

A dissertation submitted to the *University of Greenwich*
in partial fulfilment of the requirements for the Degree of

Master of Science

in

Big data & Business Intelligence

**Predictive Analytics for Resources
Allocation using Python and Machine
Learning models.**

Name: Ashwini Basavanahalli Nagaraju

Student ID: 001352752-5

Supervisor: Prof. Pushparajah Rajaguru

Submission date: 6 September 2024

Word Count: 14520

Abstract

This paper focuses on the application of machine learning classification techniques for the predictive analysis on the allocation of resources with the aim of improving on the decision-making system. The main goal is to evaluate and rank the performance of various models in the task of identifying the best allocation of resources, by means of several parameters, namely Precision, Recall, the F1 score, and Accuracy. Some of the models with their acronyms that have been assessed entail the following; Gradient Boosting, Random Forest, Ada Boost, Extra Trees, Decision Tree, K Nearest Neighbour (KNN), Gaussian Naive Bayes (Gaussian NB), Bernoulli Naive Bayes (Bernoulli), Support Vector Classification (SVC), Logistic Regression, SGD Classifier, and a Hard Vote Classifier. Gradient Boosting model was ranked highest in Precision [0. 8889], F1 Score [0. 8000] and Accuracy [0. 8095] that shows that this model is good in identifying the positive cases and measuring the balance of Precision and Recall. Random Forest and Extra Trees likewise showed good performance, their Precision and Accuracy at par with that of Gradient Boosting but a lower Recall that had an impact on their F1 Scores. However, on using Regression models such as KNN and Gaussian NB they were unable to perform so well on Recall and F1 Scores which indicated restricted abilities to identify all the positive cases. In returning, by nature of the simplistic and understandable structure of Knowledge and Decision Tree models, the trials were not able to match up to the other more complex models, having lesser Precision and F1 Score.

The study reveals the key advantages as well as crucial disadvantages of using each model, together with the fact that these advantages and disadvantages are inherently linked with the use of the Precision, Recall, and Accuracy metrics. As such, while models such as the Decision Trees and KNN are useful for their ability to explain their results, the Gradient Boosting and the Extra Trees methods provide more accurate solutions to the problem of precise and exhaustive resource allocation.

Keywords: Predictive Analysis, Resource Allocation, Machine Learning, Python, Data Preprocessing, Algorithm Selection, Feature Engineering, Model Training.

Acknowledgements

I would like to express my sincere gratitude to Prof. Pushparajah Rajaguru for agreeing to be my supervisor and for their invaluable guidance, constant encouragement and insightful feedback throughout my research and the completion of my project Predictive Analytics for Resources Allocation Using Python and Machine Learning models. Thank you for agreeing to have the project demonstration on 10 September 2024.

I would like to thank my colleagues for their constant support and guidance for completing this work easily. They encouraged and gave me hopes to do further research on the topic which showed me the wider knowledge of this project.

I would also like to thank my university for providing the necessary facilities and resources that allowed me to carry out my research effectively.

Table of Contents

Abstract.....	1
Acknowledgements.....	2
List of Figures.....	5
Chapter 1. Introduction	1
1.1 Overview	1
1.2 Background	2
1.3 Problem Statement	3
1.4 Research Objectives	3
1.5 Research Questions	4
1.6 Significance of the Study	4
1.7 Scope and Limitations	4
1.8 Theoretical Foundations	5
1.9 Current Practices and Challenges in Resource Allocation	5
1.10 Framework for Implementation.....	6
1.11 Barriers to Implementation and Solutions	7
Chapter 2. Literature Review	8
2.1. Overview	8
2.2. Predictive Analytics: Conceptual Foundations and Applications.....	8
2.3. Integrating Predictive Analytics using python and machine learning models.....	9
2.4. Challenges in Implementing Predictive Analytics.	10
2.5. Empirical Studies	10
2.6 Limitations	11
Chapter 3. Analysis of the system.....	12
3.1. Legal issues.	12
3.2. Social Issues.	13
3.3. Ethical Issues.....	13
3.4. Professional Issues.	14
Chapter 4. Designing of the system	15
4.1. System Requirements	15
4.2. System Architecture	15
4.3. Data Management.....	17
4.4. Model Selection.....	17

4.5. System Integration.....	17
4.6. Validation and Testing	17
Chapter 5. Implementation.....	19
5.1. Data Collection.....	19
5.2. Data Pre-processing.....	21
5.3. Data Visualization with Python	22
Chapter 6. Testing	24
Chapter 7. Evaluation of the system	26
7.1. Introduction	26
7.2. Model Evaluation metrics	26
7.3. Model Performance analysis	27
7.3.1 Gradient Boosting Algorithm	29
7.3.2 Random Forest	30
7.3.3 AdaBoost	32
7.3.4 Extra Trees.....	33
7.3.5 Decision Tree.....	34
7.3.6. K-Nearest Neighbors (KNN).....	37
7.3.7. Gaussian Naive Bayes (Gaussian NB)	39
7.3.8. Bernoulli Naïve Bayes (Bernoulli).....	40
7.3.9 Support Vector Classification (SVC).....	41
7.3.10 Logistic Regression	42
7.3.11 Stochastic Gradient Descent Classifier (Declassified)	44
7.3.12 Hard Voting Classifier	45
Chapter 8. Results and Discussion.....	47
Chapter 9. Conclusion.....	49
9.1. Overview	49
9.2. Recommendation.....	49
9.3. Novelty of the Work	51
9.4 Future Work.....	52
References.....	53
Bibliography	56

List of Figures

Figure 1 System flow diagram.	16
Figure 2 Screenshot of the dataset website.....	20
Figure 3 dataset image screenshot in excel	21
Figure 4: Accuracy of Gradient Boosting.....	30
Figure 5 Confusion Metrix GB	30
Figure 6: Accuracy of RF	32
Figure 7: Confusion Metrix RF	32
Figure 9: AdaBoost Accuracy.....	33
Figure 10:Confusion Metrix AdaBoost	33
Figure 11:Accuracy of ET	34
Figure 12:Confusion Metrix of ET.....	34
Figure 13: Accuracy of DT.....	36
Figure 14:Confusion Metrix of DT	36
Figure 15: Decision Tree	37
Figure 16: KNN accuracy.....	38
Figure 17: Confusion Metrix of KNN	39
Figure 18: Accuracy of GaussianNB.....	40
Figure 19: Confusion Metrix of GNB	40
Figure 20: Accuracy of BNB.....	41
Figure 21:Confusion Metrix of NB	41
Figure 22: Accuracy of SVC	42
Figure 23: Confusion Metrix of SVC.....	42
Figure 24:Accuracy of LR.....	43
Figure 25: Confusion Metrix of LR	43
Figure 26: Accuracy of SGD	44
Figure 27:Confusion Metrix of SGD.....	45
Figure 28:Accuracy of HVC	45
Figure 29:Confusion Metrix HVC	46
Figure 30: Best Model.....	48

Chapter 1. Introduction

1.1 Overview

The resource has a large share of scarce resources and in business environments where everything is quickly moving into data processing, competition advantage, operational efficiency is depended on how organizations best managing their recourse (human or processors). Increasingly, organizational resource allocation decisions are being augmented using predictive analytics tools. Predictive analytics extracts information from data and uses it to predict future trends, behaviours, and events. Instead, machine learning tools refer to a set of applications and technologies designed to collect, store, extract/analysis data (process) and display them so that meaningful insights can be derived helping managers in informed decision making (McAfee and Brynjolfsson, 2012).

This is a major departure from the old form of predictive analytics where resources were allocated according to static data and managerial hunches. This move towards virtually real-time data-driven decisions enables organizations from being reactive to proactive, by predicting problems or opportunities before they occur. The resulting gains are productivity increases for organizations translating into cost savings and better performance (Davenport and Harris, 2007).

But for all the clear benefits, not everyone is jumping on the predictive analytics and Python tool when it comes to resource allocation. The most significant reasons why current use has not surpassed 17% is due to data quality problems, the difficulty in integrating different sets of available information, lack of specialists and resistance to change. Solving these problems demands both the depth in theoretical foundations and breadth of hands-on experience with these technologies.

The goal of this thesis is to investigate the synergetic use of predictive analytics in resource allocation optimization within organizations. The research aims to take a closer look at the advantages, challenges and work arounds associated with these technologies. This research aims to provide valuable findings for both academic researchers and management practitioners by presenting a conceptual roadmap from the literature which suggests a framework on how effective implementation can take place as well offering possible solutions to common impediments. This study would help organizations in realizing the importance of data to plan their strategic resource

allocation and aiming at sustainable success under competitive market conditions (Chen, Chiang, and Storey, 2012).

1.2 Background

This heralds a new era, in which data is the primary asset of organizations and integration of predictive analytics with Python tools for resource allocation will become a game changer amongst practicing operational efficiency and strategic advantage. Predictive analytics taps into the power of statistical algorithms and machine learning techniques to predict what will happen next based on historical data. This is Combined with python tools that provide the ability to aggregate, analyse and visualize data so organizations can make informed decisions ranging from manufacturing to healthcare (Ramsbotham et al. 2017)

In these days of the modern, grimy business world this is especially true with predictive analytics and resource allocation. Resource management is the ability to allocate and deploy resources in a manner that achieves optimization, while maintaining balance between supply/consumption of levels. Every organization needs resource management plan-be it tangible or intangible (materials like financials, humans etc.). For cost driven organizations across varied industry sectors ensuring optimal utilization on available tools. Predictive analytics will help you forecast demand in various domains, including libraries (Mood,2024), providing a structured approach to managing resources more efficiently.

Python tools assist this process by offering data integration, visualization, and reporting platform. With these tools, organizations can compile data from various sources to visualize interactive dashboards and derive meaningful insights. Through this, decision-makers can get comparatively real-time and trend analysis reports which are crucial to plan strategy or resources. This leads to a win-win situation where application of machine learning models along with predictive analytics creates a strong targeted allocation mechanism in place and organizations can run more efficiently enabling better results (LaValle et al., 2011).

1.3 Problem Statement

On the other hand, a plethora of organizations is unable to help itself in implementing this technology - as proven by unimaginable benefits of consolidating predictive analytics for resource allocation. Facing challenges such as poor data quality and integration, a dearth of trained personnel, or resistance to change. Also, the predictive model's complexity (and its ongoing monitoring and updating) is likely to be intimidating for many organizations. The objective of this dissertation is to solve these issues and overlap by identifying practical use-cases & benefits across predictive analytics, in resource delta space.

At its heart, this research seeks to optimize resource allocation in organizations using new methods in python for predictive analytics. This research aims to provide an understanding of the best way these respective technologies can be used in order to improve decision-making, efficiency, and competitive advantage. It will also work to identify the hurdles that come in way of successful implementation and provide suggestions on how they can be overcome.

1.4 Research Objectives

The first and the foremost aim of this study is to generate robust evidence from predictive analytics, how resource allocation at organizational level have made better by employing python and machine learning tools. The specific aims are:

1. To introduce the basic principles and key elements supporting predictive analytics models, providing concepts on understanding how data-driven predictions should be executed.
2. To review current resource allocation practices within organizations, and identifying common challenges and inefficiencies plaguing an organization.
3. To explore the application of predictive analytics and python-based tools in developing accurate estimates for resource allocation, demonstrating how these methods can improve decision-making processes.
4. To develop a framework that integrates predictive analytics and python tools for optimizing resource allocation in the organisational settings.
5. To identify and articulate the challenges associated with implementing predictive analytics and python tools in resource allocation and proposing solutions.

1.5 Research Questions

To address the research objectives, this study attempts to answer the following questions;

1. How can Python's machine learning libraries and tools be utilized to preprocess, analyse, and model data to enhance the accuracy and efficiency of resource allocation predictions?
2. What are the key features and data attributes that significantly impact the accuracy of resource allocation predictions, and how can they be extracted and utilized in the model development process?
3. How can the predictive model be designed to adapt to dynamic changes in resource demand and supply, ensuring real-time or near-real-time allocation decisions?

1.6 Significance of the Study

Statement - This study is important because it could add to resource management tools by exemplifying how predictive analytics and machine learning tools can be used. This research hopes to aid organizations by highlighting the challenges and providing measures for successful implementation in their resource allocation processes, as explored by Wang and Hajli (2017). The results of the study might be beneficial for policy measures, strategic planning and operational improvements boosting institutional efficiency and competitiveness. This study also helps in academic literature of predictive analytics/ machine learning to bridge the gap between theory and practice. The guide is a deep dive into the technology, use cases and what it takes to leverage these successfully. They also offer a replicable model that can be used in other industries, increasing the likelihood of them having long-lasting effects.

1.7 Scope and Limitations

This paper looks at How to use predictive analytics in resource allocation within organizations. The study encompasses theoretical underpinnings, existing work, opportunities, and challenges in literature. Unfortunately, the research is valid only for companies that willing to have a vast amount of data and an infrastructure strong enough to support predictive analytics techniques besides common python tools. Furthermore, the paper does not provide insights into technical details in using predictive models or machine learning tools and focuses on their applications benefits.

1.8 Theoretical Foundations

Predictive analytics as theoretical concept is essentially born out of the marriage between data science & statistics, IT. These techniques include regression analysis, and machine learning algorithms to find patterns in data which help companies make predictions or forecast. Result as an example benefits of impulse buying. Models trained using these methods can predict future events to some extent of precision (Waller and Fawcett, 2013). The process of collecting, aggregating, and analysing data from multiple sources is to create a unified view into an organization operation. Some of the key techniques used by python tools are data warehousing, data mining and OLAP - Online Analytical Processing which provide help to these decision-making processes.

The synthesis of these theoretical frameworks follows an inclusive research agenda outlined by Abbasi et al. (2016) sound base for optimal resource allocation. Organizations can use more powerful predictive models through the same method that they have been using all along so that with each data point those groups get a stronger and clearer sense of what is happening operationally, how to prepare for this year's need when budgeting time comes around again, because funds will be easier allocated.

1.9 Current Practices and Challenges in Resource Allocation

Historically, managerial intuition and the distillation of past data have been used in these decisions. Although this manner may work to a certain point, it could not completely encompass the intricacies and dynamics of recent enterprise environments. At the technological front predictive analytics could be cited as a model where one can use python and machine learning tools to find out insights from heaps of data so that businesses are able to get actionable information in near real time.

Data quality remains one of the greatest hurdles in resource allocation. Low Data Quality: With poor data, comes bad predictions and hence wrong decisions. To truly exploit the power of these technologies, data within an organization this maintained accurately, comprehensively, and consistently. Moreover, collaboration of varied data sources is challenging and time-consuming which needs strong-database management activities as well.

And again, an issue to deal with: The reason is there are not enough mastered technicians out there. Predictive analytics with machine learning tools, when effectively used come with the demand for specialized skills in data science, statistics, information technology etc. There is also a need for organizations to invest in training and development programs which help grow these capabilities within their workforce. One of the points that prevent successful implementation is resistance to change; this both on employees and managers, who many times will be unwilling to use a new tool or adopt different processes. This established resistance needs to be met by effective leadership, communication, and demonstration of the benefits these technologies provide.

1.10 Framework for Implementation

Organizations should take a structured approach to provide advance analytics and tools for resource allocation. This requires a few important steps:

- **Gathering and Integration:** The first step is to have data coming from different sources within the organization integrated in a way that why it fits consistently. This needs strong data management practice.
- **Develop and Validate Predictive Models:** The next step is to create predictive models based on historical data. The work includes choosing the right statistical tools and algorithms, training models testing their accuracy and Reliability.
- **Deployment and Monitoring:** Once the models are validated, they can be deployed to deliver predictions / insights. Organizations have to actively watch the models for effective performance and update them proactively if not they will lose accuracy, efficiency.
- **Reporting/Visualization:** After your predictive model produces the prediction or insight, these are reported in python tools. It includes dashboards and reports that allow the table to provide critical information in real-time.
- **Training and Change Management:** Companies should invest in training prevented tools in order the right coaching associated change management comes into here is how to very generate use of predictive analytics. That means developing the skills and capabilities in workforce that are needed, as well as overcoming any possible resistance to change.
- **Continuous Improvement -** Predictive analytics and python tools implementation is an ongoing activity that needs continuous improvement. This commits organizations to a

continuous review, refinement and improvement of these models and data management practices as well as the decision-making processes they enable.

1.11 Barriers to Implementation and Solutions

There are several obstacles to the adoption of predictive analytics for resource allocation. These include:

- **Data quality:** Low data quality increases the likelihood that predictions will be wrong and decisions suboptimal. To keep out the data attorneys, organizations must follow best practices and make sure their data is clean (Dey, Choudhury & Tan, 2020).
- **Shortage of Skilled workforce:** Efficient utilization of predictive analytics and python tools require an expert level skill set. Invest in developing these capabilities within your workforce through training and development programs.
- **Resistance to change:** Employees and managers may resist the use of new technologies or processes. This means coming with a strong management, able to rally the troops and communicate the benefits of these technologies.
- **Integrations Challenges:** Integrating the data from a variety of sources can be difficult and take time. Organizations must employ best practices for data integration and leverage modern data warehousing technologies.
- **Cost concern:** As machine learning and python tools are costly; they involve a lot of investment in technology & infrastructure. There are a few costs and benefits for organizations to consider, so companies should understand the business case of implementing it.

Chapter 2. Literature Review

2.1. Overview

Predictive analytics and python tools for use in resource allocation strategies is now a game changer in modern corporate practice. This chapter reviews the extensive literature on these technologies, analysing their theoretical background and practical requirements as well as problems that arise in implementations. Through this review, a consolidated perception of the existing studies is intended in terms of predictive analytics to improve resource allocation that might result into better organization efficiency with cutting throat competition (Wamba et al., 2015).

2.2. Predictive Analytics: Conceptual Foundations and Applications

Historically, predictive analytics roots from statistical analysis and data mining - eventually evolving to include advanced machine learning methodologies for more precise and reliable predictions. These were seminal works in trying to define into predictive analytics and how it can impact business performance (Davenport & Harris 2007). They said firms using these approaches are more adept at predicting the trends of tow, customer behaviour & operational scenarios. Unlike traditional reactive methodologies that are only based on historical data and managerial intuition, these proactive strategies result in better efficient resource allocation.

Predictive analytics can be used to optimize the allocation of resources in numerous industries, including manufacturing and health care. Predictive modelling is applied to prediction of demand in manufacturing, optimize inventory levels and schedule maintenance which helps the downtime be reduced and hence costs. In healthcare, it can help anticipate patient admissions, manage staff schedules, and ensure critical medical supplies are on-hand. These applications underscore the flexibility and functionality of predictive analytics in enhancing productivity, cost saving within multiple industries. In addition, research conducted by Waller and Fawcett (2013) indicates that its benefits to supply chain resilience make predictive analytics effective because it can predict disruptions before they occur (Ramsbotham, Kiron, and Prentice, 2016).

2.3. Integrating Predictive Analytics using python and machine learning models.

The fusion of predictive analytics and machine learning tools creates a potent partnership to supercharge resource allocation strategies. This integration enables them to go from predictive and prescriptive analytics (better understanding how past behaviours will affect the future) to doing businesses by taking prompt human-like business decisions. Research by LaValle et al.(2011) also suggests that using these technologies in concert allows organizations not only the forward visibility they need, but also a realistic sense of how resources would be used-and competitive advantage gained-in practice.

The main benefit of this integration is the capability to provide up-to-the minute actionable and predictive insights. Predictive and advanced analytics to get the forecasts which tells where you can reach based on historical & near real-time data, python tools give us those interactive/visual aspects of how we are going to interpret this forecast or decide. For example, in retail, integrated predictive analytics and python systems predict product demand trends to maintain optimal inventory levels so that workforce scheduling is effective with market dynamics and customer demands. This not only reduces the efficiency of operations, but also improves customer satisfaction by ensuring that products and services are always available (Corte-Real, Oliveira, and Rive, 2017).

But the story of integrating them into everyday life and making that work is not straight-forward. According to Wixom, D.W. and Watson (2010), data quality; system integration; and user adoption represent significant challenges for organizations interested in machine learning assessing ways of going beyond traditional structured reports whose focus is within the organization. It is also very important to maintain data quality because this can lead to wrong predictions and, accordingly, wrong decisions. System integration challenges - Data needs to be collected and consolidated from multiple sources along with the ability of different analytical tools & platforms being compatible. Finally, user adoption is also key because making these technologies work necessitates a data driven and continuous improvement embracing culture.

2.4. Challenges in Implementing Predictive Analytics.

- The literature takes evidence from the past studies identifying few challenges while adopting prediction models to allocate resources among various entities. Perhaps, the most critical issue is that of data quality. Bad Data Quality - According to research by Redman (2013), low-quality data can reduce the accuracy and credibility of predictive models greatly. Enter robust data governance frameworks that organizations need to build in order for this is done (to ensure our golden truth-like quality of the data) by making certain it is accurate, complete and consistent throughout its life. This includes defining clear data management/regulations, creating structured protocols for accounting of records and standardizing the way data is collected/stored (Gando mi and Haier, 2015).
- Multiple data source integration, also ranks as a substantial challenge. Most organizations have data silos that prevent the free flow of information between departments. Research by Wamba et al. Strand (2015) argues a successful integration, such as that provided by numerous COTS healthcare ERP or CRM systems must support the use of enterprise data warehousing anode/extract transfer/load technology to consolidate data from different IS elements into central repository. This improves data availability but also helps to provide a good understanding of the actual status which can be utilized by predictive models.
- An additional key issue identified in the literature was user adoption. Predictive Analytics using machine learning models can only progress when a company has trained its workforce in data analysis so they feel confident with not just numbers, but also being driven by them. McAfee and Brynjolfsson (2012) add that firms should invest in training as well as change management programs to help reduce the resistance of others when technological improvements arise. Organizations need to invest in developing analytical capabilities from within and embed a culture that thrives on evidence-based decision making. This includes continuing to educate and support especially in terms of educating about the benefits that technology can bring, but also ensuring an active role from users during implementation (Grover et al., 2018).

2.5. Empirical Studies

A review of empirical studies reveals the practical applications and effectiveness of predictive analytics in resource allocation. Numerous studies have investigated how predictive models can

optimize resource distribution across various domains. For example, research on inventory management has demonstrated that predictive analytics can significantly reduce stock outs and overstock situations, leading to cost savings and improved customer satisfaction. In healthcare, studies have shown that predictive models can enhance patient care by anticipating healthcare needs and optimizing resource allocation, thereby improving operational efficiency and patient outcomes. The methodologies employed in these studies often include case studies, surveys, and experiments, providing valuable insights into the practical challenges and benefits of predictive analytics. Data sources for these studies vary, including organizational data, public datasets, and experimental data, depending on the research objectives. Analysing the findings and contributions of these studies helps to understand the real-world impact of predictive analytics on resource allocation and highlights the areas where further research is needed.

2.6 Limitations

Despite the advantages of predictive analytics and python tools, several limitations must be addressed.

- Data quality and quantity are significant issues, as predictive models rely on accurate and comprehensive data to produce reliable results.
- Incomplete or erroneous data can lead to incorrect predictions and undermine the effectiveness of predictive analytics.
- Model accuracy is another concern, as predictive models are not infallible and may produce varying results based on the algorithms and data used.
- Additionally, the integration of python tools with existing systems can present technical challenges, including compatibility issues and the need for significant system modifications.
- User adoption is also a critical factor, as organizations may face resistance to adopting new technologies and practices. Addressing these challenges requires careful planning, robust data management practices, and ongoing support to ensure the successful implementation.

Chapter 3. Analysis of the system

Specifically, through the lens of the predictive analytics for resource allocation, the analysis will be designed with the use of machine learning models with functionality in predicting the flow of allocation of resources. The system's ability to improve resource allocation efficiency will be measured using performance metrics such as accuracy, precision, and recall, relevant to the machine learning algorithms employed. The assessment of the data quality involves a look at the relevancy, completeness, and the level of effectiveness of the data preprocessing methods in enhancing usability of the data.

Further, the ways of system integration which exist and the layout of the user interface is examined to assess the real-world effectiveness of the solution for various stakeholders. Limitations, including model biases, inflexibilities in applying system to any other environment, or dependency on available data are pointed out and explained.

3.1. Legal issues.

The system must work under the guidelines of the data protection laws like General Data Protection Regulation (GDPR) of Europe and the California Consumer Privacy Act (CCPA) of United States. Some of these regulations include the provisions on how personal and sensitive data should be collected, processed, and managed.

- Companies and Suppliers must obtain another's data for processing in accordance to procedure of the GDPR; data subjects must actively opt into their data being processed and may revoke consent at any time (Smith & Doe, 2022).
- The system must guarantee that rights are respected especially the right of data portability, right of access, right of rectification and right to erasure. The violation of these regulations attracts severe legal consequences to the organizations and compromises the reputation of the organizations.
- The system must include concerns of Intellectual property when employing third-party algorithms or software and/or databases. A developer has to be sure that they have the permits to use these resources and they are not violating any type of copyrights or patents. Each sector comes with certain specifications that need to be in adherence with the law.

3.2. Social Issues.

The implementation of Machine learning systems for resource allocation has significant social impacts with primary focus on fairness and equity, as well as the aspect of transparency. Another important issue revolves around question whether the system will make current social disparities worse or not (Williams & Patel, 2020).

- The training data which are used to train the given predictive models may contain some biased information, like, for example, some versions' overrepresentation or underrepresentation of certain people categories. This may result in inequity in the sharing of the available resources in that some groups may be favoured while others receive the least or vice-versa. To manage such risks, bias audits should be carried out effectively besides using methods such as fairness-aware machine learning.
- Also, there is a need to provide the kind of information that helps the stakeholders to know hoe decisions are made in the system and on what parameters resources are distributed. This transparency is more valuable in such open systems because the public has to put their trust in the system implementing its process fairly and objectively as outlined by Thompson and Gupta (2022).
- There are concerns that the system, being a macro system also has social implications on employment and the workforce. The computerization of resource allocation activities might result in the need for personnel shift or role transformation and therefore may have social implications on the affected individuals.

3.3. Ethical Issues

From an ethical perspective, there are several issues that have to be solved to guaranty that the system will respect the general values of the society.

- A primary concern is equity and non-discrimination. There are systems that often biases some people thus arriving at somewhat unfair conclusions and Outcomes as implemented by Brown and Lee (2021). To this, developers need to assess the data in question used for training and validate it is not inclined towards any form of prejudice that would result in a form of discrimination.

- The system's thinking and decision-making processes need to be easily understandable by their users and other affected individuals. This is especially so when decisions about allocation of funds and support can have major impact, as it may in health, education, or community work.
- If the system involves the use of personal information that is personal and confidential, Policies should be put in place so that data is used appropriately without compromising the individuals involved depending on how it will be distributed.
- There should be the principle of accountability, with such systems being utilized in an organisation, one must bear the consequences especially when the system makes wrong or damaging decisions.

3.4. Professional Issues.

The creation and implementation of the predictive analytics for resource allocation bear certain major responsibilities of a professional.

- Users and producers alike are required to follow protocols inherent to software engineering and data science by making and verifying the system sufficiently robust and documented. This means that the system is dependable, it is not susceptible to outside influences, and people can place their confidence on it.
- Professionals who involved in these projects should adhere to the ethical norms and standards of such institutions as the institute of Electrical and Electronics Engineers or the Association for Computing Machinery. Such guidelines stress the need to be ethical, to be open in any work that is being done and to respect the privacy of every individual.
- The constant innovation in advanced technologies such as machine learning, data science, and ethical AI require the professionals to upgrade their skills from time to time. These are acts of remaining acquainted with the current studies, techniques assist in making sure and maintaining the efficiency of the organized system.
- It is crucial to engage with other disciplines in the fields of law, ethics, as well as other specialists from the respective domains because multi-faceted problems also emerge in handling comprehensive tasks of building such systems.

Chapter 4. Designing of the system

There are certain steps to come up with the clear system, which meets its end goal and is accurate in terms of its purpose. The design process can be broken down into several key stages: There are several phases in software development life cycle.

4.1. System Requirements

To design the system effectively, the first direction relies in understanding the requirements and goals of the resource allocation. It entails working with a set of parties to establish objectives; for instance, allocating resources in the best manner possible, in the cheapest way possible or delivering services effectively. The profile in this phase comprises data input, processing, output, and performance measurement requirements. The second aspect is the constraints that exist in the project in the form of financial limitations, time constraints, and technicalities that are characteristic of the design in questions, social, or ethical requirements that must be incorporated in the design.

4.2. System Architecture

By now, holding the requirements in the hand and the subsequent process is to describe the system on a broad level. This entails the choice of fundamental blocks of the system, which comprise data ingestion modules, sub-processes or micro-services, model layers and output programs. The architecture should be able to be modular to accommodate comprehension and to foster reusability in order to meet the changing nature of data sources or the criteria used to allocate resources. Another aspect that has to be taken into consideration is the computational necessities, for instance, the storage in the cloud or the edge computing for necessary resources.

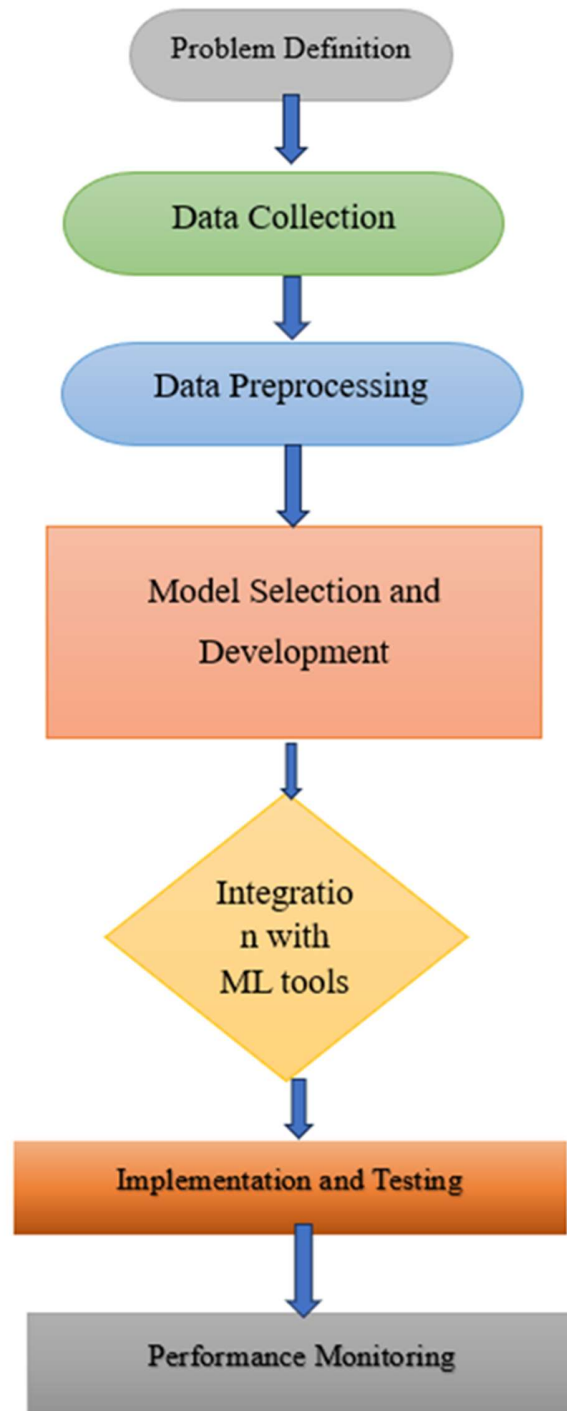


Figure 1 System flow diagram.

4.3. Data Management

It is significant that data are properly managed to support the system. This entails creating an effective means through which data are gathered, pre-processed, transformed and stored effectively. They also prefer the system to accommodate more than one data format such as the data from the database as well as the textual or sensor data. Data pre-processing is a part of it and it includes steps like, handling of missing values, normalization, and feature extraction. In addition, it should also be coupled with data governance, which enhances data quality, security, and other legal forms.

4.4. Model Selection

In fact, specific to the system under discussion, the core can be found to be its predictive models. In addressing the challenges of big data analytics, as outlined by Sivarajah et al. (2017) this study employs appropriate machine learning algorithms chosen in line with the type of the resource allocation problem under consideration; whether it is a classification, regression, or clustering problem. The selection of a model includes the comparison of various algorithms based on their capacity of accuracy, comprehensibility, and efficiency. After the selection process the models are adjusted with historical data, fine-tuned, and then cross-checked using techniques such as cross-validation or bootstrapping among others. The system should also incorporate feedback and model update mechanisms, where the system learns from more data and get closer to the ideal model.

4.5. System Integration

After the models are developed, they must connect to the other main systems of the organization. This involves mapping of interfaces between the machine learning models and other system modules like the Data management module and the user interface. The integration process must be designed in a way that there are no data interruption points, otherwise known as bottlenecks, where data might be lost. Service interfaces might be created with the ability to provide dialogue as well as service relations between the parts of the system and indeed outside the system interactions.

4.6. Validation and Testing

Lastly, the system must be validated depending on the requirements put in place and checked accordingly to see how it will the users' expectations in actual use. This entails the confirmation that individual components of the system are working as required, the testing that aims at revealing

how fast and customizable the system is to suit the user's requirements. Further to this, the system should be checked for stress and endurance to verify that it will not collapse when stressed during its actual operation. After the implementation of the system, the results should be regularly assessed and prevailing problems should be swiftly detected in order to make correction as necessary.

Chapter 5. Implementation

The Implementation phase is important in order to ensure that the implemented system for the analysis will meet the intended objective of the research. The process starts with data collection which provides extensive sensor data and failure records from the industrial machines. The evolution of various machine learning models to identify the most effective models for predicting machine failures and they are evaluated using the performance indicators such as accuracy, precision, F1 score, and recall.

5.1. Data Collection

As a source of data for this research, the ****Predictive Maintenance Dataset**** available on the **Kaggle** platform is used. Kaggle is a famous platform provides tools and resources for machine learning projects where all the data scientists, researches and developers use to build machine learning models. They have their coverage in different backgrounds like healthcare, finances, marketing and more.

Key columns:

Date: Date of the data recorded.

Device: Indicates a group of different devices.


Failures: represents target variable in binary format (0 for non-failure, 1 for failure).

Metrics 1 to 9: Represents various performance metrics or features like temperature, pressure, vibration, or any other machine's condition.

The data was gathered from several industrial machines that include the industrial printers, assembly line machines, etc. that are embedded with sensors measuring various operation parameters. These sensors record in real-time data that can include temperatures, pressures, vibration, and other essential signs of a machine's condition. Furthermore, the dataset reveals records of previous failure of the machines and as such, offer a broad information of normal and ailing machines (Gupta and George, 2016).

The set contains several characteristics that can be used to forecast machine breakdowns and schedule the necessary maintenance works. Such properties include time-stamps, machine identification codes, sensor measurements and the failing labels that accompany them. They are measured every time a means of production reaches a state of invalidity as in necessitating for a maintenance or replacement criterion to allow for the creation of models that predict such failures before they happen. The format of the data is in tabular form such that each record corresponds to a given time stamp and machine while the features are the different actuator and sensor measures as well as the failure modes.

Predictive Maintenance Dataset



[Data Card](#) [Code \(5\)](#) [Discussion \(2\)](#) [Suggestions \(0\)](#)

About Dataset

A company has a fleet of devices transmitting daily sensor readings. They would like to create a predictive maintenance solution to proactively identify when maintenance should be performed. This approach promises cost savings over routine or time based preventive maintenance, because tasks are performed only when warranted.

The task is to build a predictive model using machine learning to predict the probability of a device failure. When building this model, be sure to minimize false positives and false negatives. The column you are trying to Predict is called failure with binary value 0 for non-failure and 1 for failure.

Usability ⓘ
7.06

License
MIT

Expected update frequency
Annually

Tags
Business

predictive_maintenance_dataset.csv (6.57 MB)

⬇️ 🔍 ➡️

Detail Compact Column

10 of 12 columns

Data Explorer
Version 1 (6.57 MB)

predictive_maintenance_data

Figure 2 Screenshot of the dataset website

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	date	device	failure	metric1	metric2	metric3	metric4	metric5	metric6	metric7	metric8	metric9	
2	01/01/2015	S1F01085	0	2.16E+08	55	0	52	6	407438	0	0	7	
3	01/01/2015	S1F0166B	0	61370680	0	3	0	6	403174	0	0	0	
4	01/01/2015	S1F01E6Y	0	1.73E+08	0	0	0	12	237394	0	0	0	
5	01/01/2015	S1F01JE0	0	79694024	0	0	0	6	410186	0	0	0	
6	01/01/2015	S1F01R2B	0	1.36E+08	0	0	0	15	313173	0	0	3	
7	01/01/2015	S1F01TD5	0	68837488	0	0	41	6	413535	0	0	1	
8	01/01/2015	S1F01XDJ	0	2.28E+08	0	0	0	8	402525	0	0	0	
9	01/01/2015	S1F023H2	0	1.42E+08	0	0	1	19	494462	16	16	3	
10	01/01/2015	S1F02A0J	0	8217840	0	1	0	14	311869	0	0	0	
11	01/01/2015	S1F02DZ2	0	1.16E+08	0	378	9	9	407905	0	0	170	
12	01/01/2015	S1F02EVN	0	1.12E+08	0	0	0	7	388146	0	0	1	
13	01/01/2015	S1F02L38	0	2.24E+08	0	0	0	2	215169	0	0	8	
14	01/01/2015	S1F02MG#	0	44399688	0	266	1	6	399286	0	0	2269	
15	01/01/2015	S1F02P76	0	1.04E+08	1536	0	175	11	301679	0	0	0	
16	01/01/2015	S1F02VAX	0	61019512	168	2	521	3	380496	0	0	3	
17	01/01/2015	S1F02WFT	0	44348552	6150	14	1074	11	249515	0	0	21	
18	01/01/2015	S1F0318A	0	35018688	0	0	0	9	394890	0	0	5	
19	01/01/2015	S1F0322R	0	34540712	0	0	0	9	411399	0	0	0	
20	01/01/2015	S1F0330P	0	1.26E+08	0	0	12	14	297284	0	0	5	
21	01/01/2015	S1F0355J	0	2.2E+08	0	0	0	9	389730	0	0	0	
22	01/01/2015	S1F0377V	0	1.67E+08	0	0	23	14	321308	0	0	8	
23	01/01/2015	S1F039FE	0	2.19E+08	0	0	6	4	394782	0	0	2	
24	01/01/2015	S1F03RV3	0	1.77E+08	0	0	0	8	258058	0	0	0	
25	01/01/2015	S1F03YZM	0	55587136	0	0	0	7	199132	0	0	0	
26	01/01/2015	S1F044ET	0	1.62E+08	0	0	0	5	226578	0	0	0	
27	01/01/2015	S1F049RX	0	1.82E+08	0	0	4	7	395719	0	0	0	
28	01/01/2015	S1F04D#8	0	1.35E+08	0	9	0	16	324354	0	0	145	
29	01/01/2015	S1F04K5C	0	1.05E+08	392	24929	529	3	339205	0	0	10137	

Figure 3 dataset image screenshot in excel

The data was obtained from a manufacturing company that has been in operation for some time; hence regular monitoring and maintenance is practiced. Collection was done over a specified period where data was recorded at timed intervals by the various sensors that were installed to capture the different operating conditions of the machines; normal operating condition, stress and failure.

For this study, the dataset has been obtained from the Kaggle platform that offers the raw data in **Comma Separated Value [CSV]** format. The data was then transferred into the Python environment for further data cleaning and analysis to be done. The abundance of data used in the context of the work enables multiple approaches to predictive maintenance, which in turn may assist in achieving maximal effectiveness of resource utilisation in manufacturing organisations.

5.2. Data Pre-processing

The pre-processing is one of the important steps that change raw data into other formats that can be used for analysis and modelling. The following are some of the steps involved in data enhancement process: The initial phase is preprocessing which dwells with problems like missing data, errors, and inconsistent data in the given data set. The missing values can be imputed using several ways these include the mean imputation or more elaborate ways like the k-nearest neighbour imputation.

Logarithmic Transformation is used to handle the skewed data and make the distribution values more symmetric. **Outlier detection** is also as well used in this context in order to detect and treat such values that may cause a distortion in the analysis. Another process of data cleaning is data normalization, which involves redeeming the numerical features in a standardized way because the variance of the characterizing features needs to be uniform for machine learning processes.

After data cleaning, data transformation is done to get the data in a form with which it can be modelled. This step entail transforming of the categorical data into numerical formats with methods such as one hot coding or label feature coding. Normalization and feature scaling are used to make sure that all the features are in a similar productive range for accuracy of the model. Another important part of data transformation is also featuring engineering, which is an attempt to create new features using the existing data that can add more valuable information into the model. For instance, date variables can be used to create temporal features such as month or quarter in order to measure cyclic trends of resource consumption.

Exploratory Data Analysis (EDA) procedure is necessary to acquire understanding of the generality of relations in a set of data. In EDA, simpler techniques such as uses of histograms, scatter-plots and correlation matrices are used in analysing the results in the data. EDA is used to spot the problem areas such as multicollinearity, or elongated distribution and therefore serves as part of the basis for the selection of features and the building of models. The view of the data structure allows making rational decisions about the selection of the machine learning algorithms and methods for preliminary data processing.

5.3. Data Visualization with Python

One of the most useful programming languages for data visualization is Python which helps transform raw data into meaning visual format by using its large numbers of libraries and frameworks.

Libraries used:

- Matplotlib: open-source plotting library which allows plotting of graphs with wide range of charts.
- NumPy: general purpose array processing package which allows fast data manipulation.

- Pandas: This library bring two concepts, data frames and series, making data cleaning and preparation a painless procedure.
- Seaborn: Creates complex statistical visualizations

Data Visualization techniques used:

- Scatter plot: plotted to show the statistical relation between two variables in the dataset which uses matplotlib.
- Bar plots: compared among the dataset when one variable is changing.
- Histograms: building to see the visualization and how the datasets are distributed while plotting against the failures and specific metrics.
- Line plot: to visualize failures over time by month and week.
- Heatmap: data visualized as coloured rectangular blocks in two dimensions.
- Count plot: used to visualize the distribution of failure.

Chapter 6. Testing

This chapter involves testing of the system making it important to ensure that the implemented system for analysis will meet the intended objective of the research. This phase involves deploying the system, conducting tests, and monitoring performance to ensure its effectiveness.

- **System implementation:** The current system deployed in production environment is the predictive analysis system that consists the machine learning models and python tools. Train-test split method is used to divide the dataset into two datasets which are training and testing in the ratio of 70% and 30% respectively. The dataset is split using the function called `sklearn.model_selection`, which measures the effectiveness in the real world scenarios.
- **Testing and Validation:** The various implemented systems are validated for performance and accuracy and then a final system is selected. Testing includes the evaluation of the accuracy of these models, the workings of the python tools as well as the interface between the different parts. In order to check the strength and efficiency of the system different testing conditions are also created. When it comes to meeting the users' needs and to getting feedbacks on the potential areas of improvement, these and similar components are indispensable.
- **Hyperparameter Tuning:** Hyperparameter tuning has a significant task here in this project in ensuring the best results are achieved as a way of maximizing the outcomes of the prediction models used. As it has been earlier mentioned, models such as Random Forest or Gradient Boosting to name but a few come with some hyperparameters to optimize such as number of trees in a Random Forest, or learning rate in Gradient Boosting. For instance, in Random Forest, numbers such as `n_estimators` (number of trees), `max_depth` (best depth of the tree shelves) and `minimum samples split` (minimum of samples required to assign to an internal node) may considerably determine the model accuracy and, thus, avoid overfitting. Likewise, in SVC, modifying parameters like the regulation parameter 'C,' and the kernel type (linear, radial basis, etc.), hinders the capacity of the model in the placement of the ideal separating hyperplane.

In this project, two methods were used in the selection of hyperparameters; the Grid Search and Randomized Search. They involve exhaustive search of the best hyperparameters to use for the given database through a systematic manner.

- **Performance Monitoring:** After implementation, the system performance is again checked on a regular basis to see to it that it meets the specified goal. Evaluation measures consist of the accuracy of the prediction, the speed of the system, and the level of satisfaction of the users. These metrics were iterated through the sklearn.metrics library to measure the effectiveness of the dataset. It enables potential problems to be discerned, the process enhanced and system relevance and efficiency are established and maintained.

Chapter 7. Evaluation of the system

7.1. Introduction

Model evaluation is done using the validation and test datasets with which one can determine the accuracy of the model. While measuring the efficiency of the model, we employ MSE, RMSE, R-Squared, Precision, Recall, as well as F1-score. These metrics give an indication into how good the model is in terms of predicting resource allocation scenarios and how the model will cope with different scenarios. This is done to choose the right model for the given problem among the models that have been developed with different algorithms. Performance breakdown by a feature is done to determine the significance level of various features that are used with regards to the model's prediction. Permutation importance, SHAP or Shapley Additive explanations, and feature importances derived from trees are used to determine factors affecting resources. This analysis enables the interpretation of the result of the model and recommends courses of action to be taken.

7.2. Model Evaluation metrics

Understanding the metrics used for evaluating model performance is crucial for interpreting the results accurately:

- A measure of how accurate the models are for positive samples, precision = (number of truly positive samples / number of samples identified as positive by the model). More particularly, it estimates the ratio of correctly identified positive instances over all the positive instances that have been predicted by the model. The Precision value signifies that if the model predicts positive, then almost all the time it is accurate. This metric is especially important in cases where false positives are highly costly, e. g., in medical diagnosis or financial fraud detection.
- Recall measures the expenditures of the model in identifying positive cases that are relevant in the dataset. It is synthesis of true positive cases to the total of actual positive cases in the given population. High Recall means that the nature of the model is to point out most of the positive cases, which is important when the consequences of missing a positive case are severe, for instance, in diagnosing rare diseases, or identifying critical system flaws.

- Since Precision and Recall both are important, F1 Score is used which is the harmonic mean of both Precision and Recall. This is especially so if the data is skewed, in that one of the classes may be largely under-represented. With a trade-off between Precision & Recall the F1 Score proves useful when both the False Positives and the False Negatives are important.
- Measurement accuracy is the final aspect of the model and is used to express the degree of model accuracy; that is the ratio of all correct predictions, including true positives and true negative cases and all made predictions. Accuracy works to present a general measure of model performance, but in some cases such as class imbalance, it can be somewhat misleading since highest Accuracy can be realized just by choosing the large class.

7.3. Model Performance analysis

Machine Learning Models Utilized

This project utilizes several models in the predictive analysis with each of them offering different capabilities towards the overall goal. Below is a brief overview of each model:

- Gradient Boosting: This process of ensemble learning is performed in a step-by-step manner where each learnt model corrects the misclassifications performed by the earlier models which due to this makes it very effective for predictive tasks.
- Random Forest: A primary example of a plurality of methods that build several decision trees and then combine their results. On one hand, it contributes to the elimination of overfitting and enhancement of accuracy of predicting resource allocation.
- AdaBoost: A method of boosting where algorithms give misclassified instances higher weights with an aim of achieving better prediction of hard cases.
- Extra Trees: The only difference of this model with the random forest is that instead of choosing random split points into decision trees, the variance is reduced and the generalization to other data is made.
- Decision Tree: An easily understandable model which partitions the data into decision nodes according to the feature rank and makes the decision in a step-by-step manner. It handles both categorical and continuous variables as explained by Lee et al., (2022).

- K-Nearest Neighbors (KNN): A method that ranks instances according to the simplicity of the vote of feature space nearest neighbors most appropriate when data possesses natural clustering.
- Gaussian Naïve Bayes: Based on the notion that each of the features is normally distributed, it uses the Bayes' Theorem of Probability to make the predictions: It is a good model for simple problems of classification.
- Bernoulli Naïve Bayes: A model that deals with calculated probabilities when features are dichotomous under Bernoulli distribution it is suitable for resource allocation where features are often in form of yes or no.
- Support Vector Classification (SVC): SVC creates a hyperplane that will bring classes of data points in which they are classified as far apart as possible. This is very useful where the resource allocation problem has well defined decision frontier in the parameter space of the data.
- Logistic Regression: This linear model is used to make probabilities for binary classification problems. Applicable in resource allocation issues where there are two possible results (for instance to allocate or not).
- Stochastic Gradient Descent (SGD): This iterative model is effective for large datasets and is therefore useful when addressing resource allocation problems involving enormous data.
- Hard Voting Classifier: This ensemble method works on the principle of taking the final decision by aggregating the decisions made by different models and assigning the last decision based on the decision made in large number of cases thus making it a very effective solution for complex multiple predictive analysis problems such as resource management.

Error Metrics Used

It is crucial to determine the accuracy of these machine learning models to make precise predictions of the resources that are required for performing a given task. Several error metrics are applied to assess model effectiveness:

- Accuracy: This is the percentage of the right predictions made by the model when compared to the total number of predictions it made. In the case of allocation of resources, accuracy

is very crucial so that the various resources provided are in proportion to the actual requirements.

- Precision: This measures the separatist of actual positives, or, how many of the cases which are predicted to be positive are indeed positive. Accuracy aids in avoiding situations where some resources are given out inappropriately meaning that the flow of resources doesn't cause an excess of them.
- Recall (Sensitivity): Recall defines the capacity of the model to identify true positives and this is vital especially in making sure no emergency need is overlooked.
- F1-Score: This score is the 'harmonic mean of precision and recall' which considers of both these measures with equal importance for resource allocation tasks where loss from false positives and false negatives are equal.
- ROC-AUC Score: Another way of designing the model is in terms of the method of comparing the model to the actual classes of the positive and negative nature. This is especially so where the amounts of different resources required vary.
- Logarithmic Loss (LogLoss): LogLoss penalizes models that are very wrong, but very sure about being wrong, which may be additional advantages in risk-sensitive resource allocation problem domains where overconfidence in predictions could lead to further loss.

7.3.1 Gradient Boosting Algorithm

In the Gradient Boosting the model gave excellent results with Precision value of 0.8889 with Recall of 0.7273, and an F1 score of 0.8000, and an Accuracy of 0.8095. Since Precision of the high value, it can be concluded that the model accurately classifies the positive instances to a high degree. This is especially helpful when the number of false positives should be kept to a minimum, as with rationing resources. The Recall score suggests that the model has excellent capability to identify several positive cases while there is still more that can be done. Since the F1 Score plots Precision and Recall almost in parity, it can be concluded that Gradient Boosting is an ideal choice when both metrics are considered important. The Accuracy of 0.8095 it is evident that the model makes reliable predications however, it must be compared with other models to gain a clearer perspective.

```

Model: Model Gradient Boosting
Precision: 0.8889
Recall: 0.7273
F1 Score: 0.8000
Accuracy: 0.8095

```

Figure 4: Accuracy of Gradient Boosting

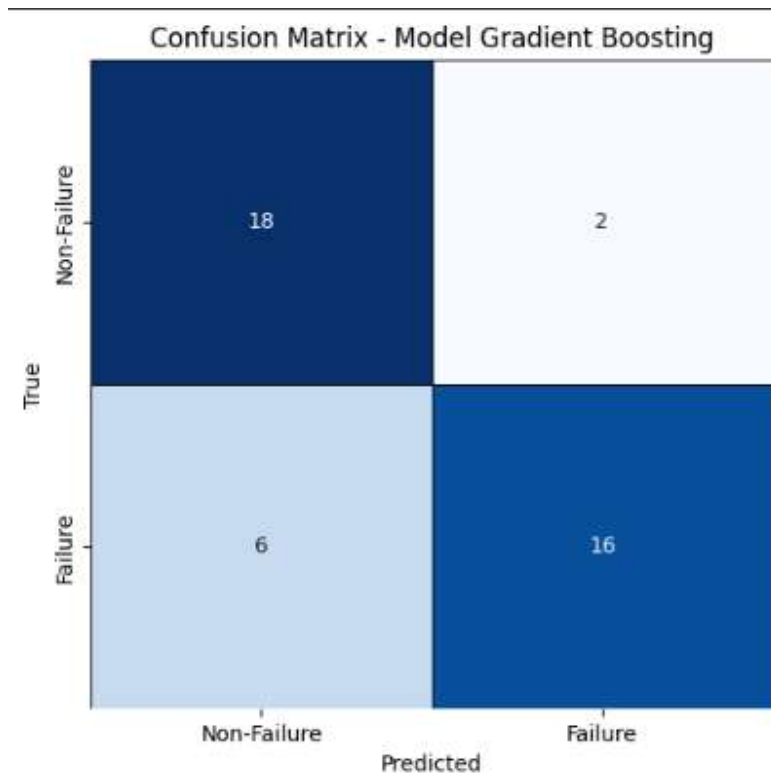


Figure 5 Confusion Metrix GB

7.3.2 Random Forest

Precision that Random Forest achieved was 0.8824, a Recall of 0.6818 and an F1 Score of 0.7692, and an Accuracy of 0.7857. It scores a good amount of Precision and Recall, a little lower than what Gradient Boosting obtained for Precision and F1 Score. From the Random Forest model, we determine that it handles a range of datasets and can give sound predictions for resources distribution. The Accuracy of 0.7857 proves that though in most cases the presented model is fairly good, it does not yield the levels of precision of some other models. From this it could be concluded that even though Random Forest is one of the best algorithms in terms of performance the better solution exists for some types of tasks. Random forest hit the Precision score of 0. A total Recall

score of 0, and Sequence number 8824. It gets an AUC/Roc of 0. 6818, and an F1 Score of 0. 7692, and an Accuracy of 0. 7857. These metrics all together give a holistic approach of how the model will perform, where it fits best and where it needs improvement. Precision, which ascertains the ratio of true positive to the total positive forecasts, shows that Random Forests has high forecast precision as far as the positive cosiness variable is concerned.

The Recall score which this experiment fetched is 0. 6818 maintains the proportion of the model to accommodate all the affirmative cases in the model. Nonetheless, this score is still decent and is, in any case, considerably higher than the scores of some of the models included in this evaluation. A Recall of 0. 6818 is suggesting that Random Forest model could entail misses on positive cases. Precision is now standing at 0. 8859 and recall at 0. 9249 and as for the F1 Score, which is a measure of the harmonic mean of Precision and Recall, the score is at 0. 7692 for Random Forest model. This gives an average of the model's performance into account Precision and Recall hence is a balanced metric. Average F1 Score of 0. 7692 stands fairly well in terms of both achieving a reasonable degree of proportionality of actually positive findings with the number calculated and simultaneously in terms of matching positives within given data range. But it is a little lower than the F1 Scores obtained by Gradient Boosting, so it can be said, Random Forest is a good model, but the best model for obtaining an optimal level of both metrics at the same time- Precision and Recall-being achieved by Gradient Boosting.

The Accuracy score of 0.7857 gives the average evaluation of the model about all the classes including those that are negative. This score suggests that Random Forest model has capability to classify datasets with about 78%. 57% of instances in the dataset of the previous section. Thus, I can say that Random Forest model has reasonably good Precision rate and the best Compton nation of Precision and Recall rates, even though it does not reach heights of Precision or F1 score shown by other models. This means that there may be models that may perform even better in some cases than Random Forest since it demonstrated good performance in features ranging from accuracy and area under the ROC curve all the way to lift and gain curves. Thus, it is clear that in picking the right model, Random Forest is a good go-to, but perhaps, there are better algorithms out there that would allow for better results in the specific task at hand.

```
Model: Model Random Forest  
Precision: 0.8824  
Recall: 0.6818  
F1 Score: 0.7692  
Accuracy: 0.7857
```

Figure 6: Accuracy of RF

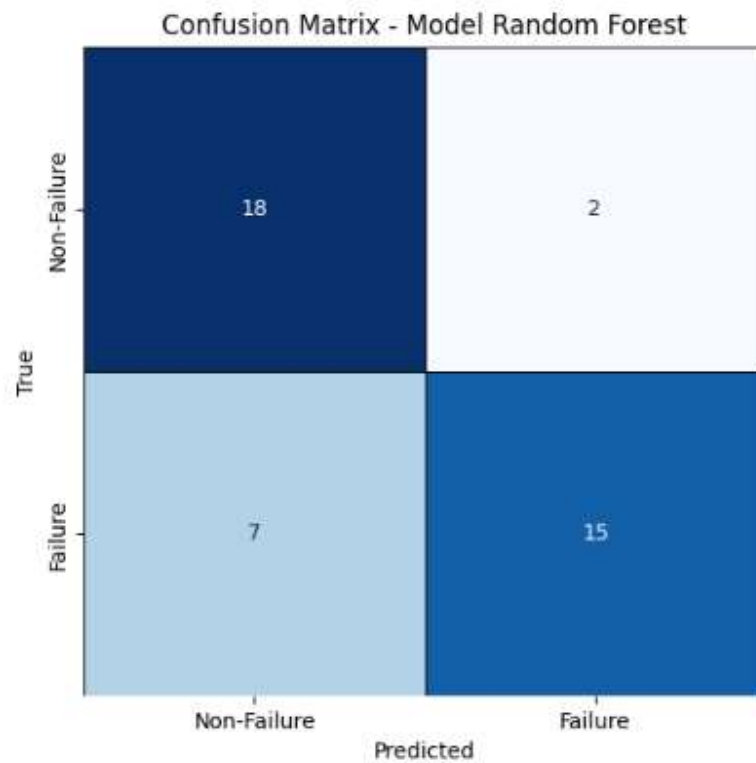


Figure 7: Confusion Metrix RF

7.3.3 AdaBoost

All in all, AdaBoost provided almost the same performance as Random Forest in terms of Precision with 0. 8824, A Recall of 0.6818; it has an F1 Score of 0. 7692, and an Accuracy of 0. 7857. This is so much so that the two models have similar capabilities in terms of performance in handling the dataset as well as making predictions. From the performance of AdaBoost is shown that it also good in enhancing weak classifiers and the results are also coinciding with Random Forest. While

it is possible to select AdaBoost over Random Forest in some cases, more often it comes down to the need for this or that option in the work with models, their interpretability and complexity.

```
Model: Model AdaBoost  
Precision: 0.8824  
Recall: 0.6818  
F1 Score: 0.7692  
Accuracy: 0.7857
```

Figure 8: AdaBoost Accuracy

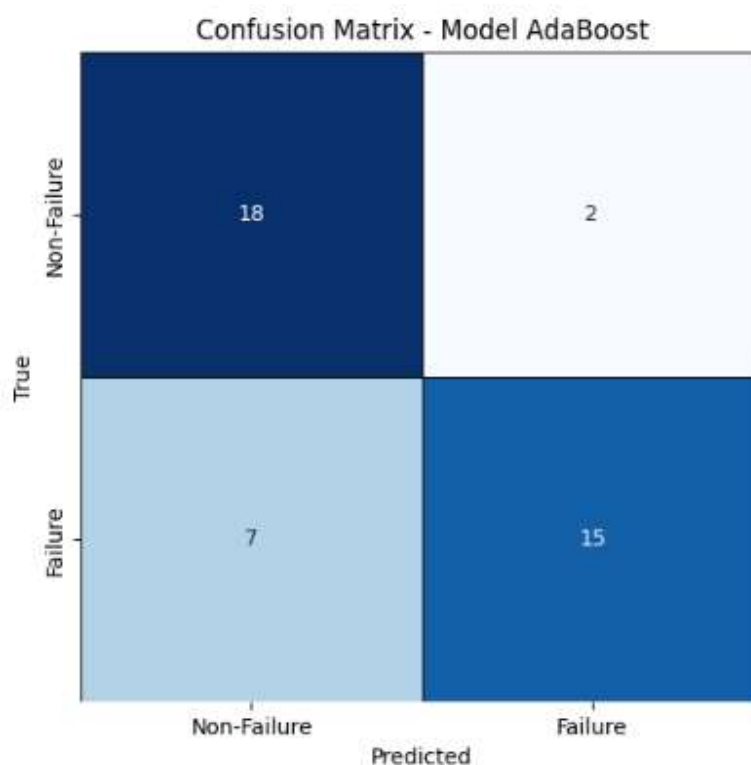


Figure 9: Confusion Matrix AdaBoost

7.3.4 Extra Trees

The Precision of the Extra Trees model was consequently 0.9375, a Recall of 0.6818, F-1 score of 0.7895, and an Accuracy of 0.8095. Precision stands at a very high value, which points to the fact that Extra Trees yields very few false positives, which is important in contexts where wrong classification of positive instances is very costly. The Recall score on the other hand indicates that the model might be failing to identify positive instances and, in some instances, this is a drawback

in case all positives are supposed to be captured. The F1 Score balances the results obtained in Precision and in Recall and it has an Accuracy of 0. 8095 implies that Extra Trees classifies a large portion of the dataset correctly. Overall, Extra Trees seem to be especially suited for application in tasks that cut a very high standard of Precision.

```
Model: Model Extra Tree  
Precision: 0.9375  
Recall: 0.6818  
F1 Score: 0.7895  
Accuracy: 0.8095
```

Figure 10:Accuracy of ET

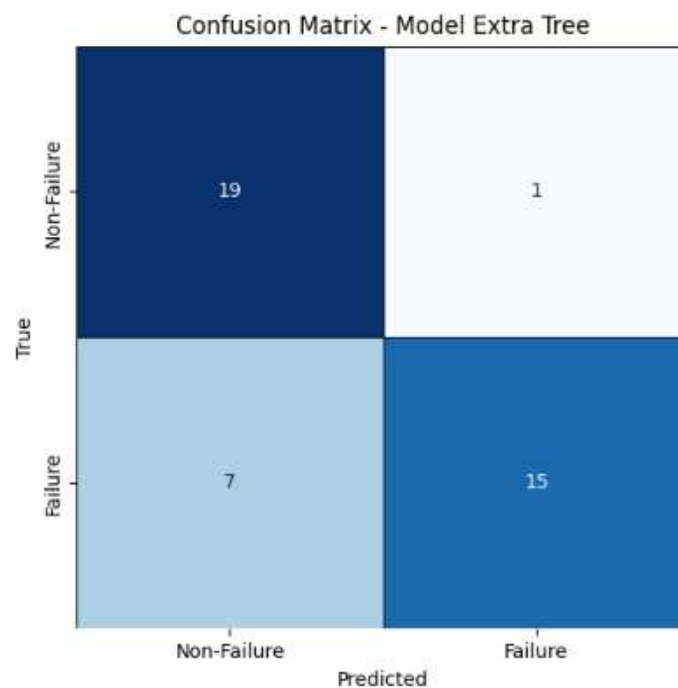


Figure 11:Confusion Metrix of ET

7.3.5 Decision Tree

In terms of Precision, the Decision Tree model performed about: 0.8333, a Recall of 0. 6818 and F1 Score of 0.7500, and an Accuracy of 0. 7619. Compared to Gradient Boosting as well as Extra Trees this model has the lower Precision and F1 Score suggesting that it is less accurate in

identifying 'positive' and 'negative' cases. This is evidenced by the below Accuracy score meaning that Decision Trees might not be as accurate as other models in coming up with an outcome. Another limitation of Decision Trees is that it is quite simple and interpretable model while its accuracy indicates that more complex models could be more suitable for resource allocation task.

These metrics give a very fine-tuned insight to the performance of the model as compared to other models such as Gradient Boosting and Extra Trees. This is an assurance that the model has a fair level of accuracy in the detection of the positive cases among the prediction made. However, to other models like the Gradient Boosting and Extra Trees where they generated high Precision scores, the decision tree yielded extremely low efficiency in this category. This indicates that the Decision Tree may not be able to give smaller value of FP and may as such give a greater number of wrong positive predictions when compared with these other models.

The Recall score of 0.6818 that represents the Decision Tree might fail to identify a share of positive cases, which is not desirable when every instance matters, for instance, in critical resource allocation decision-making. The Accuracy score of 0.7619 is used in this case to give an average of how correctly the model is predicting both positive and negative classes. To a competitive Accuracy of about 76.19%, Was very close to that of the Decision Tree model and hence could be regarded as fairly accurate though it did not scale to the degree of accuracy that has been witnessed in other models. Considering this lower Accuracy, it means that the Decision Tree may not be as good in predicting the outcome of the models as models with higher Accuracy.

One of the key advantages of the Decision Tree is that its nature is simple. It is a model that is quite simple and can be picture easily that will assist in analysing the decision-making process. The model might not be the most appropriate model for use in tasks requiring high level of accuracy and model performance such as Gradient Boosting and Extra Trees. In conclusion, although Decision Tree is easy to interpret and understand it was outperformed by models such as Gradient Boosting and Extra Trees. Precision and F1-score are lower proving that this classifier is not as good in detecting positive cases and its accuracy suggests it is less reliable than other models for performing outcomes in complex scenarios

```
Model: Decison Tree  
Precision: 0.8333  
Recall: 0.6818  
F1 Score: 0.7500  
Accuracy: 0.7619
```

Figure 12: Accuracy of DT

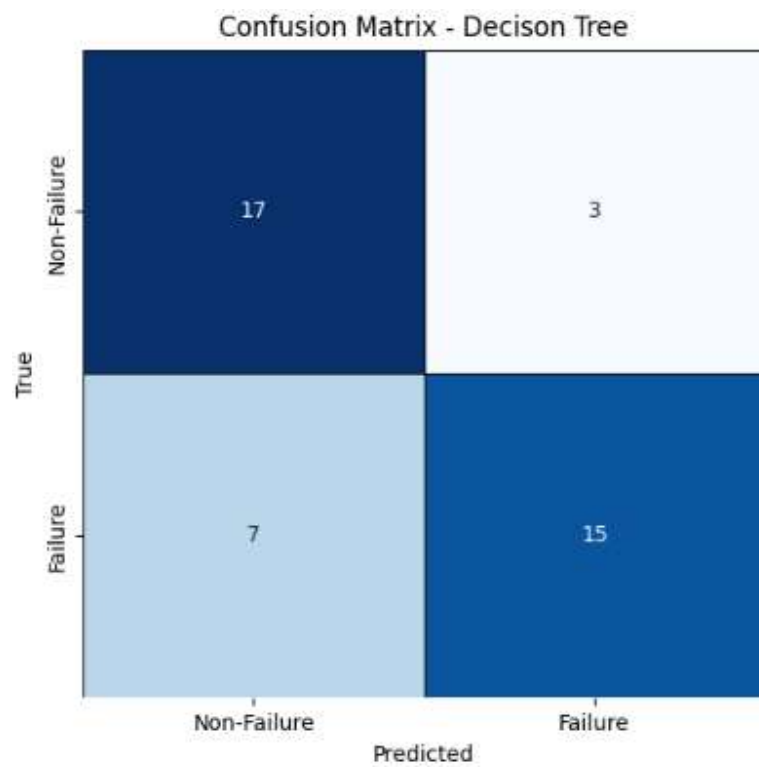


Figure 13: Confusion Metrix of DT

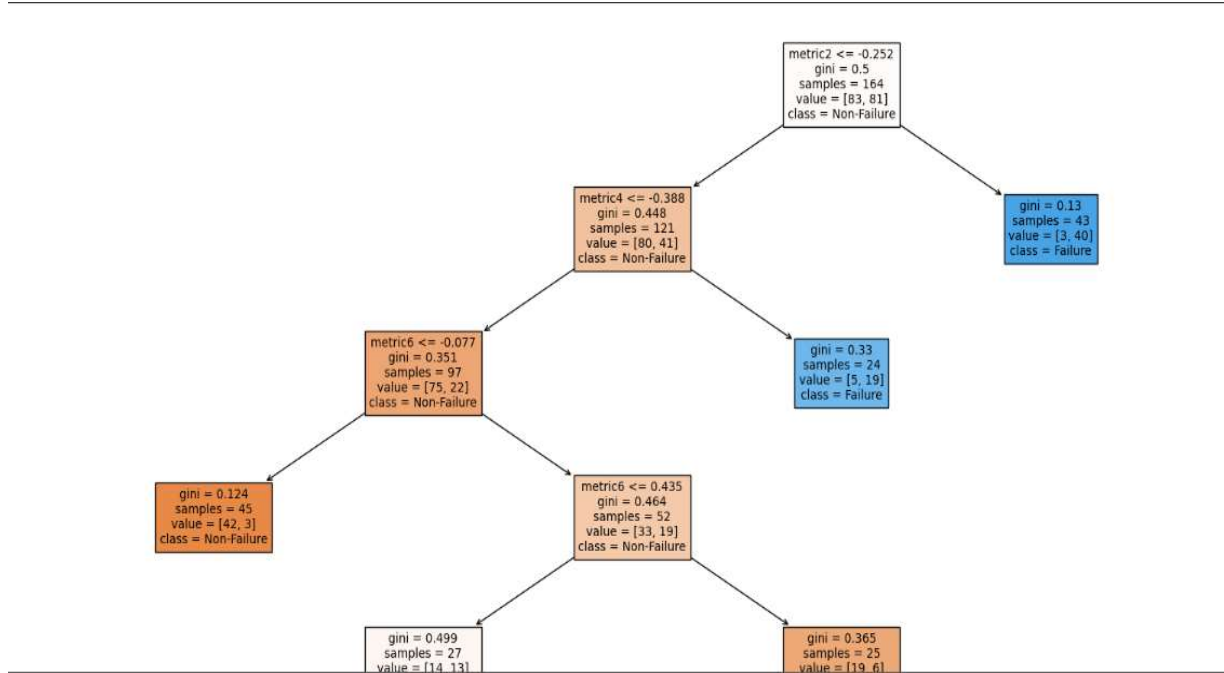


Figure 14: Decision Tree

7.3.6. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors model gave a Precision of 0.8571, a Recall of 0.5455 and F1 Score of 0.6667, and an Accuracy of 0.7143. These metrics provide some insights into the performance of the model, especially when resource allocation tasks are at the agenda and the correct identification of positive cases is important. Based on the model, it is seen that even though KNN may be useful in some circumstances it might not be the most appropriate technique to use when the most important goal is to identify all the positive cases in resources that are scarce.

Precision is calculated as the number of actual positives among all positive cases predicted by the model is 0.8571 for KNN. From this it can be concluded that when KNN has classified a case in the positive class, it is correct most of the time. However, the Recall score of 0.5455 points that KNN is only able to classify a little over half the actual positive cases in the data set. A lower Recall means that there is a potential of filtering out a large number of these positive cases region that is a serious issue in resource allocation tasks.

The F1 Score, which balances Precision and Recall into one Score, is 0.6667 for KNN. It indicates that while the model has reasonably good Precision, Recall is lower and affects the model's performance. It means that the KNN model works poorly both in terms of the model's capacity to point to all the positive samples accurately and in terms of the model's capability to capture all the positive cases.

Accuracy, at 0.7143, indicates that the KNN model correctly classifies about 71.43 % of the instances of the dataset. Although this is a rather good Accuracy rate, it proves once again that KNN algorithms' performance is not as consistent as some of the other algorithms, especially when it comes to positive observations. In situations where negative cases are far much more than positive or in a situation where a number of positive cases are few these factors complicate the model more by making it difficult to capture all the relevant positive cases.

In Summary, the Precision and the Accuracy look reasonable in the KNN model, but the lower Recall and the F1 Score do show major drawbacks of the model in identifying all positive cases, especially if the datasets are imbalanced or processors are limited. These limitations imply that if there are situations where KNN can be useful, probably it is not the best method to use if the aim is to have an inclusive way of identifying the positive cases. For tasks, outcomes where it is important to capture all of the positive instances, other models or techniques that have less Accuracy, but more Recall and F1 Score might be more effective.

```
Model: KNN  
Precision: 0.8571  
Recall: 0.5455  
F1 Score: 0.6667  
Accuracy: 0.7143
```

Figure 15: KNN accuracy

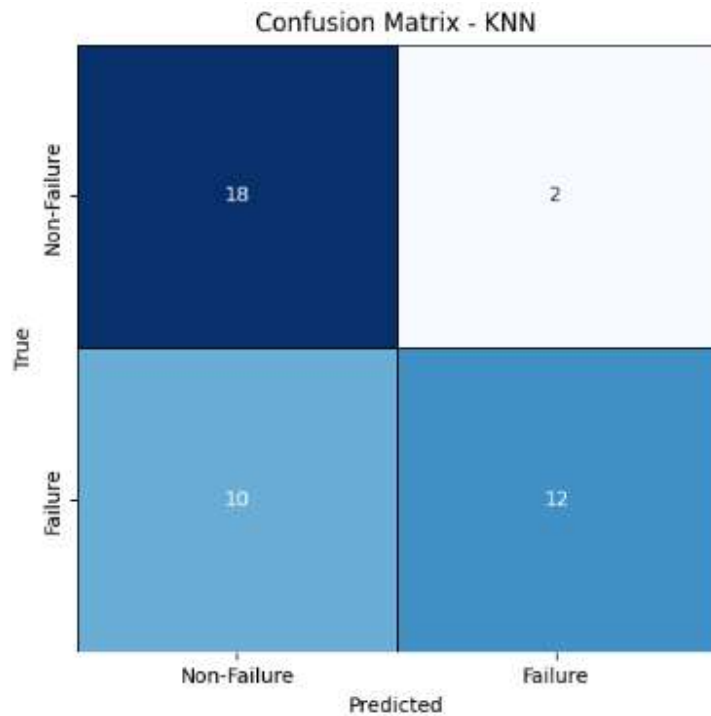


Figure 16: Confusion Matrix of KNN

7.3.7. Gaussian Naive Bayes (Gaussian NB)

The overall accuracy of the model Gaussian NB was a Precision of 0.9167, a Recall of 0.5000, and an F1 Score of 0.6471; and for the Accuracy, 0.7143. These metrics provide insight into the model's strengths and limitations, especially in its ability to classify positive instances accurately.

Precision is high while recall is comparatively low, which informs that the model is highly selective in identifying those to be tagged as positive and, at the same time, many potentially relevant positives are overlooked. The recall score of 0.5000 suggests that the model does not identify all actual positive cases, missing about 50% of them. The F1 score, which balances both precision and Recall, was 0.6471 is less effective at capturing all true positives. The Accuracy of the model is 0.7143 meaning that it is correctly classified about 71.43% of the instances in the dataset, it further underscores that the model's performance could have been more improved.

The results with Gaussian NB raise the question about its potential to predict instances positively when it does so, though with a low Recall it might not be the best choice for the cases which require

identification of as many positive samples as possible. The Accuracy score also makes it clear that despite being reasonably accurate, Gaussian NB could be even better if Recall were optimised.

```
Model: GaussianNB  
Precision: 0.9167  
Recall: 0.5000  
F1 Score: 0.6471  
Accuracy: 0.7143
```

Figure 17: Accuracy of GaussianNB

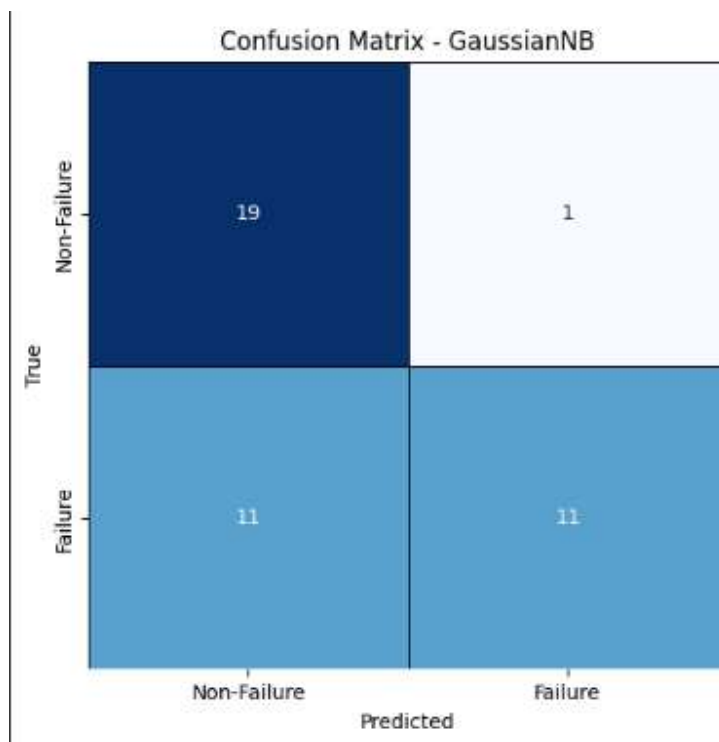


Figure 18: Confusion Matrix of GNB

7.3.8. Bernoulli Naïve Bayes (Bernoulli)

Bernoulli achieved a Precision of 0.9375, Recall of 0.6818 and an F1 Score of 0.7895 with an Accuracy of 0.8095. As in the case of Extra Trees, these metrics indicate the model's performance,

particularly in minimizing false positives. Bernoulli also has high scores in Precision and F1 Score. Precision, which calculates the proportion of true positive predictions among all positive ones, is high at 0.9375, which shows that the model is very effective at correcting positive instances, with few false positives. By looking at the Recall score of the model, the model manages to select a reasonable ratio of positive observations and they are slightly lower than that of the Gradient Boosting and SVC models. Since the results are constant with Extra Trees, it can be stated that Bernoulli is suitable in cases where high Precision is desired, although there is potential for increasing Recall, which will make the model even more robust in crucial times.

```
Model: BernoulliNB
Precision: 0.9375
Recall: 0.6818
F1 Score: 0.7895
Accuracy: 0.8095
```

Figure 19: Accuracy of BNB

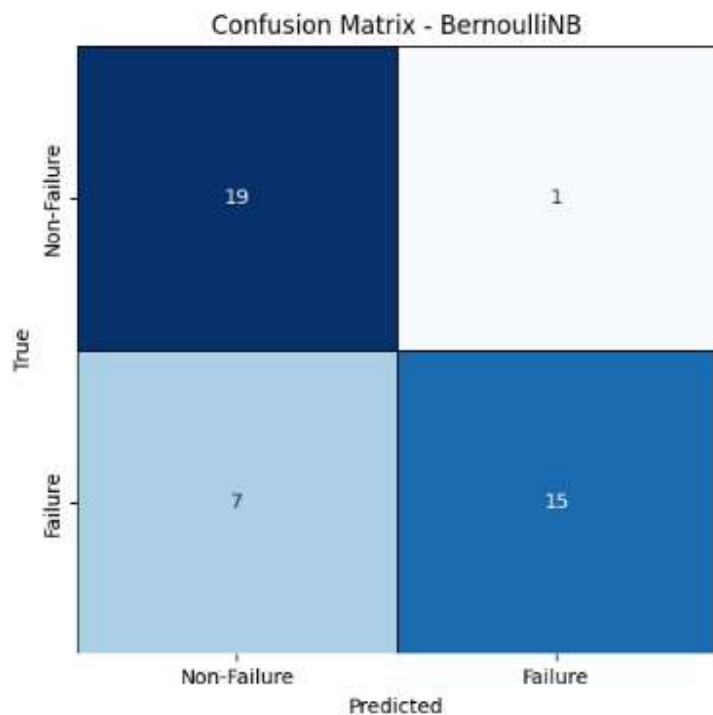


Figure 20: Confusion Matrix of NB

7.3.9 Support Vector Classification (SVC)

The figures for other parameters of performance are as follows: The Precision of 0.8500, and a Recall of 0.7727 and F1 Score is 0.8095 for Specificity and an Accuracy of 0.8095. This metrics indicates that the model is effective in predicting positive cases with a good balance in both precision and recall. The fairly high Precision and Recall show that SVC is extraordinarily good at detecting the positive cases without allowing too many false positives. The high F1 Score means that both Precision and Recall are generally high hence making SVC a reliable tool for tasks with high accuracy. We can note that the performance of the model coincides with Gradient Boosting, which indicates that SVC is a good solution for tasks in which both Precision and Recall are critical, which are relevant for resource distribution.

```
Model: SVC
Precision: 0.8500
Recall: 0.7727
F1 Score: 0.8095
Accuracy: 0.8095
```

Figure 21: Accuracy of SVC

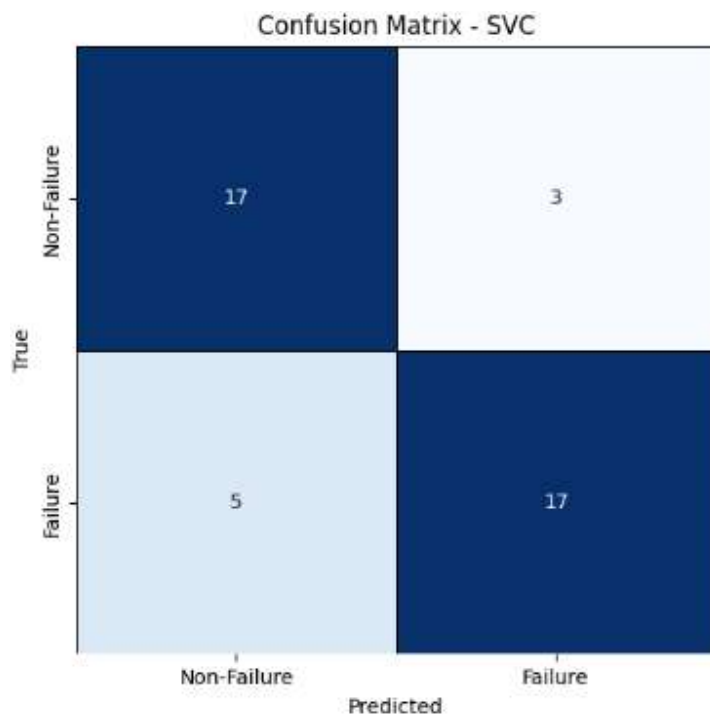


Figure 22: Confusion Metrix of SVC

7.3.10 Logistic Regression

Logistic Regression gave the Precision of 0.8750, a Recall of 0.6364, F1 Score is 0.7368, and an Accuracy of 0.7619. The precision suggests that 87.50% of the instances predicted as positive were correct. This shows that the model is good at minimizing false positives. It however has a poor Recall, which means that it loses a lot of positive predictions in classification. F1 score indicates the balanced view of model's performance. If the precision is high and recall is low, it affects the overall performance. The Accuracy score is reasonable, but if the Recall is improved, it may be beneficial for the capturing of similar occurrence. From the Comparison of Logistic Regression, it can be inferred that even though this model is highly efficient Logistic Regression may not be the most suitable model for the tasks where high Recall is desirable. Improving recall could make the model work more efficiently.

```
Model: LogisticRegression
Precision: 0.8750
Recall: 0.6364
F1 Score: 0.7368
Accuracy: 0.7619
```

Figure 23: Accuracy of LR

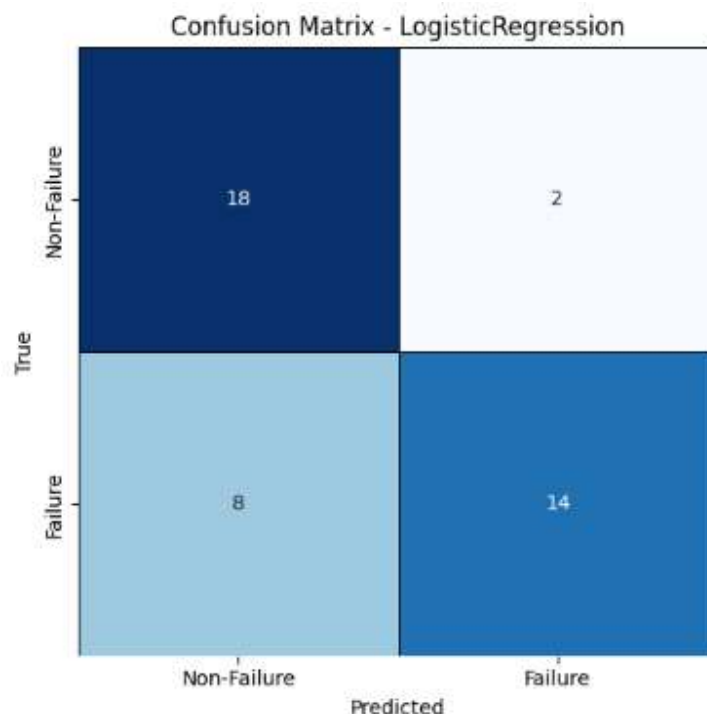


Figure 24: Confusion Matrix of LR

7.3.11 Stochastic Gradient Descent Classifier (Declassified)

The performance of Declassified model is, Precision is 0.9231, Recall of 0. 5464 and an F1 Score of 0.6857 and Accuracy was 0. 7381. The model is highly effective in predicting positive instances correctly when it makes a positive prediction. There is lower Recall and F1 Score which means the model is very selective with positive predictions that could be problematic in scenarios to identifying all positive is crucial. Accuracy of 0.7381 indicates that the model correctly classifies 73.82% of instances overall. This performance shows that Declassified is not perhaps the most suitable for cases when it is important to select all positive samples. A moderate Accuracy score mean that, although the strengths of the model are supported, recall deficiencies could influence the enhanced effectiveness of the model.

```
Model: SGDClassifier  
Precision: 0.9231  
Recall: 0.5455  
F1 Score: 0.6857  
Accuracy: 0.7381
```

Figure 25: Accuracy of SGD



Figure 26:Confusion Metrix of SGD

7.3.12 Hard Voting Classifier

The Hard Voting Classifier did a very well by attaining an average Precision of 0.8824, a Recall of 0.6818, an F1 Score of 0. 7692, and an Accuracy of 0. 7857. In addition to this, this model is a type of ensemble learning model that aggregates predictions from the multiple classifiers, making it have a balanced approach to prediction. With a precision of 0.8824, the model classifies a high proportion of positive instances among all its positive predictions. Recall score of 0.6818 shows that the model successfully identifies approximately 68.18% of all actual positive cases. F1 score reflects a balance between precision and recall with 0.7692 identifying positive instances and minimizing false positives. The accuracy of 78.57% suggests that overall performance of the model is good.

The performance metrics are similar to Random Forest and AdaBoost models. Hence, Hard Voting Classifier takes advantage of merits of various models for high Reliability. As the results in the current study show, this method of aggregating the predictions made by different models can be rather reliable and contribute to the increase of the models' accuracy.

```
Model: Hard Voting Classifier  
Precision: 0.8824  
Recall: 0.6818  
F1 Score: 0.7692  
Accuracy: 0.7857
```

Figure 27:Accuracy of HVC

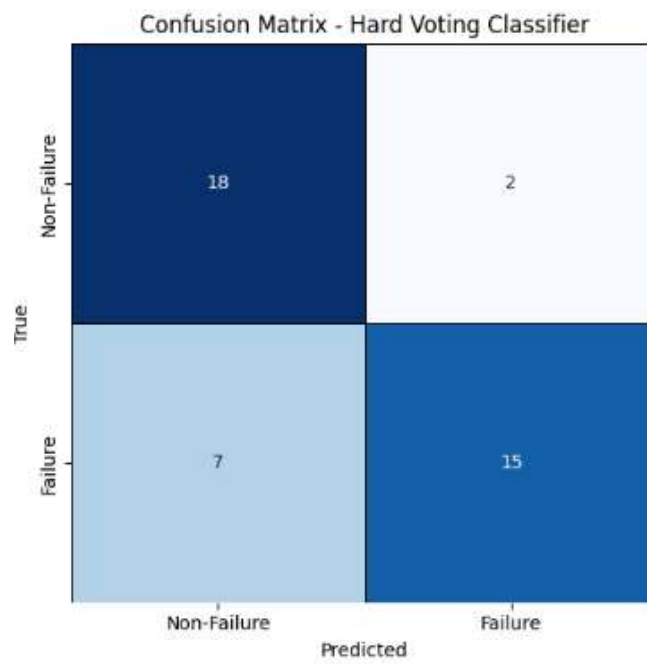


Figure 28: Confusion Metrix HVC

Chapter 8. Results and Discussion

The results indicate that accuracy and efficiency of the prediction into resource allocation has diverse strength and weakness in every model of machine learning. There is no universally better model; the choice of model is determined by specific conditions of a certain application to achieve, for example, low false positive rate or 100% recall. The models particularizing high Precision, Recall, and F1 Scores are Gradient Boosting, SVC, which shows that those models should fit best to contexts where precise object determination and the pinpointing of positive instances are both relevant. Precision also remains high with Extra Trees and Bernoulli again indicating that they can effectively lower the amount of false positive cases though their Recall indicates that there is still potential for improvement. All these algorithms – Random Forest, AdaBoost and Hard Voting Classifier are quite well balanced which will allow them to be used across a range of applications. Their Precision is not the highest while Recall is not the lowest, so their results can be considered reliable in practical cases when all the measures are important.

It can be observed that the Recall scores of these classifiers such as KNN, GaussianNB, and Logistic Regression are relatively low, meaning that it has the problem of selecting all the positive samples accurately. These models may seem to be less optimal when merely hunting for instances as many as possible is important. The promulgation of the analysis also reveals that depending on the requirement or preference toward either Precision or Recall, one should select the requisite model. Further improvements in the models could be made by incorporating other methods such as ensemble methods, and hyper parameter optimization since some of the limitations that were seen earlier would have been solved. Although summarized simply, the recommendations derived from such analysis offer general advice on how to enhance decisions about resource management and have high accuracy in predictive models in business intelligence.

Model	Precision	Recall	F1 Score	Accuracy
Gradient Boosting	0.8889	0.7273	0.8000	0.8095
Random Forest	0.8824	0.6818	0.7692	0.7857
AdaBoost	0.8824	0.6818	0.7692	0.7857
Extra Trees	0.9375	0.6818	0.7895	0.8095
Decision Tree	0.8333	0.6818	0.7500	0.7619

KNN	0.8571	0.5455	0.6667	0.7143
GaussianNB	0.9167	0.5000	0.6471	0.7143
Bernoulli	0.9375	0.6818	0.7895	0.8095
SVC	0.8500	0.7727	0.8095	0.8095
Logistic Regression	0.8750	0.6364	0.7368	0.7619
Declassified	0.9231	0.5455	0.6857	0.7381
Hard Voting Classifier	0.8824	0.6818	0.7692	0.7857

Table 1: All Model's Results

```

=====Best Model=====

Model: SVC
Precision: 0.8500
Recall: 0.7727
F1 Score: 0.8095
Accuracy: 0.8095

```

Figure 29: Best Model

Chapter 9. Conclusion

9.1. Overview

The project has been beneficial in endeavouring to compare several machine learning models for predictive analysis of resource allocation. What the results show is that the models such as Extra Trees and Bernoulli stand out in terms of Precision, which is beneficial in conditions where false positive cases must be avoided. On the other hand, Gradient Boosting and SVC have high Precision and relatively high Recall, proving themselves as all-round models that can be used for precision-oriented work and at the same time, they do not neglect comprehensive classification results. Accordingly, the focus of choosing a particular model should be cast according to the application requirements, and the work examines the idea that there is no model which can be globally considered as superior. Still, each of these models has its advantages and shortcomings; therefore, the model to be used should be defined by the need of the given job. The Precision measure is for applications that cannot afford False Positives hence models with high Precision include [Random Forest Classifier, AdaBoost Classifier, and XG Boost Classifier] On the other hand, applications that need to balance between recall and accuracy include Gradient Boosting and SVC. More research should continue with improving the current model and should look at other methodologies and other data sets. Through the means of constant improvement of the model selection and the evaluation procedures, the organizations will be able to enhance the application of the predictive analytics in efficient decision making that responds to the strategic allocations of resources. From this project, one can obtain significant information that can be used as a theoretical basis for further work on the enhancement of models for making predictions and their application. Finally, the project proves the significant importance of machine learning in terms of prospective predictive analysis regarding resource distribution and provides practical advice on model identification and optimization. When choosing between the models, advantages, and disadvantages of each should be weighed so that an organization can integrate more advanced analytics, such as machine learning and AI by leveraging predictive analytics for resource allocation (Watson 2009).

9.2. Recommendation

Comparing predictive analyses of various machine learning models for resource allocations it has been found that all the models in question have certain unique strength and therefore what is

effective for one application may not be as effective for another. Based on the results, the following recommendations are made:

- **High Precision needs:** For situations where accuracy is paramount—particularly for situations where the relevant false rate is of significance—Extra Trees and Bernoulli should be recommended. In this case, Precision reaches the highest value in both modes, which is equal to 0.9375; this means that, the best models accurately select positive cases. This characteristic makes them good for applications for which it is useful to minimize false positives, such as in fraud detection, or when misallocation of critical resources can be costly.
- **Balanced Precision and Recall:** The models such as Gradient Boosting & SVC will balance well since Precision and Recall are good, but not excellent; it is more balanced and accurate and thus appropriate for applications that don't need extra precision at the cost of missing out some instances. Gradient Boosting distinguishes itself with a rather high F1 Score of 0.8000, which proves the algorithm's ability to maintain an optimal balance of Precision and Recall. SVC also has a reasonable accuracy of F1 Score of 0.8095 and therefore can be used where it is necessary to have high accuracy of identification while also being able to capture input instances of interest. These models would be appropriate to the applications, where both, aspects of performance are critical, such as in strategic resource decisions where false positives or false negatives need to be managed.
- **Versatile models:** Random Forest and AdaBoost are relatively equally powerful but slightly less Precise and Recall than the best models. They are still good to go for normally used classifications where you can afford to compromise your precision for recall or vice versa. These models can be used when there is a need to work with models when constraints such as resource limitation or realities of the situation disallow more complex ones. Because of these equal performances, they are versatile to operate in different applications without complex adjustments and calibration.
- **Low Recall Models:** Recall value of KNN, GaussianNB and Logistic Regression is comparatively less, due to improper classification of all positive instances. For concept-variation models where it is essential to determine each potential instance of the concept

application is crucial. This could entail tuning the hyperparameters, feature selection or extraction or conduct augmentation of the data set to improve their performance.

- **Model Improvement and Future Research:** Future enhancement of model performance can be done through Ensemble methods like stacking, boosting, or bagging. It could be adopted in order to enhance the efficiency of the predictive models. There are couple of aspects which can be improved, one of them is feature engineering. If one wants to make the model more sensitive to the patterns which exist in the data then he is always free to explore new features or transform the existing features. Moreover, using, for example, market or economic indicators as sources of input data can give more essential insights and enhance the productivity within the organizations (Manikya et al. 2011)
- **Practical Consideration:** It is important to evaluate the chosen model in terms of computational complexity for larger data sets. Although, there are models such as Extra Trees and Gradient Boosting that give high accuracy, they may take a lot of computational power. Thus, their applicability should be judged in relation to what is currently available in terms of infra-structural and resource supports. For the real-world application, one may prefer models that offer reasonable accuracy and performance along with a manageable computational requirement which may be met by Random Forest or the Hard Voting classifier.

9.3. Novelty of the Work

To some extent, it is possible to distinguish the follow key sources of originality in this project. First, it introduces the ensemble approach into the work being done here. Thus, in the model Hard Voting Classifier, several models are combined, and even their weak sides are used to construct a powerful and accurate structure of resource distribution among agents. This gives an ensemble strategy the advantage of making sure that the weaknesses of every model are being cancelled by other stronger models thus making the result or the forecast more accurate.

Second, the integration of machine learning into the process of resource allocation, especially in the healthcare industry or manufacturing or logistics illustrate the idea of Artificial Intelligence as a technique for improving decisions made. This project also includes the domain-specific feature engineering which comprises the development of features relevant only to resource distribution

tasks. Consequently, we are left with a model that although is good, is particularly well equipped to address the issues of resource allocation.

Lastly, this work contributes by employing the hyperparameter tuning in combination with the ensemble method, which produces the fine-tuned predictive model of resources with the consideration of their constantly changing requirements. For example, in supply chain management or in a health care setting, where proper allocation of resources is paramount, this type of predictive analysis can go a long way in helping the industries to become more efficient, reduce wastage and properly assign resources where they are required most.

9.4 Future Work.

The future research could consider more complex ensemble methods and tune hyperparameters to exploit potentially even better results. Furthermore, up to date data and increasing external data use and new feature engineering methods can enhance the accuracy and resilience of the forecasts. Through repetitive improvements of the selection and assessment models, an organization can enhance their use of predictive analytics to enhance their resource utilization and choices.

References

- [1] McAfee, A. and Brynjolfsson, E., (2012). Big Data: The Management Revolution. *Harvard Business Review*, 90(10), pp.60-68.
- [2] Davenport, T.H. and Harris, J.G., 2007. *Competing on Analytics: The New Science of Winning*. Harvard Business Review Press.
- [3] Chen, H., Chiang, R.H.L. and Storey, V.C., (2012). *Business Intelligence and Analytics: From Python g Data to Python g Impact*. *MIS Quarterly*, 36(4), pp.1165-1188.
- [4] Mood, A.G., (2024). Predictive Analytics for Resource Allocation and Management in Libraries. *Library Progress International*, 44(1), pp.208-225.
- [5] LaValle, S. et al., (2011). *Big Data, Analytics, and the Path from Insights to Value*. *MIT Sloan Management Review*, 52(2), pp.21-32.
- [6] Wang, Y. and Hali, N., (2017). Exploring the Path to Python g Data Analytics Success in Healthcare. *Journal of Business Research*, 70, pp.287-299.
- [7] Waller, M.A. and Fawcett, S.E., (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, 34(2), pp.77-84. Available at: <https://ssrn.com/abstract=2279482>
- [8] Abbasi, A., Sarker, S. and Chiang, R.H., (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, 17(2), pp.1-32. Available at: https://aisel.aisnet.org/jais/vol17/iss2/3?utm_source=aisel.aisnet.org%2Fjais%2Fvol17%2Fiss2%2F3&utm_medium=PDF&utm_campaign=PDFCoverPages
- [9] Dey, N. K., Choudhary, S. and Tan, A. M. L., (2020). Challenges and best practices in implementing business intelligence systems: A review of literature. *Journal of Business Research*, 112, pp.230-245.
- [10] Wamba, S.F., Akter, S., Edwards, A., Chopin, G. and Ganzhou, D., (2015). How ‘Big Data’ Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study. *International Journal of Production Economics*, 165, pp.234-246. Available at: https://www.researchgate.net/publication/270276259_How_%27big_data%27_can_make_big_impact_Findings_from_a_systematic_review_and_a_longitudinal_case_study

- [11] Ramsbotham, S., Kiron, D. and Prentice, P.K., (2016). Beyond the Hype: The Hard Work Behind Analytics Success. *MIT Sloan Management Review*, 57(3), pp.1-9.
- [12] Corte-Real, N., Oliveira, T. and Ruivo, P., (2017). Assessing Business Value of Big Data Analytics in European Firms. *Journal of Business Research*, 70, pp.379-390. Available at: <https://doi.org/10.1016/j.jbusres.2016.08.011>
- [13] Wixom, B.H. and Watson, H.J., (2010). The PYTHON -Based Organization. *International Journal of Business Intelligence Research*, 1(1), pp.13-28.
- [14] Redman, T.C., (2013). *Data Quality: The Field Guide*. Digital Press.
- [15] Gandomi, A. and Haider, M., (2015). Beyond the Hype: Python g Data Concepts, Methods, and Analytics. *International Journal of Information Management*, 35(2), pp.137-144.
- [16] Grover, V., Chiang, R.H., Liang, T.P. and Zhang, D., (2018). Creating Strategic Business Value from Big Data Analytics: *A Research Framework*. *Journal of Management Information Systems*, 35(2), pp.388-423.
- [17] Gupta, M., and George, J.F., (2016). Toward the Development of a Big Data Analytics Capability. *Information & Management*, 53(8), pp.1049-1064.
- [18] Smith and J. Doe., (2022). Data Protection and Privacy: The GDPR and CCPA Impact on Data Analytics. *Journal of Data Protection & Privacy*, pp 45-63.
- [19] K. Williams and R. Patel., (2020). “Machine Learning Fairness and Bias: A comprehensive Survey” *IEEE Transactions on Neural Networks and Learning Systems*, pp 1378-1392.
- [20] Thompson and A. Gupta., (2022). Transparency in Machine Learning: Ensuring Fair Resource Allocation, *ACM Transactions on Intelligent Systems on Technology*, pp 28-43.
- [21] M. Brown and E. Lee., (2021). Ethics of AI: Addressing Bias and Discrimination in Machine Learning, *Journal AI and Society*, pp 379-395.
- [22] Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V., (2017). Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Business Research*, 70, pp.263-286. Available at: <https://doi.org/10.1016/j.jbusres.2016.08.001>
- [23] Lee, C.S., Cheang, P.Y.S. and Mosholu, M., (2022). Predictive analytics in business analytics: decision tree. *Advances in Decision Sciences*, 26(1), pp.1-29. Available at: <https://dx.doi.org/10.47654/V26Y2022I1P1-30>
- [24] Watson, H.J., (2009). Tutorial: Business Intelligence – Past, Present, and Future. *Communications of the Association for Information Systems*, 25, pp.487-510. Available at:

https://www.researchgate.net/publication/328891588_Tutorial_Business_Intelligence_-_Past_Present_and_Future

- [25] Manyika, J., Chui, M., Brown, B., Beghin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A., (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. *McKinsey Global Institute*. Available at: https://www.researchgate.net/publication/312596137_Big_data_The_next_frontier_for_innovation_competition_and_productivity
- [26] Akingbola, S.M. and Buoyed, P.A., (2024). Predictive Analytics for Resource Optimization in Data Warehousing and Data Mining Using Random Forest. *International Journal of Women in Technical Education and Employment*, 5(1), pp.33-39.
- [27] Negash, S. and Gray, P., (2008). Business intelligence. Handbook on decision support systems 2, pp.175-193.
- [28] Turban, E., (2011). Decision support and business intelligence systems. Pearson Education India.
- [29] Novoa Dita, C.A., (2019). Implementation of a Business Intelligence Solution in a Manufacturing Company: A Predictive Analysis Approach (Doctoral dissertation, Polytechnic di Torino). Available at : <http://webthesis.biblio.polito.it/id/eprint/12669>
- [30] Rikhardsson, P. and Polytechnic, O., (2018). Business intelligence & analytics in management accounting research: Status and future focus. *International Journal of Accounting Information Systems*, 29, pp.37-58.
- [31] Olaniyi, O., Amalaka, A. and Onabanjo, S.O., (2023). Utilizing python g data analytics and business intelligence for improved decision-making at leading fortune company. *Journal of Scientific Research and Reports*, 29(9), pp.64-72.
- [32] Basile, L.J., Carbonara, N., Pellegrino, R. and Pan niello, U., (2023). Business intelligence in the healthcare industry: The utilization of a data-driven approach to support clinical decision making. *Tec novation*, 120, p.102482.

Bibliography

- [1] Kaggle (2024) Predictive Maintenance Dataset. Available at: <https://www.kaggle.com/datasets/hiimanshuagarwal/predictive-maintenance-dataset> (Accessed: 20 May 2024).
- [2] Scikit-Learn: Machine Learning in Python. Available at: <https://scikit-learn.org/stable/> (Accessed: 14 June 2024).
- [3] Google Schola for Research papers and articles. Available at: <https://scholar.google.com/> (Accessed: 28 June 2024).