreneurs in advancing the startup ecosystem.

Keywords : Startup, startup ecosystem, Mergers and Acquisitions (M&A), Initial Public Offering (IPO), startup success, machine learning, prediction, classification, success factors, success rate, funding.

Résumé

Les startups jouent un rôle crucial dans la croissance économique, l'innovation et la création d'emplois. Toutefois, l'avenir incertain de ces projets suscite des inquiétudes quant à leur taux de réussite. Les statistiques indiquent qu'environ 90% des startups échouent, ce qui met en évidence les défis auxquels elles sont confrontées, tels que les ressources financières, la dynamique d'équipe et la demande du marché. Pour relever ces défis, l'apprentissage automatique et l'intelligence artificielle ont gagné en intérêt pour prédire le succès des startups. Cette étude vise à développer un modèle prédictif utilisant diverses techniques d'apprentissage automatique pour classer les startups comme réussies ou non. Elle explore les approches de classification binaire et multi-classes et évalue divers algorithmes pour déterminer leur efficacité. En contribuant à la compréhension des facteurs de réussite et en fournissant des informations aux décideurs, aux investisseurs et aux entrepreneurs, cette recherche vise à faire progresser l'écosystème des startups en Algérie. Notre analyse a révélé que le financement joue un rôle important dans la détermination du succès, le montant du financement, son calendrier et sa durée étant des facteurs très influents. En outre, le pays dans lequel une startup opère influence également ses chances de succès. Ces informations permettent de mieux comprendre les facteurs de réussite et fournissent des indications précieuses aux décideurs, investisseurs et entrepreneurs pour faire progresser l'écosystème des startups.

Mots clés : Startup, écosystème des startups, fusions et acquisitions, introduction en bourse, succès des startups, apprentissage automatique, prédiction, classification, facteurs de succès, taux de réussite, financement.

ملخص

تلعب الشركات الناشئة دورا مهما في دفع عجلة النمو الاقتصادي و الابتكار و خلق فرص العمل. و مع ذلك, فإن المستقبل غير المؤكد لهذه المشاريع يثير مخاوف بشأن معدلات نجاحها, تشير الاحصائيات الى ان ما يقارب من 90 بالمئة من الشركات الناشئة تفشل, مما يبرز التحديات التي يواجهونها كموارد مالية و ديناميكيات الفريق و طلب السوق.

لمواجهة هذه التحديات, اكتسب الذكاء الاصطناعي اهتماما كبيرا فيها يخص التنبؤ بنجاح الشركة الناشئة. تهدف هذه الدراسة إلى تطوير نموذج تنبؤي باستخدام تقنيات التعلم الآلي المتنوعة لتصنيف الشركات الناشئة على أنها ناجحة أم لا.

خلال هذه الدراسة, تستكشف مناهج التصنيف الثنائية و متعددة الفئات و تقيم الخوارزميات المختلفة لتحديد فعاليتها في هذا المجال. من خلال المساهمة في فهم محركات النجاح و تقديم رؤى لصناع القرار والمستثمرين و رواد الأعمال, تهدف هاته الدراسة الى تعزيز بيئة الشركات الناشئة في الجزائر.

كشفت نتائج دراستنا أن التمويل يلعب دورا مهما في إمكانية النجاح, حيث أن مقدار التمويل و توقيته و مدته هي عوامل مؤثرة للغاية. بالإضافة الى ذلك, البلد التي تعمل فيه الشركة الناشئة يلعب دورا كبيرا في تحديد فرص نجاحها.

تساهم هذه الأفكار في فهم دوافع النجاح و توفر إرشادات قيمة لصانعي القرار والمستثمرين و رواد الأعمال في تطوير النظام البيئي للشركات الناشئة.

الكلمات المفتاحية: مؤسسة ناشئة, النظام البيئي, عمليات الدمج والاستحواذ (M&A), التعلم الآلي, التنبؤ, التصنيف, عوامل النجاح, التمويل.