A Decentralized Employee Performance Appraisal Framework for Recruitment, Performance Prediction and Ranking using Permissioned Blockchain and Ensemble Learning

by

Afra Antara Anjum
18101220
Sadaath Islam
18101227
Shaikat Majumder
18101630

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Afra Antara Anjum

18101220

Sadaath Islam

18101227

Shaikat Majumder

18101630

Approval

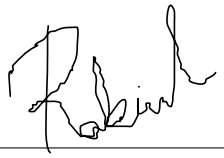
The thesis/project titled "Decentralized Employee Performance Appraisal Framework for Recruitment, Performance Prediction and Ranking using Permissioned Blockchain and Ensemble Learning" submitted by

1. Afra Antara Anjum (18101220)
2. Sadaath Islam (18101227)
3. Shaikat Majumder (18101630)

Of Fall, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 15, 2022.

Examining Committee:

Supervisor & Program Coordinator:
(Member)



Dr. Md. Golam Rabiul Alam

Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Professor
Department of Computer Science and Engineering
Brac University

Abstract

Recruitment is a crucial task for Human Resource Management (HRM) and determines the creation of a competent workforce that eventually brings tangible and intangible benefits for companies. Employees are key elements in determining a company's success and employees perform well when their skill set complements their job requirements. However, the current system fails to provide a single solution that verifies employee records and predicts employee-company compatibility. This paper proposes an recruitment system using a private permissioned blockchain architecture and ensemble learning algorithms. The paper proposes a permissioned blockchain architecture using permission protocol and smart contracts to store employee records in an immutable ledger. Development of Data and processing decentralization is inspired and in accordance with the Hyperledger Fabric system design, thus creating a decentralized data sharing system that is used to hold comprehensive employee performance records in a peer-to-peer system that allows employee data verification and retrieval by organizations in the blockchain consortium. The applicant records and previous performance appraisal records can be retrieved by a company in the consortium following smart contract rules and can be used to predict employee performance ratings based retrieved previous performance appraisal records. To predict the performance score, we used machine learning models namely supervised and ensemble learning. The system also ranks eligible candidates, based on predicted performance scores and other relevant applicant data via Multi-Criteria Decision Making Algorithm (MCDM). Finally, a Streamlit application is created where performance score predicting and ranking are done automatically with a suitable user interface for final result output.

Keywords: Human Resource Management; Recruitment; Verification; Machine Learning; Ensemble Learning; Blockchain; Ranking; Prediction; Performance Appraisal; Decision Tree; Random Forest; Recursive Feature Elimination; XGBoost; Streamlit; Hyperledger Fabric; Smart Contracts: Permission Protocol

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor Dr. Md. Golam Rabiul Alam sir for his kind support and advice in our work. He helped us whenever we needed help and his constant supervision and guidance allowed us to learn about different frameworks used and helped us complete our thesis.

And finally we would like to dedicate our research to our beloved family members. Without their unwavering support, encouragement and faith in us, we would not be able to do our task successfully. With their kind support and prayer we are now on the verge of completing the final phase of our undergraduate thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The world is moving towards a point where diversity is becoming a necessary concept in every organization. To handle all of the problems posed by people working from different cultures and backgrounds, the human resource (HR) department has become an integral part of any given organization. One of the jobs that they handle is recruiting new people for their respective organizations.

However, the hiring process can be very tedious, time-consuming, and may frequently be based on unverified information provided by the candidates. A company's business functionalities depend on its human resources. Key attributes, such as a candidate's expertise, knowledge, and work satisfaction determine their work efficiency and performance. The selection of a suitable candidate is highly prioritized as their performance determines the efficiency and functioning of the company, which in the long-run impacts business [19].

The hiring process is also an investment since the company needs to fund the human resource to organize employer branding programs, recruitment competitions, and interviews to attract potential employees and find suitable candidates from the applicant pool. Currently, we still recruit by running web-based or manual recruitment campaigns and then selecting candidates from the applications. Later selected candidates are verified and called for an interview by HR. But, the current process fails to guarantee the effective hiring of suitable candidates for job roles. Therefore, making this entire process slow, expensive and inefficient [19].

A bad hire may cost the company to invest their time and resources on the wrong candidates and more importantly lose out on the candidates that could have made a difference in their organizations. Thus, turning all the investments and efforts of hiring company valueless.

1.2 Research Problem

Today, the recruitment process mostly involves screening candidates from a pool of applications based on a fixed set of criteria and making them sit for a handful of assessments. Not only does this process hugely involve people working in HRM but

also gives them the power to select candidates who will contribute to future business goals. This process therefore often turns out to be biased as certain candidates participating in recruitment campaigns often receive added advantages or get selected in terms of favoritism. Moreover, conducting several assessment tests consumes the time, energy, and resources of both HRM and applicants, therefore slowing down the economy.

Additionally, screening through thousands of applications is hectic and slow [25]. Past records are often verified after the selection process and pave a way for fraud applicants to pass the initial stage of recruitment. This could have been avoided if verification could be done along with selection but with the current system, only manual verification methods are supported, hence, verifying thousands of applicants is an unimaginably cumbersome process, therefore, is often skipped.

In the present state, there are not many tools for employee verification since we are solely relying on the availability of previous records, sources, and references. It can be said the process is still manual as we have to contact the previous employer or reference to verify employee records. The lack of an automated system for employee verification has given rise to both wastage of HRM resources and fraudulence [22]. Firstly, to check employee records, the company either employs someone to check a candidate's previous records or hires a third party such as recruitment companies to do the task, therefore making this simple process expensive as now external vendor cost is involved. Many companies are small and do not have the budget for this cost and hence tend to skip the verification process altogether. Moreover, the process of contacting the company and the person in charge of employee verification through email or telephone is long since there might be many constraints and end up taking weeks or months to get back responses. Additionally, the verification process can also get slower than necessary due to holidays, close down periods, time zone differences, and unreachable key employees [22].

When employee verification takes such a long process, the entire hiring process will become inefficient, as a result, candidates might lose interest in the hiring company and take other jobs they are being offered in the current job market. It also imposes a threat of losing potential candidates who are better suited for the job role. Additionally, as the verification process is time-consuming, many companies tend to avoid a full background check and only contact the last employer. Companies are also at risk of losing valuable employee records due to natural disasters or fires which destroy office premises and hence databases are destroyed. Employee records may also be lost during relocating offices. Similarly, digitally stored employee records are lost due to malfunctioning storage devices. Companies often face a lack of space and therefore tend to routinely get rid of unnecessary employee information of former employees. Due to all these, many companies lose past employee records and cannot help in the employment verification process. Lastly, sometimes a company stops its business and ceases to exist, therefore losing its employee records and hence the data becomes inaccessible. In this case, former employers also might become unapproachable. Applicants without verification from former employers have fewer chances of getting hired and face the threat of becoming jobless.

Data regarding positions held by a candidate throughout his tenure in the company is often skipped during verification [22]. Many companies haste during the verification stage and are not able to verify all data and performance history. Therefore, the lack of information leads to companies misjudging a candidate's key qualities and failing to assign them appropriate job roles.

The current process also carries a major risk of spreading falsified information. Hiring employees just for verifying candidates or hiring third parties does not necessarily reduce the risk of misinformation. Many centralized systems for verification exist which take information from third parties. However, these are still at risk of cyber-attacks thus making participants' private information vulnerable to unauthorized parties who can modify the information. Many third parties are also known to sell data for marketing purposes [22].

According to [15], surveys conducted showed that during the recruitment process over 70% of job seekers apply with hidden and fraudulent information. Moreover, applicants tend to submit fake résumés, diplomas, and certificates of qualification, while others deliberately exaggerate their abilities while some even tamper their employee records and appraisal data. Participants also forge reference letters from companies they never worked in. The current system also does not solve the problem of managerial bias or the case of tampering of the recorded data by organization hence it fails to be secure and reliable [46]. Therefore, we can conclude that the current system is under the threat of self-promoting attacks, slandering attacks, and whitewashing attacks [48].

Social media background checks from recruitment sites such as LinkedIn are also a means of employee verification where candidates are judged based on their work-related information, network, and endorsements in a social setting. However, social media sites are prone to falsified information as many users put fake work information which these platforms cannot automatically verify unless reported. This not only misleads organizations but also paves a way for criminals to deceive general people as it puts their data at risk. Therefore, having a decentralized secure database system that can be accessed by such social media sites is necessary.

The final stage of recruitment is finalizing candidates based on their application, assessment, and interview results. In this stage, employers face the challenge of assigning the right talents to fit job roles so they can deliver maximum productivity at the workplace which will help boost business. However, recruiters often fail to understand candidate competency with job role requirements and favoritism also plays a factor in decision making. Therefore, computer software-based HR tools are used for predictive placement of candidates that ensures employee-job competency [19]. But due to the unavailability of reliable work history, precise decisions cannot be made and companies end up hiring employees who only have expected skill levels and hence have a contrasting resource pool compared to business objectives [19]. Moreover, many candidates might decline after being offered the job and according to [43], such an instance turns the total investment in the recruitment process into a loss. To avoid the problem, the prediction process also needs to consider both explicit and implicit data and analyze the candidate's willingness to join the orga-

nization.

Lastly, in the final stage recruiters may become indecisive while selecting candidates based on prediction value. The only way to aid this is to rank employees suitable for job roles after in-depth comparison and hence the most neutral method to rank is by using machine learning algorithms. Ranking helps to differentiate between candidates based on their performance prediction and aids recruiters to make decisions based on complete analysis.

1.3 Research Contributions

This research aims to develop a decentralized database system for recording employee information of individual organizations with the help of blockchain architecture, along with creating an all-in-one system to predict employee performance to aid in the applicant selection procedure for job vacancies. Currently, employee records can not be easily verified and the applicant selection process is inefficient. Therefore, our research aims to make the human resource system more structured and automated. The objectives of this research are:

- To store and verify employee records to aid in further decision-making and added transparency using suitable blockchain architecture.
- To evaluate the usage of machine learning algorithms to predict applicant performance appraisal using employee records.
- To rank candidates using MCDM model to further aid in comparison and selection process.

1.4 Research Methodology

This research is set to bring change in the recruitment process of human resources through developing an efficient and automated system using blockchain and machine learning. But first, data collection was our requirement to train our different types of machine learning algorithms to evaluate their performance. Therefore, we collected the open-source IBM HR Analytics from Kaggle which contained 36 features. In our dataset, there is the feature 'PerformanceRating" that we will be predicting throughout this research.

We have three parts in our research, first, our goal is to develop a Blockchain system to store employee records in a consortium of companies which allows peer-to-peer sharing of records through transactions. The stored data in blockchain is therefore immutable and allows verification. Employee performance appraisal records can be retreived from blockchain architecture to use with Machine learning models for performance appraisal prediction and lastly, we will rank our candidates based on their features and predicted performance scores. In the end, we create a Streamlit application to automate the prediction and ranking process.

1.5 Scope and Limitation

The scope of this research is to enhance the recruitment process of human resource management. Through this research, we hope to find a decentralized blockchain system where data is easily stored and shared maintaining security protocols, within a consortium of companies. The evaluation of the machine learning algorithms hopes to find the best model for datasets to provide for a more efficient system.

The dataset that we have utilized in this research is not a real-life dataset but rather an engineered one. On top of that, the dataset does not contain sufficient data entries which otherwise could have further boosted the research insights. Finally, our research is only concerned with the architectural and algorithmic controls and not any physical aspects such as machines and hardware analysis for the blockchain structure.

1.6 Document Outline

The rest of this dissertation is designed as follows: Chapter 2 contains the literature review. It describes in detail all related work that has been done using blockchain and machine learning in human resource management systems. It also focuses on different types of blockchain frameworks and their comparisons. This chapter also highlights different machine learning algorithms and uses of Multi-Criteria Decision Making (MCDM) and Topsis method in ranking candidates.

Chapter 3 focuses on the methodologies used for our system. It discusses all the machine learning algorithms and blockchain architectures that we have analyzed and used.

In Chapter 4 we have discussed data preparation for the prediction and rating of candidates. We stated our source for data-set, carried out data pre-processing to make it usable for the algorithms.

The architecture of the developed blockchain model is described in Chapter 5. The entire methodology is clearly described and its application is related to the recruitment process of HRM.

Chapter 6 describes the implementation of our proposed methodologies and analyses their results. This chapter discusses all the implementations of machine learning models in different stages of development. Moreover, it also describes how the developed Streamlit app works.

Lastly, Chapter 7 consists of the conclusion we arrive at from the research and discusses further scopes of development.

Chapter 2

Literature Review

2.1 Blockchain Architecture

In the research work [10] it is found that only 20% of papers work with applications of blockchain technology while the rest 80% are based on the bitcoin system. According to [46], Blockchain technology has scope for massive innovations due to its evolving nature. The paper also states that the current human resource management (HRM) system requires a technology that is secure and transparent for verification and validation of employee records, and therefore proposes the use of blockchain technology.

There are 4 broad groups or types of blockchain networks. Although all of these 4 are peer-to-peer networks, there are fundamental differences among each type. Public blockchains are networks that allow the public to be added to the network as an actor. According to [17], this means that this type of network is much more decentralized in nature as compared to the private blockchain network. The private network only allows certain actors to interact with the blockchain or to be a party in the blockchain. This means that the private network is faster in translation as the parties are much more limited. The consortium blockchain can open parts of the network to public actors while keeping others close to the private actors in the network. The hybrid blockchain contains the characteristics of both private and public blockchain making the cases according to the given circumstances.

The research work [15] states that the concept of blockchain was first presented in the paper "Bitcoin: A Peer-to-Peer Electronic Cash System," which was published in 2008 by Satoshi Nakamoto. According to [15], blockchain is a technical program that through its distributed node can store, validate, transfer and communicate network information independently. Therefore, [15] recognized blockchain technology to be a secure and transparent system to be combined with the current HRM system to deliver an authentic platform for employee data. Research work [38] also proposed a blockchain-based HRM due to the technology being transparent, distributed, and decentralized in nature. Research work [47] discusses the development of the Interplanetary File System (IPFS) with Blockchain to store information without any file size constraints in a distributed and decentralized environment, therefore, ensuring transparency. Here the IPFS system creates a distinctive hash of the document based on its contents and uses a Smart Contract to record the hashes, therefore creating an immutable system.

On the other hand, according to [48], trust management is an essential component to ensure the reliability of the organization and its employees. Without trust in the processes and the organization, the system becomes susceptible to three types of attacks, namely - slandering attacks, self-promoting attacks, and whitewashing attacks. Therefore in [48] a system is proposed to verify the information for organizations and employees and is implemented through four steps. Firstly, by the establishment of a consortium where all organizations are blocks and decide the addition of another block through negotiation and then rate the new block to indicate trustworthiness. The second is the generation of service data when requested by the user. The third is block synchronization where a time-stamped series of immutable blocks contain the transaction record. Blocks are immutable since a hash function is used to connect the blocks to form chains. Consensus protocol also verifies information entered in a new block. Finally, trust is created from the point of view of the user with itself, service, and platform. This is how the blocks will decide whether to cooperate with the others in the system through the trust links therefore ensuring a secure and trustworthy platform for data storage.

According to research work, [22], even though centralized solutions can verify someone's work history, those normally depend on third parties and hence are not tamperproof or do not eradicate falsified information. According to the same source, a basic requirement in the recruiting process is to be able to verify any applicant's work history and oftentimes, this can be very expensive. The use of blockchain can be used to solve this issue as the data entered into the blockchain can be verified by past employers and colleagues.

There are certain limitations to using blockchains. Since the data inserted into the blocks are immutable, if any invalid entries are made and somehow approved, there is no way to edit that entry. According to [44], blockchain ledgers are real-time and hence consume power while contacting all nodes whenever a new node is created. Hence, the sustainability of blockchain technology is questionable. According to the same source, big computing power is necessary because of the need for signature verification.

Although blockchain is a new technology in the computer science domain, the technologies have gained significant traction which is not limited to simple transactions of bitcoins which it started with initially as stated in [24]. Ethereum, which was announced in 2014 and was released and launched in 2015, fulfilled the need for a universal blockchain platform.

According to [14], ethereum can be thought of as a system with several nodes that store the state of the network in themselves. The information in the node stores the complete transaction history of the others and hence the global view or state is distributed and saved in all the nodes. According to [14], it is composed of two types of nodes which are externally owned and contract accounts. These are implemented via a combination of a public-private key or smart contracts. Every node or block has the following components; nonce, balance, storage roots, code has. A transaction can occur between blocks. However, to prevent an attack like a Dos

attack(distributed denial of service) there is a certain limit on the ability of transactions given by the GasPrice and Gas Limit. This means that the user needs to spend money on purchasing ether for the transaction to occur as the stated unit GasPrice and the transaction has a predefined amount of GasLimit that it will be consuming within which it must complete the transaction. This makes sure that the system is reliable and any anomaly can be detected by the system in a feasible time interval from breakout. According to [52], the Ethereum network supports two types of transactions in its model; namely the contract creation for the creation of new nodes and blocks. The other is the message call transaction between the already existing blocks.

Initially, the blockchain structures were developed as a better alternative to the cryptographic methods that could undergo decentralized transactions as in the implementation of bitcoin. However with time, especially after the launch of Ethereum in 2015, there has been a sharp increase in the domains where blockchain structures have been implemented. Sharing of employee information like details of their resume, work performance, or medical and private records without consent will violate the local data security laws of the country. According to [29], it is to be mentioned here that several locations do not have well-suited laws and regulations in place to sustain the network that is proposed. For these sensitive locations, the local laws and regulations are to be considered when drafting and programming the smart contract. Employee information and data sharing factors are to be communicated with all stakeholders. The infamous attack on DAO caused the abrupt decrease in the reliability of the ethereum network. According to [16], the attack caused the formation of a new ethereum classic abruptly in a short while. Furthermore, it is to be noted that the attack was possible as there was a logical error in the written smart contract. Hence the governance of the system and the parties governing the system are to be taken into special consideration. According to [41], it has been seen that the array of services involved and the number of businesses and mechanisms make the process more difficult for the users. It is a general goal for the framework to make the smart contracts as user-friendly as possible however it is often not the case. Special attention is hence needed in this part too.

The research [49] collected 108 testing networks as the testing dataset in the experiment. The paper discusses the reasonableness of the Hyperledger Fabric. The paper identifies reasonableness as are imperfections that cannot satisfy users' requirements during network configuring and bootstrapping, The paper states that it was not able to find any tool to analyze the reasonableness of the blockchain system, especially for the Hyperledger Fabric. The paper focuses on the static network reasonableness that needs checking, however, recommends working in research for dynamic network reasonableness in the future.

Hyperledger Fabric, which is one of the most prominent research domains for blockchain, is analyzed via performance diagnosis and optimization [51]. The paper takes into consideration some of the abnormal scenarios and tries to find out the performance effect due to these scenarios namely heavy I/O issues, complex business logic, heavy peer load, and complex business transaction, and suggest some of the performance optimization methods for each abdominal issue. It has been said that when the

transaction throughput reaches maximum level then we observe a bottleneck situation and some in-chain methods are suggested as solutions. Research [50] discusses the performance of Hyperledger Fabric and runs performance diagnoses in order to optimize the performance of the network. The paper claims to present a solution for performance diagnosis and optimization for a given runtime Hyperledger Fabric network. The authors also claim to optimize the network parameters, transaction proposals, and node resources.

Research [12] is a comparative analysis between the performances of Hyperledger Fabric and Ethereum. The findings of the research show that Hyperledger Fabric outperforms Ethereum in terms of latency and throughput and also consumes less hardware. However, Ethereum has a higher rate of a successful transaction according to the research. The paper also delves into further research scopes which should include other blockchain architectures.

Research [45] is a Systematic Literature Review(SLR) using 121 primary papers. Decentralization, audibility, and persistence are considered the 3 most important and prevalent features of any blockchain according to this paper. Researchers have identified two ways to solve the issues seen in the blockchain system. One is the on-chain system and the other is the off-chain system. The on-chain approach is essentially about making changes to the elements inside the blockchain to improve the efficiency of the system. The improvements can be such as increasing block size and sharding and consensus protocol-related approach. On the other hand, an off-chain solution is there to improve the blockchain efficiency via increasing the throughput by executing the transaction outside of the main blockchain via lightning network compatibility.

The research [33] talks about collecting the data for several thousands of articles for analysis purposes and then using data clustering to find out the trend of the research in the blockchain research space. The articles were taken from the period 2008 and 2019. Latent Dirichlet Allocation (LDA) was used to inductively identify topics from the corpus. It was seen that about 200 publications have a connection between business systems and Blockchain systems by using design science and prototyping approaches. It was said that the researchers in the articles that were analyzed used conceptual and prototyping methodologies to assume the initial consequences of blockchain and its design implications. But the value configuration, intermediate actors, and other third parties were last researched in these papers. They recommend more direct effort to be put into the research of the value of blockchain and the integration cost to putting it in the existing system. It has been mentioned that due to lack of research it is specifically unclear as to what are the private and public threats to blockchain systems. Strategic alignment of research is considered the most important factor when researching unconventional topics like blockchain.

2.2 Prediction and Rating

Research paper [18] focused on Green HRM using artificial intelligence where it is stated that Artificial Intelligence not only helps to reduce errors from 20% to minimum but also has reduced 55%-60% workload of HRM. Currently, Artificial

Intelligence has taken over 80% of the HRM system and plans to expand this to 100%. Hence, the addition of machine learning in HRM proves to be efficient, secure, authentic, and cost-effective.

In the research work [19], a web-based system is created to assist the recruitment process which consists of a skill assessment subsystem, an employee details verification subsystem, a performance predicting subsystem, and an attrition predicting subsystem. This system uses Blockchain for verification and several machine learning algorithms for candidates' performance prediction. Moreover, the system allows organizations to upload a set of resumes or employee history which are identified to be the key requirements for the vacancies. Yet, this system is not fully computerized as human interaction is required during verification. Moreover, the proposed system only predicts individual employee efficiency but fails to compare them in a pool of applicants for the job role, thus failing to suggest the best candidate to take the position. A rating system could have been used to solve this problem.

Research [43] predicts joining efficient candidates before selecting processes based on relevant features and uses statistical methods to select features and applies machine learning models. However, the system does not consider decision changes in humans and carries a risk of eliminating candidates with high competency.

Similarly, research work [20] states that while HR managers only consider a handful of basic criteria to predict employee performance, a machine learning-based system does that based on hundreds of criteria. Their proposed system focuses on establishing vacancies and setting competencies for each vacancy upon labor market demand and business requirements. The system also evaluates candidates' performance by comparing data from applications and competencies set for vacancies. But the system fails to be credible in terms of performance prediction due to a lack of past work appraisal records in evaluating candidates. Therefore, blockchain is considered to store past work evaluation records of candidates.

Research [32] is a study of feature elimination techniques using different classifiers to predict the land suitability for crop cultivation. Feature elimination is an essential part of machine learning as not all of the features or variables are useful or carry the same level of weight. One of the techniques is Recursive Feature Elimination (RFE) which is the most widely used one. RFE begin with the entire dataset as a whole and starts to remove weak features by ranking its importance and will continue to do so until only the most important features remain. Zulfikar et al shows in [9] how recursive feature elimination (RFE) affects the performance of decision tree model by running decision tree model with and without RFE. The results showed that the results have higher accuracy when decision tree is run with RFE.

Research [39] attempts to aid the human resource department by evaluating performance using computational intelligence. Their proposed system combines fuzzy logic with various classifying techniques such as Naive Bayes, decision tree, and artificial neural network to give predictions. The predictions are then translated into grades, Insufficient, Regular, Good, and Excellent.

Research [30] worked with another important aspect of the HR department, employee turnover. Their research attempts to do a risk assessment by predicting employee turnover with the help of computational tools such as random forest, decision tree, correlation matrix, etc. They recorded useful information of the employees such as age, marital status, job type, tenure, etc., and used these parameters to predict the turnover rate.

Research [57] is a study that aims to predict fraud in supply chain based on the machine learning algorithm XGBoost. According to the paper, XGBoost is a tree ensemble model which adds the results of all the trees it creates to predict the final predicted value. The algorithm first sorts the elements and then accesses the data sequentially and as a result, it accumulates gradient statistics in terms of loss function. According to the same paper, XGBoost has two major advantages, the first one being that it adds a regularization term to the objective function and the second one is that it uses second order derivatives along with the first order derivatives in order to make sure that the loss is more accurate. The paper also compares the XGBoost algorithm with Logistic regression algorithm and Gausian Naïve Bayes algorithm and the experimental data shows that XGBoost performs better than the other two in detecting fraud in supply chain.

A smart applicant ranker is developed in paper [21] which uses Ontology to compare candidates' Educational qualifications, skills, and work experience with job requirements and therefore aids in the candidates' selection process. However, the system fails to prove reliable since this system is solely based on information provided by candidates, and no fact-checking and additional information gain is involved.

In research work [7], teachers are selected and ranked using the Fuzzy TOPSIS method and therefore helps to differentiate among conflicting criteria. The paper states that Multicriteria Decision Making (MCDM) selects and ranks teachers. Fuzzy AHP is a type of analytical hierarchy process (AHP) which is an MCDM method, whereas Fuzzy TOPSIS is itself an MCDM method. In this research, weights are applied using Fuzzy AHP whereas Fuzzy TOPSIS is used in ranking. Technique for Order of Preference by Similarity to Ideal Solution also known as TOPSIS is based on the ideology that competing candidates should be closer to a positive ideal solution, which maximizes benefit and minimizes cost, and farthest from a negative ideal solution, which minimizes benefit and maximizes cost, in terms of distance. To deal with vague information Fuzzy TOPSIS is used which is an extension of TOPSIS.

Research [13] is a comparative analysis between two MCDM models, TOPSIS and VIKOR. The research established a product aspect ranking system to compare the performance of the two methods. The strengths and weaknesses of the two methods are also highlighted. Both the methods performed well but the datasets used were very small. Comparative analysis should be done on much bigger datasets.

As far as we have seen, there is a lack of research on the integration of blockchain technology and machine learning to make a secure and unbiased recruitment system that is fully automated and aids in decision-making based on verification and

ranking.

Chapter 3

Methodology

Our aim was to develop a system that includes a permissioned blockchain architecture that registers companies in the consortium thus allowing them to store employee records in a decentralized system and ensure data security using smart contracts. After verification, a job applicant's previous work record can be retrieved to be used in a Streamlit web application for performance appraisal prediction using ensemble learning techniques and ranking candidates using MCDM. In the Streamlit application, after performance appraisal prediction, the predicted values are used along with other applicant data to rank employees using TOPSIS. Figure 3.1 shows the sequence of steps in our proposed framework.

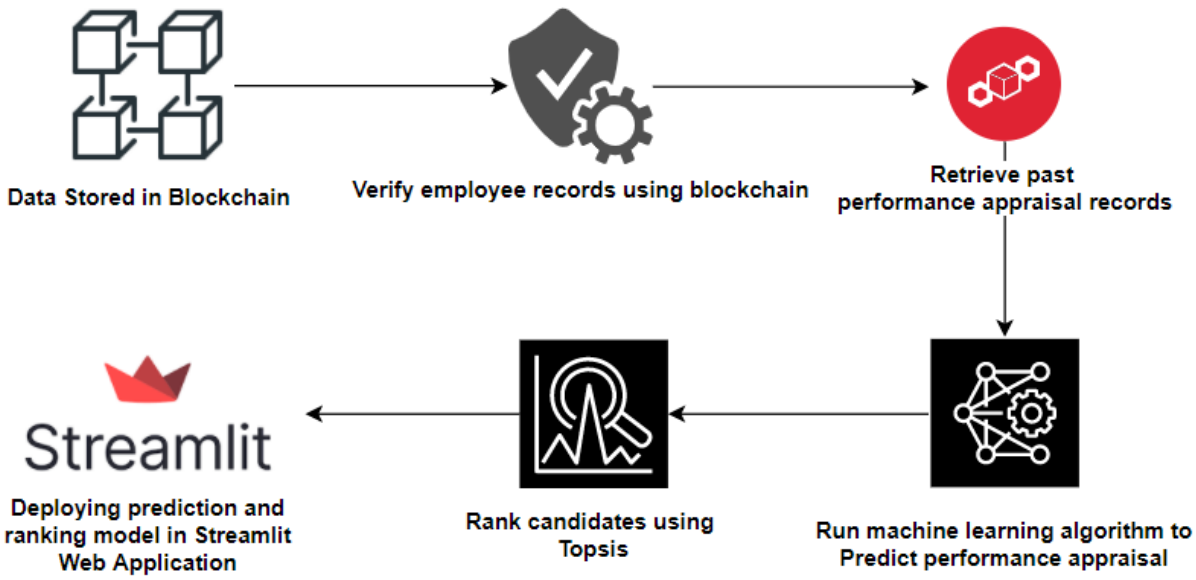


Figure 3.1: Steps in Recruitment Framework

In this section, we will be explaining our research methodology where we discuss the machine learning algorithms for prediction and ranking as well as the blockchain architectures being used in our research.

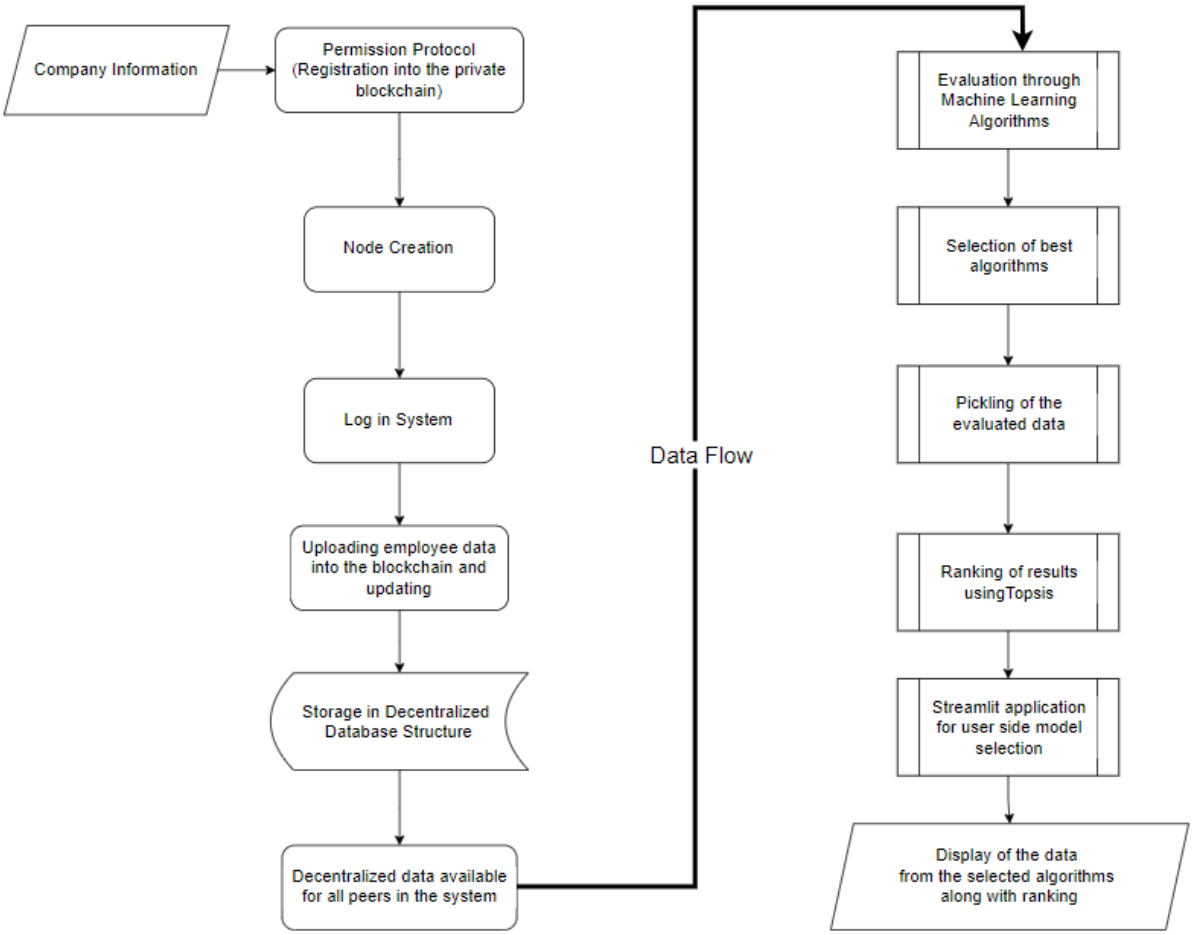


Figure 3.2: Workflow

Figure 3.2 shows a simplification of the workflow to design our system architecture of the complete proposed system. As we can see here the system contains one part concerning blockchain for decentralization and the other part for machine learning and ranking for performance evaluation and ranking.

On the left section, the blockchain architecture methodology is illustrated. Data is first entered into the system after a company gains access as a peer via following the permission protocol. Companies or peers can then update the employee performance dataset accordingly and this will be stored in a decentralized fashion. When the system acquires the information, applicant data can be verified from companies that are part of the consortium. The data can also be retrieved manually by companies authorised by smart contract to use in machine learning models to predict future performance appraisal score of a candidate.

Therefore, we have compared 6 different classifiers, namely Naive Bayes, Neural Network, Logistic Regression, Decision Tree, Random Forest, and XGBoost. After the first evaluation, decision tree and ensemble algorithms are chosen for a later

stage of research. After further feature reduction with recursive feature elimination, we re-ran our models and evaluated their performance. The models were then pickled and made available to the Streamlit application that we developed. Users can select the model they want to run from the options in the user interface of the application to predict performance scores along with selecting the ranking option that ranks the employees based on cumulative score which is calculated using the MCDM model. Therefore, the final output is the list of candidates sorted in a way that ranks the most suitable candidate first.

3.1 Machine Learning Models for Prediction

Research Work [35] has used Logistic Regression, Decision Tree, and Random forest algorithms to make predictions. Our research work took inspiration from their work and added three new algorithms which are Logistic Regression and Neural Network Classifier and XGBoosting to find out the best-performing classification model among them to predict the performance rating of candidates with the highest level of accuracy.

In the following sections we look at different machine learning algorithms we have used in this research.

3.1.1 Logistic Regression

Gladence et al. [8] state that logistic regression is a classifier algorithm that can be used to model or predict a certain class or event. Logistic regression is a well-performing algorithm if the predicted values are discrete in nature [2]. It is very easy to understand and implement and the accuracy of the predicted values can be drastically improved sometimes upon using normalization techniques.

3.1.2 Decision Tree

The Decision Tree is a supervised learning algorithm and is utilized to solve classification and regression problems. A decision tree has a tree structure that is used for classification and prediction. The data is broken down into smaller subsets of data and arranged like a tree, with the outcomes of the testing conditions being at the branches. A Decision Tree is very useful when there can be multiple courses of action for one particular case [55].

3.1.3 Random Forest

Random Forest is a prediction algorithm that uses the concept of bagging - a method of constructing classifier compositions that are trained independently of each other. As a result, multiple trees are constructed and only a fixed number of features from the training set is used and each leaf node of the tree contains observation of only one class. The working architecture for random forest id shown in figure 3.3. In

classification problems, the final decision comes down to a majority vote [30].

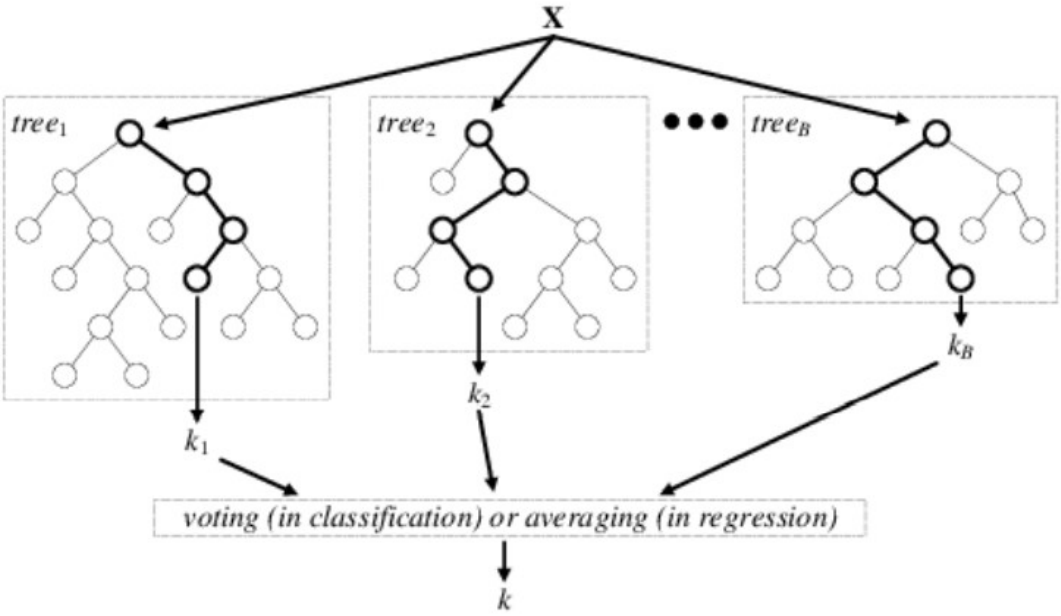


Figure 3.3: Random-Forst Architecture

3.1.4 Naive Bayes

This model is another useful tool to predict values. It uses Bayes theorem and assumes that there are no correlations between the predictors and that the predictors are independent in nature. The dataset used in this case has mostly independent predictors. So, the Naive Bayes Classifier is an excellent tool to use to predict the performance rating [31].

3.1.5 Neural Network Classifier

Zhang [3], states that Neural Network has become an important tool for classification. According to the same source, Neural Networks can adjust themselves according to the data as they are data-driven methods. Neural Networks are also non-linear models and this makes them very flexible in modeling complex data and relationships. Neural Networks are also able to estimate posterior probabilities, which enables them to make classification rules and perform statistical analysis.

$$P(Q|X1...Xn) = \frac{P(Q)|P(X1...Xn|Q)}{P(X1...Xn)}$$

3.1.6 XGBoost

After viewing the results from the first analysis we also included a comparison with the XGBoosting algorithm. According to research [56] tree boosting is highly effective and widely used in machine learning. Chen and Guestrin (2016) proposed a novel sparsity-aware algorithm for sparse datasets and a weighted quantile sketch for approximate tree learning. In their research [56], they described XGBoost as a scalable machine learning system for tree boosting. XGBoost, which in full form is called extreme gradient boosting, is an ensemble machine learning algorithm that makes use of gradient boosting structure to predict and determine values.

XGBoost and Gradient Boosting (GBM) work in similar ways. However, XGBoost takes it to a whole new level as it improves upon the GBM architecture through various optimizations and algorithmic enhancements. Some of the most prominent system optimizations include parallelization, tree pruning, and hardware optimization and some of the algorithmic enhancements include regularization, sparsity awareness, weighted quantile sketch, and cross-validation. Fig 3.4 shows how Gradient Boosting works.

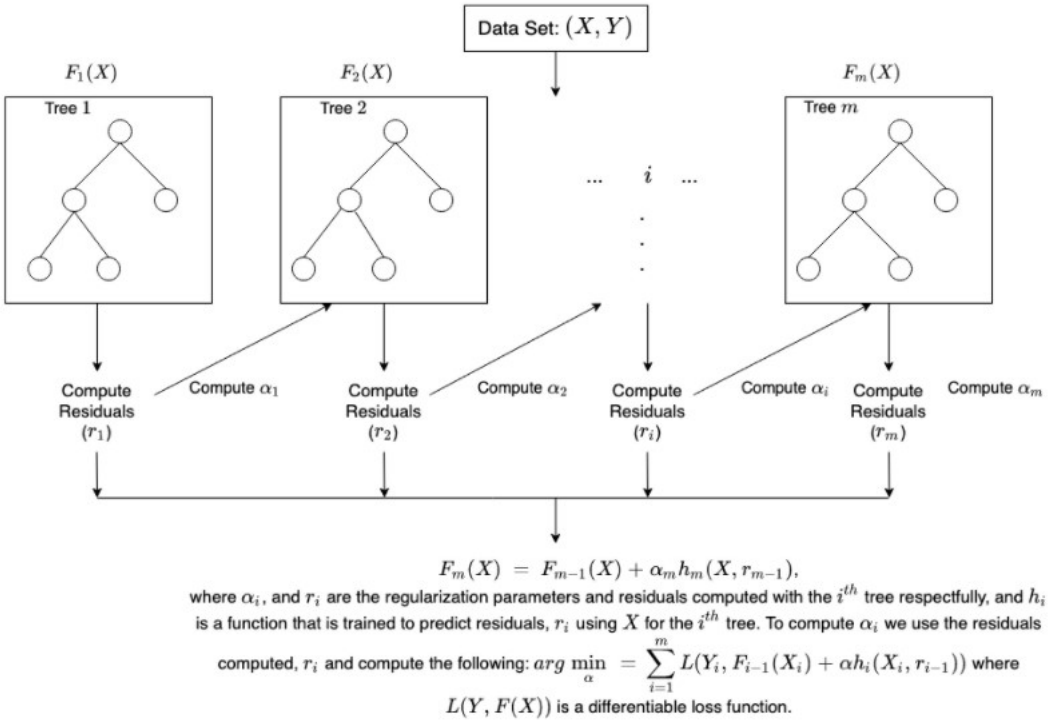


Figure 3.4: Gradient Boosting Architecture

3.2 Rating and Ranking Candidates

3.2.1 MCDM

MCDM is a multi-criteria decision-making algorithm that takes into account the multiple criteria which can affect the result of a decision. Alrababah et al. [13] stated that Multi-Criteria Decision Making can consider the different weights of criteria altogether at a time. As decision-making often involves imprecision and vagueness in terms of the certainty of the results, the MCDM can be seen as a way of effectively coming up with a more reliable and logical conclusion. MCDM techniques are specifically suitable for implementation in circumstances where selection is required from a pool of options or alternatives and figure 3.5 shows the steps in MCDM implementation. Since our architecture specifically requires the effective selection of employees on the basis of multiple performance metrics such as 'Age', 'DistanceFromHome', and 'EducationLevel'. Therefore it is crucial to implement MCDM techniques in this solution. Placing the human resource according to the competence of the employee is the priority of any company and hence to accommodate a more fair and just process that is also efficient, it is crucial to utilize the algorithm that takes into account the ability to compare and find out the best employee from a pool of applicants.

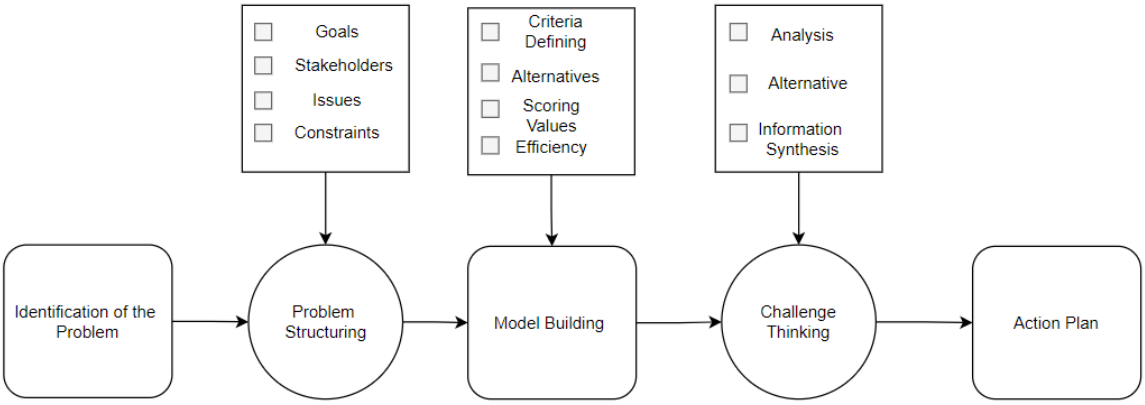


Figure 3.5: Steps in the MCDM methods

3.2.2 TOPSIS

TOPSIS (a technique for order preference by similarity to an ideal solution) was presented by Hwang and Yoon [1]. The TOPSIS method is our preferred algorithm because of the number of steps in the algorithm being limited to seven irrespective of the number of criteria or the situation. Fuzzy algorithms make the case by considering a positive best value (positive ideal) in accordance with all the criteria and a negative best value (negative ideal). The algorithm's base logic lies in the fact that the algorithm takes into consideration the distance from the positive ideal and the negative ideal whereby the alternative farthest to the negative ideal and closest to the positive ideal is taken as the best alternative and the opposite as the least preferred alternative [6]. For making the case of identifying the importance of the criteria, different weightage is given according to the relative importance of the criteria. In our system, the weights are determined by the interviewer or the person in charge of the interviewing process. Figure 3.6 below shows a conceptual diagram of TOPSIS while figure 3.7 shows the steps in the TOPSIS algorithm.

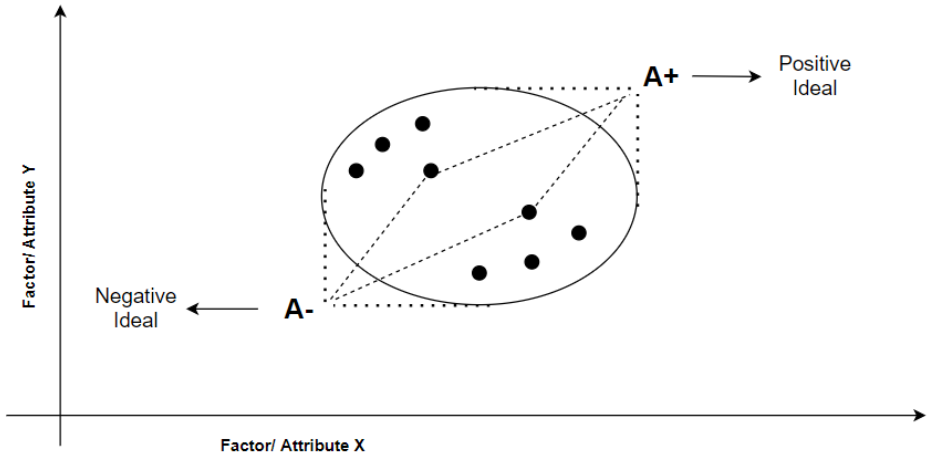


Figure 3.6: Concept diagram showing the positive and negative ideals as a result of multiple factors

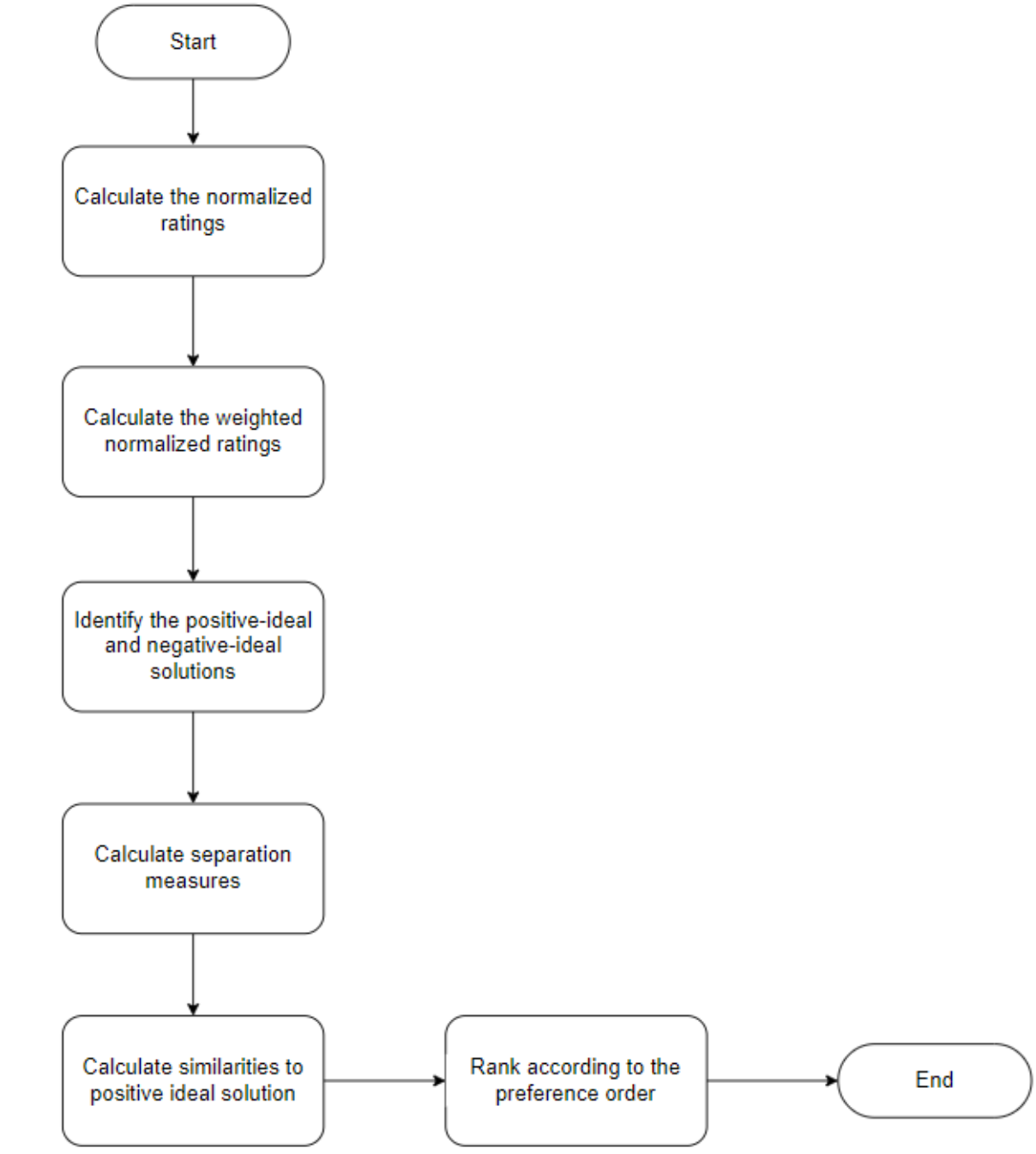


Figure 3.7: Steps in the TOPSIS method

3.3 Blockchain Architecture

3.3.1 IOTA

IOTA is a form of ledger open for anyone to use which is designed for the Internet of things (IoT). A directed acyclic graph (DAG) is used to store transactions in its ledger. According to [37], IOTA is a cryptocurrency but not developed in the blockchain. IOTA uses a series of cryptographic algorithms for verification in the form of hashes which increases the security of transactions and prevents fraudulent transactions from taking place. "Tangling" is introduced to minimize the cost of IOTA. Unlike Blockchain, IOTA is open source and only gets faster with more transactions. Moreover, IOTA does not require the high computational power that traditional blockchains such as etherium require as IOTA does not require any miners.

3.3.2 Ethereum

According to research [56], Ethereum is a blockchain that uses a peer-to-peer network in order to establish the connection. It makes use of smart contracts to aid in its operations and transactions. Private Enterprise Entereum networks consist of actors and the smart contracts will dictate the interactions and transactions inside the network accordingly.

3.3.3 Hyperledger Fabric

According to research [36], Hyperledger Fabric is a permissioned blockchain network hosted by Linux Foundation. Fabric has a lot of functionalities such as peers, chain code, ordering service, and state database. In order for any transaction to occur within the network, a consensus protocol is required and as it is a permissioned blockchain network, in order for an actor to be a part of the network, it needs to have a membership. It is extremely versatile as it supports a wide range of industry use-cases.

3.3.4 Proposed Blockchain Methodology

As we have already seen that there are some shortcomings of IOTA for which we cannot choose the system for our blockchain feature. This is due to the continuous data pruning in tangle causing the data holding capacity to be very short-lived. Furthermore, since the system is not yet robust at the time of writing this paper, we cannot rely on the system to do all of our desired functions properly as it is subject to high levels of change at this phase. Since we want a system where there is complete accountability of every action or transaction, IOTA does not serve this purpose well too. Finally, the difficulty of developing smart contracts or chain codes in an IOTA system coupled with additional hassles in the deployment of a decentralized app makes the IOTA system less desirable.

The use of Ethereum, although more complete, is not recommended for this research. This is due to the architecture of Ethereum being mainly a complex blockchain architecture that is used for end-to-end systems. Our system should suffice with a

more specific private enterprise-level blockchain.

Hence we took inspiration from the research blockchain work Hyperledger Fabric and developed a blockchain simulation tool that replicates the functions of a blockchain built in python. The advantage of building a simulation tool from the bottom up is that it helps understand the architecture of a private enterprise blockchain system better with deep insights in terms of processing and data decentralization alongside understanding the on-chain and off-chain performance issues of blockchain to greater depths.

This blockchain simulation achieves a decentralisation and immutability of transaction records in a private permissioned blockchain to ensure accountability and transparency, therefore creating a secure way of data sharing among enterprises in the consortium to allow data verification and retrieval for usage in machine learning algorithms for performance appraisal prediction.

Chapter 4

Dataset & Pre-Processing

4.1 Data Collection

When it comes to datasets from human resources, it is very hard to find human resource data from actual organizations due to privacy concerns. Therefore, we have used the IBM HR Analytics Employee Attrition and Performance dataset which was developed by IBM data engineers. Hewage et al. [35] used the same IBM dataset for a similar purpose of predicting employee performance. The dataset used to run and implement the machine learning models was appropriated from the website, "Kaggle". This dataset has 1470 rows and 35 columns and consists of several useful features to measure performance rating related to both the employee such as their age, education field, and education level and their performance-related data such as the number of projects worked, days since the last promotion and performance rationing. The dataset can be downloaded from the link:
https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

In Kaggle, a list shown in the following figure was given for the meaning of encoded values for a few features. Since our dataset did not have an employee ID, we created a column for 'Employee_ID' to label each row.

Table 4.1: Information of raw dataset.

Selected Features	Data Type
Age	int64
Attrition	object
BusinessTravel	object
DailyRate	int64
Department	object
DistanceFromHome	int64
Education	int64
EducationField	object
EmployeeCount	int64
EmployeeNumber	int64
EnvironmentSatisfaction	int64
Gender	object
HourlyRate	int64
JobInvolvement	int64
JobLevel	int64
JobRole	object
JobSatisfaction	int64
MaritalStatus	object
MonthlyIncome	int64
MonthlyRate	int64
NumCompaniesWorked	int64
Over18	object
OverTime	object
PercentSalaryHike	int64
PerformanceRating	int64
RelationshipSatisfaction	int64
StandardHours	int64
StockOptionLevel	int64
TotalWorkingYears	int64
TrainingTimesLastYear	int64
WorkLifeBalance	int64
YearsAtCompany	int64
YearsInCurrentRole	int64
YearsSinceLastPromotion	int64
YearsWithCurrManager	int64

Table 4.1 shows a the features in our used dataset.

After checking we found our dataset was free of null values. Moreover we have also found the details of each column in our dataset.

4.2 Data Analysis

Data pre-processing is used to transform raw data in a useful format to be used in data analysis. Among our 36 columns, we realized there were a few columns that were not as important. We found these columns through finding a correlation between the features in our dataset by generating a correlation matrix and showing correlation using a heatmap.

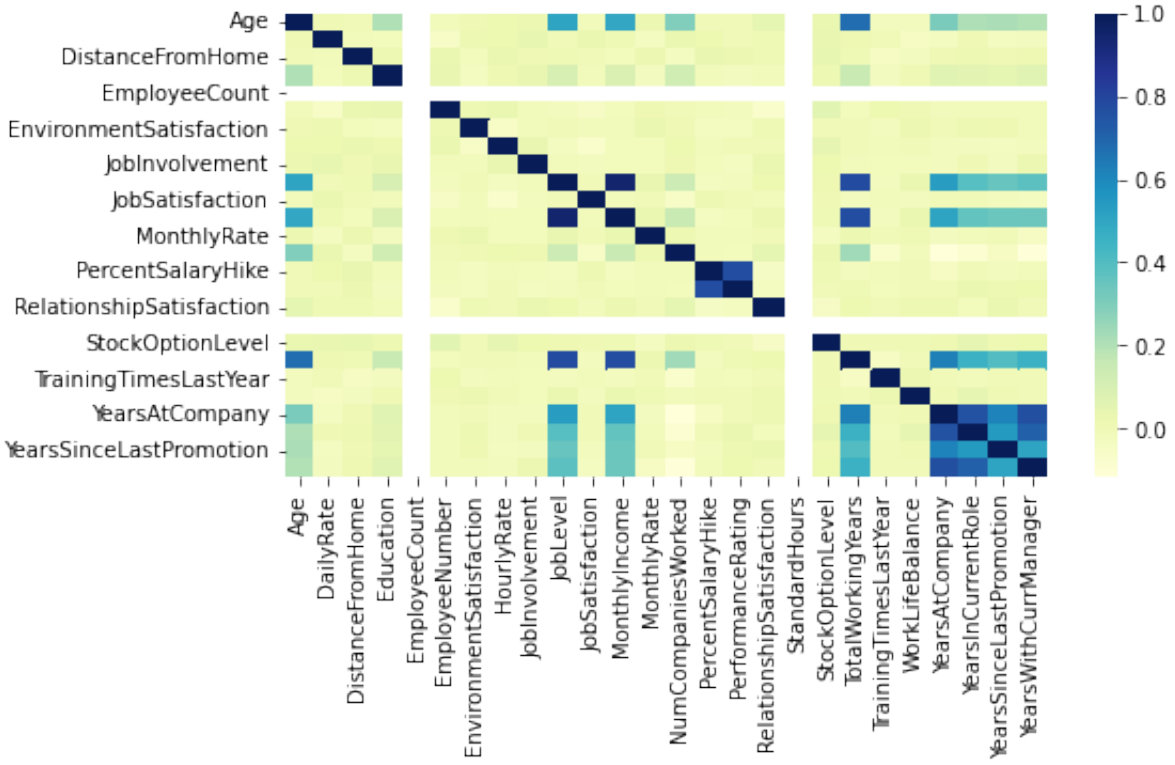


Figure 4.1: Correlation Matrix

Figure 4.1 shows PerformanceRating has the highest correlation with PercentSalary-Hike, while having low correlation values with most of the other features. Hence in the next stage, we dropped all the columns having zero correlation or a very small level of correlation. Since the value ranges for each feature are different, we have created a box-plot to evaluate the data, even more, however, we have omitted 'MonthlyRate' and 'MonthlyIncome' since the values are much greater than other features and continuous. Below figure 4.2 shows the box-plot.

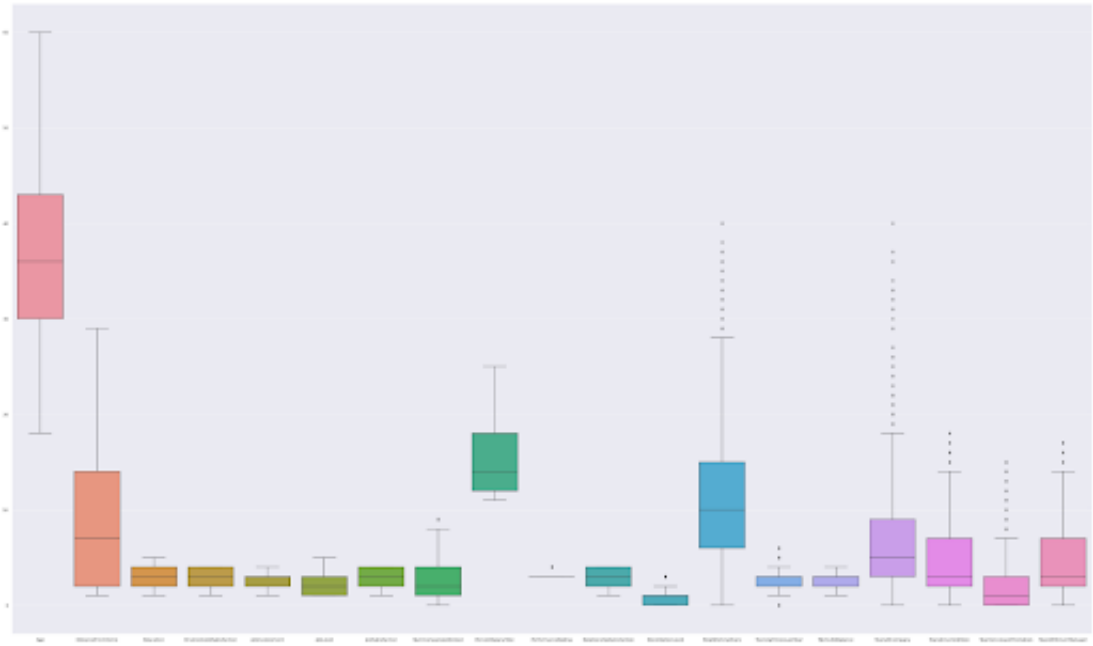


Figure 4.2: Box-Plot

On the other hand, our dataset has been pre-encoded for certain features previously which are ‘WorkLifeBalance’, ‘RelationshipSatisfaction’, ‘PerformanceRating’, ‘Education’, ‘EnvironmentSatisfaction’, ‘JobInvolvement’ and ‘JobSatisfaction’. The encoded value description is given in Kaggle and the screenshot is shown in figure 4.3.

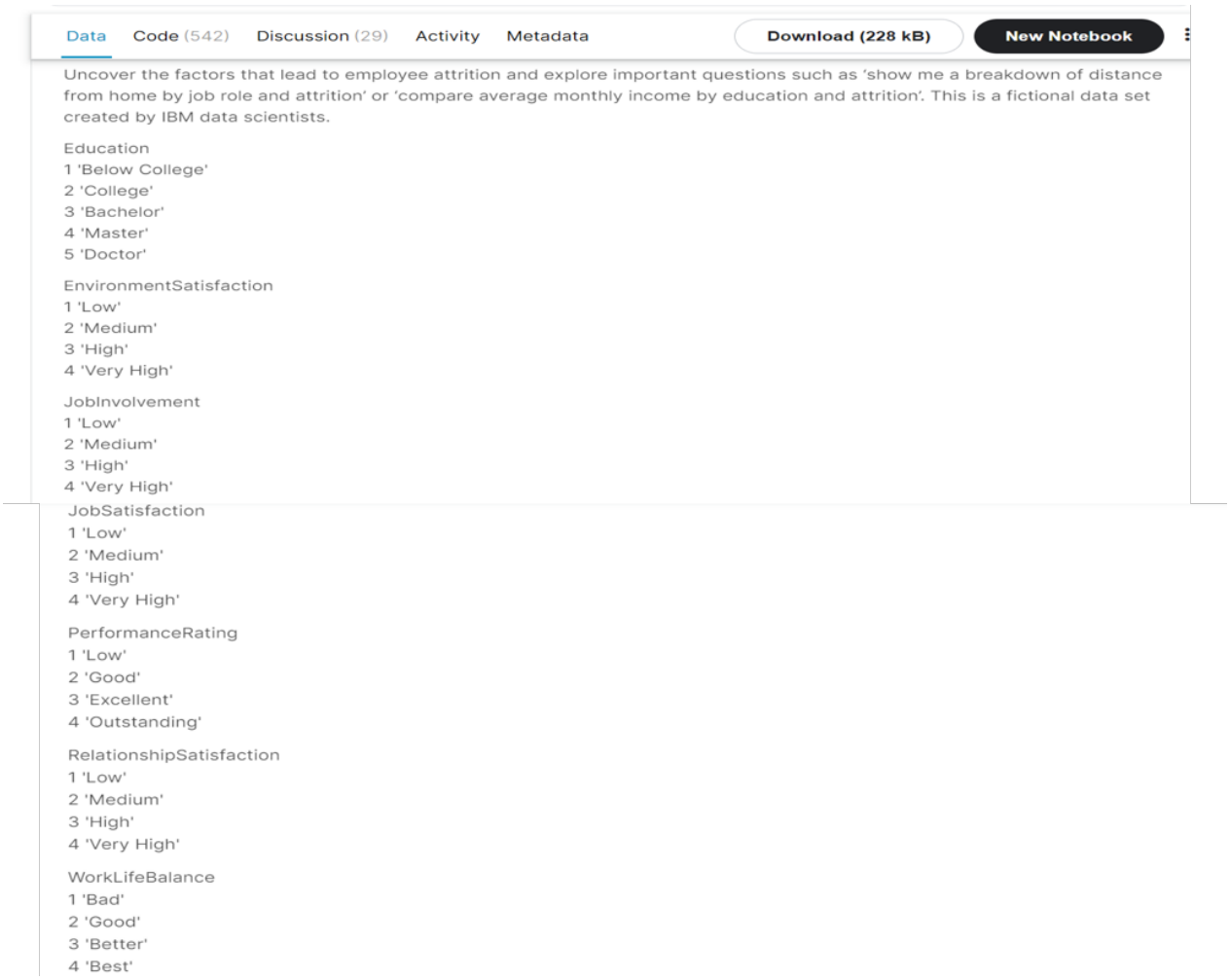


Figure 4.3: List of encoded data in Kaggle for IBM HRM Attrition data-set

Apart from these features, there are other features such as 'EducationField', 'JobRole', 'MaritalStatus' and 'Gender' that have string values. We have found the unique string values held by each feature to analyze our dataset even further to help us encode them for Prediction and Ranking. Table 4.2 shows the unique values in each feature found using .unique().

Table 4.2: List of Unique vales for Features with String Data Type

Feature	Unique Values
BusinessTravel	Travel_Rarely Travel_Frequently None_Travel
Department	Sales, Research & Development, Human Resources
EducationField	Life Sciences, Medical, Marketing, Technical Degree, Human Resources, Other
JobRole	Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources
OverTime	Yes, No

4.3 Feature Selection

Feature selection is an essential part of data pre-processing as it enables the machine learning algorithms to train faster, decreases complexity of dataset, and improves accuracy of model. In our initial model comparison, we only removed featured that were redundant or made our system biased. After first comparison we selected the best performing models and ran those with recursive feature elimination to select the best features.

4.3.1 Initial Feature Selection

Initially, we dropped eight features that seemed redundant or would make our system biased and are listed below:

- **StandardHours** was dropped because it had NA values.
- **EmployeeCount** was dropped because it had NA values.
- **Over18** was dropped because it had only one unique value.
- **MaritalStatus** was dropped because to prevent biasedness.
- **Gender** was dropped because to prevent biasness.
- **Attrition** was dropped because it was irrelavant to our usecase.
- **EmployeeNumber** was dropped because the feature did not add any value to research.
- **EmployeeID** was placed as index and then dropped as a separate column.

4.3.2 Recursive Feature Elimination

According to [9], RFE improves the accuracy of Decision Tree, therefore we chose this algorithm for feature selection. RFE is a feature selection model that removes all the weakest features in a data set after fitting a model. It employs filter-based feature selection internally as well as wrapper-style feature selection. RFE seeks to reduce dependencies and collinearity in the model by iteratively deleting a limited number of features in every loop.

According to [9] works by first training the classification model on the dataset, ranking features by importance, discarding the least important features, and fitting the model again. This procedure is repeated until only a certain number of features are left. In RFE the Gini importance ranking technique is used to eliminate features. The following is the pseudocode for RFE.

Algorithm 1 The Pseudo Code of the RFE Algorithm Is Given Below

Input:
Training set T,
Set of p features F = f1,...,fp
Ranking Method RM(T,F)
Output:
Final ranking R
Code:
Repeat for i in (1:p)
    Rank set F using RM(T,F)
    f ← last ranked feature in F
    R(p i + 1) ← f
    F ← Ff

4.4 Data Pre-Processing

We pre-processed our data twice, once for performance prediction and a second time for applying a Multi-Criteria Decision Making Model.

4.4.1 Prediction

For performance prediction, our dependent variable is 'PerformanceRating' in our dataset which is performance scores given to each employee. According to Kaggle, the Performance score has been given under four criteria and a certain score is given to each which are 1 for 'Low', 2 for 'Good', 3 for 'Excellent', and 4 for 'Outstanding'. However, in our dataset employees are rated only 3 or 4. Therefore to get a closer look at the relationship of performance rating with each feature we created a Histogram of each feature depending on 'PerformanceRating' and got the following results.

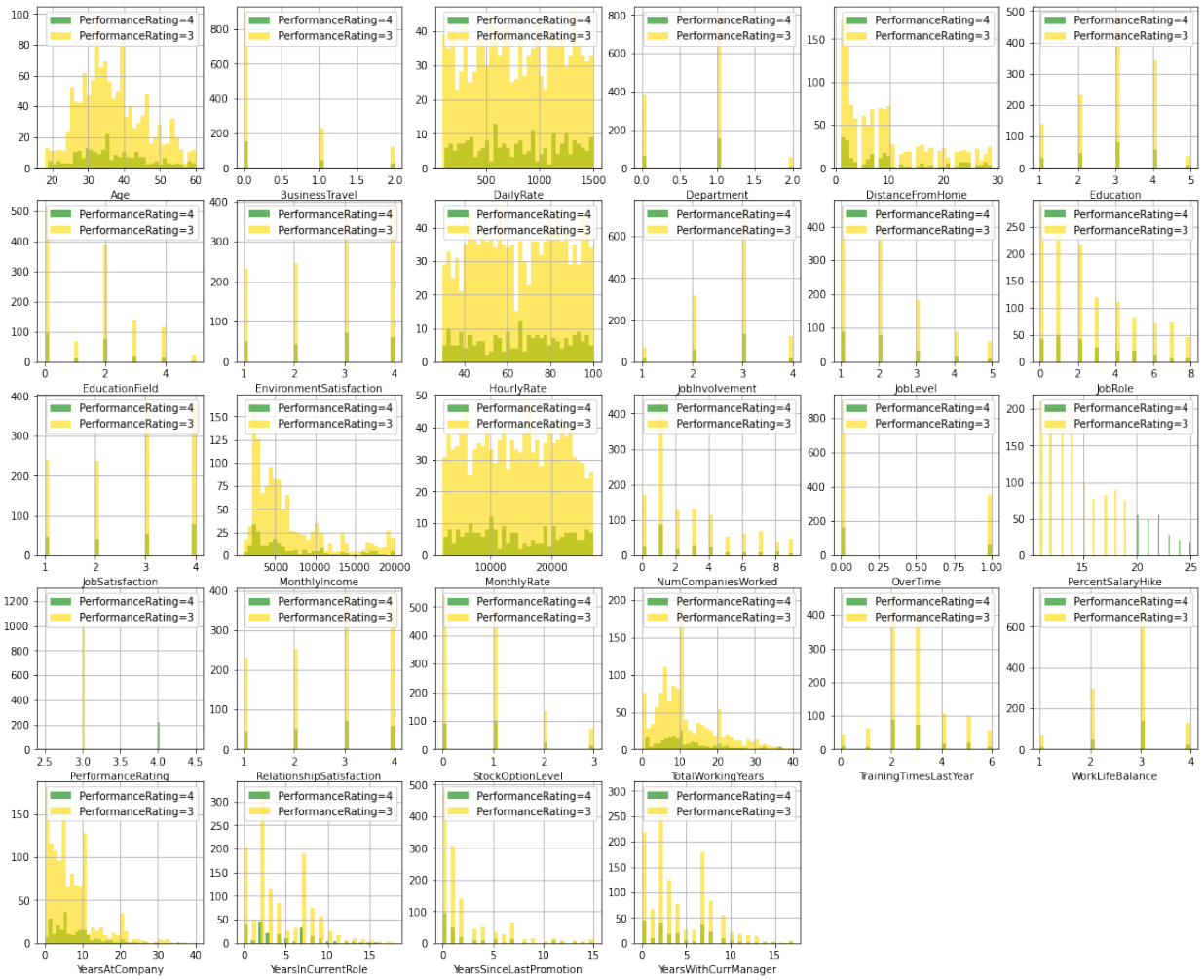


Figure 4.4: Histogram of each feature depending on 'PerformanceRating'

Label Encoding

Since prediction is value-based, the unique values we found for all features in table 4.2 need to be encoded to strings. Thus, we will perform label encoding by turning all the string values for features in table 4.2 into int by assigning numbers. Therefore, table 4.3 shows the features included in our cleaned dataframe on which we will run our machine learning models. The table also shows the data types of the features after label encoding and feature selection. This dataframe will be used for the initial stage of machine learning model evaluation.

Table 4.3: Information of cleaned dataset.

Selected Features	Data Type
Age	int64
BusinessTravel	int64
DailyRate	int64
Department	int64
DistanceFromHome	int64
Education	int64
EducationField	int64
EnvironmentSatisfaction	int64
HourlyRate	int64
JobInvolvement	int64
JobLevel	int64
JobRole	int64
JobSatisfaction	int64
MonthlyIncome	int64
MonthlyRate	int64
NumCompaniesWorked	int64
OverTime	int64
PercentSalaryHike	int64
PerformanceRating	int64
RelationshipSatisfaction	int64
StockOptionLevel	int64
TotalWorkingYears	int64
TrainingTimesLastYear	int64
WorkLifeBalance	int64
YearsAtCompany	int64
YearsInCurrentRole	int64
YearsSinceLastPromotion	int64
YearsWithCurrManager	int64

Train-Test Split The dataset is split into two parts, with 70% of the data being allocated for training purposes and 30% of it for testing purposes. The 'y' values consist of the 'PerformanceRating' and the 'x' values are everything else. Using regression models, and by training the datasets, the predicted values of 'y' are found out using 'x'.

Data Pre-Processing for XGBoost

For XGBoost, we had to pre-process our data differently. In XGBoost one hot coding is required. Therefore, to make a faultless learning model, we will re-encode the pre existing encoded features back to original string values.

One-Hot Encoding

The categorical data is transformed into a binary vector representation using one-hot encoding and is done using the pandas library. One-hot encoding creates a new

column for each unique value in a column and if that particular data is present in a row, the value '1' is shown under that column for that row. Table 4.4 shows the list of features encoded using one-hot encoding.

Table 4.4: One-Hot encoded features.

One-Hot Encoded Features
BusinessTravel
Department
Education
EducationField
EnvironmentSatisfaction
JobRole
JobInvolvement
JobSatisfaction
RelationshipSatisfaction
WorkLifeBalance
OverTime

4.4.2 Rating

For TOPSIS our data has been processed to solely calculate scores of certain factors which will allow us to rank the candidates. The dataset that we have contains values like distance from home and age. These values cannot be fed directly into the algorithm for topsis. Feeding these values directly into the dataset will not result in a fair evaluation since the distance magnitude will be dominating the further steps. Hence we have converted these values into numerical ranges using bins. According to our method, for the Distance From Home feature, we have set the range from 0-5, where the employee who lives the closest gets a higher score. Here an employee staying within a range of 0-5km gets a score of 5, whereas an employee residing within 5 to 10 km is assigned a score of 4. Furthermore, the rest of the values go through similar modeling. On the other hand, for Age, the range values are put in the fashion that the lower the age the higher the value, from a score range of 4 to 1. Since according to our evaluation criteria younger employees are to be larger assets due to their higher working capacity. The ranges set for Age and DistanceFromHome are shown in Table 4.5 and 4.6.

Table 4.5: Setting Range for Age

Age in years	Score
18-30	5
30-40	4
40-50	3
50-60	2
60 and above	1

Table 4.6: Setting Range for DistanceFromHome

Distance in km	Score
0-5	5
5-10	4
10-15	3
15-20	2
20-25	1
25 and above	0

Next, we figure out the features on which evaluation can be done in an unbiased way. Therefore, we dropped several features and chose only 10 features to use our TOPSIS algorithm to create a score. Apart from creating ranges for Age and DistanceFromHome, we did label encoding, for rest features that had object datatype, using integer values. The final input data frame is shown in table 4.7.

Table 4.7: Information of data frame used for TOPSIS model

Selected Features	Data Type
Age	int64
DistanceFromHome	int64
Education	int64
JobInvolvement	int64
JobLevel	int64
NumCompaniesWorked	int64
OverTime	int64
PerformanceRating	int64
TotalWorkingYears	int64
TrainingTimesLastYear	int64

4.5 Data Scaling

Feature scaling is one of the most common steps in data processing algorithms. According to research [27], the magnitude of features in a dataset may vary by a huge margin. As a result, the features have to be scaled so that the features are closer to one another in magnitude.

4.5.1 Min-Max Scaler

Minmax scaler is a feature scaling technique. According to the research paper [27], minmax scaler scales the features within a fixed range, either from -1 to 1 or from 0 to 1. This takes minimum and maximum sample as m_max and m_min. The formula is given below:

$$\frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.1}$$

4.5.2 Standard Scaler

Similar to minmax scaler, standard scaler also scales dataset features to an acceptable range. According to research [42], standard scaler expects the information to be ordinarily appropriated inside each component. According to the same paper, it will scale them to such an extent that the dissemination revolves around 0, and standard deviation becomes 1. The mean and standard are calculated and later it is scaled using the following formula:

$$X_{scaled} = \frac{X - mean}{StandardDeviation} \tag{4.2}$$

Chapter 5

Blockchain Architecture Implementation

The blockchain architecture aims to create a permissioned, decentralised and immutable system to store employee records from different companies that are part of the consortium, thus allowing the companies to share data for verification purposes and retrieval for further steps of this framework.

Due to the restrictions found in IOTA and Ethereum, we are no longer using them. Instead, now we aim to work with a prototype permissioned enterprise blockchain simulation developed in python. This architecture is heavily inspired by the working methodology of hyperledger fabric and we developed it using python to make it more adjustable with our system. The structure of this blockchain system is described below.

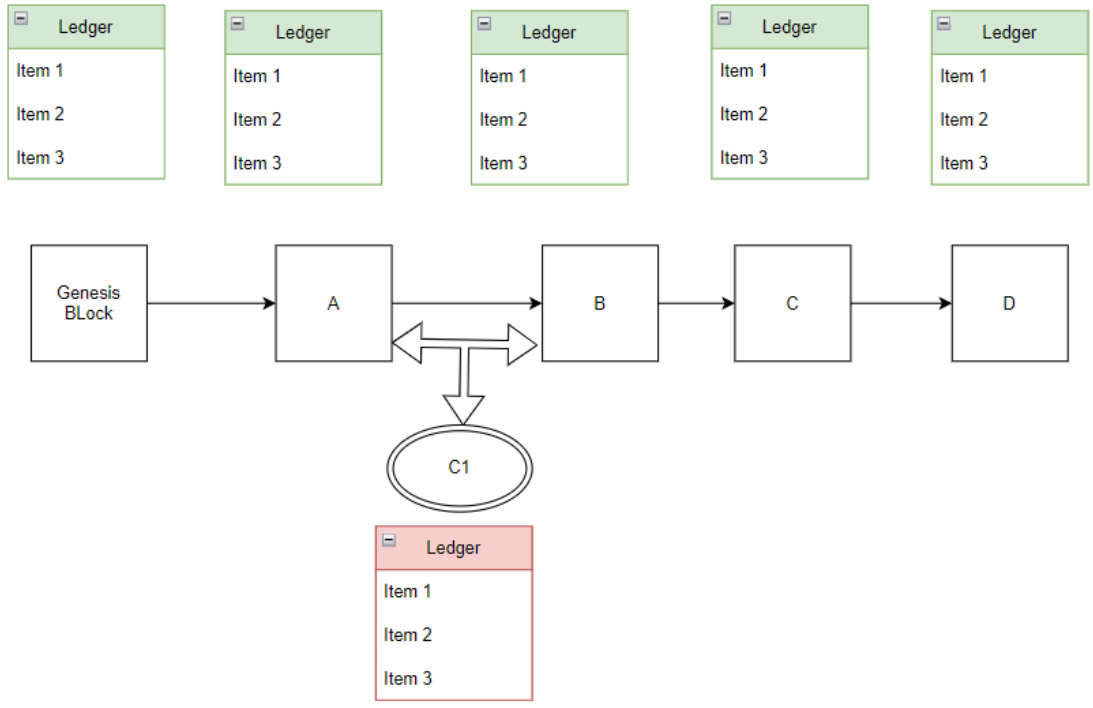


Figure 5.1: The basic structure of the system and private channel

This is the fundamental structure of the blockchain. As we are replicating the model of hyperledger fabric we aim to make an almost similar structure that works as the method depicted in Figure 5.1. As we know, according to [40], hyperledger fabric uses a modular design to create a neutral space for collaboration, our system is also one which replicates this ability. In Figure 5.1 we can see the main system channel which is a legacy process in hyperledger fabric as of now. This means that it is not absolutely necessary to have a legacy channel. We can see that the genesis block precedes several other nodes creating the network. There is a scope for a private channel whereby two nodes can be connected to form this channel. When a private channel is created just as shown in the diagram between A and B, the data can be shared between A and B and cannot be accessed by the other nodes C and D as they are not members of the private channel. The advantage of the private channel is that special contracts can be curated specifically to the needs of the clients in the channel which can be different from the ones in the other channels.

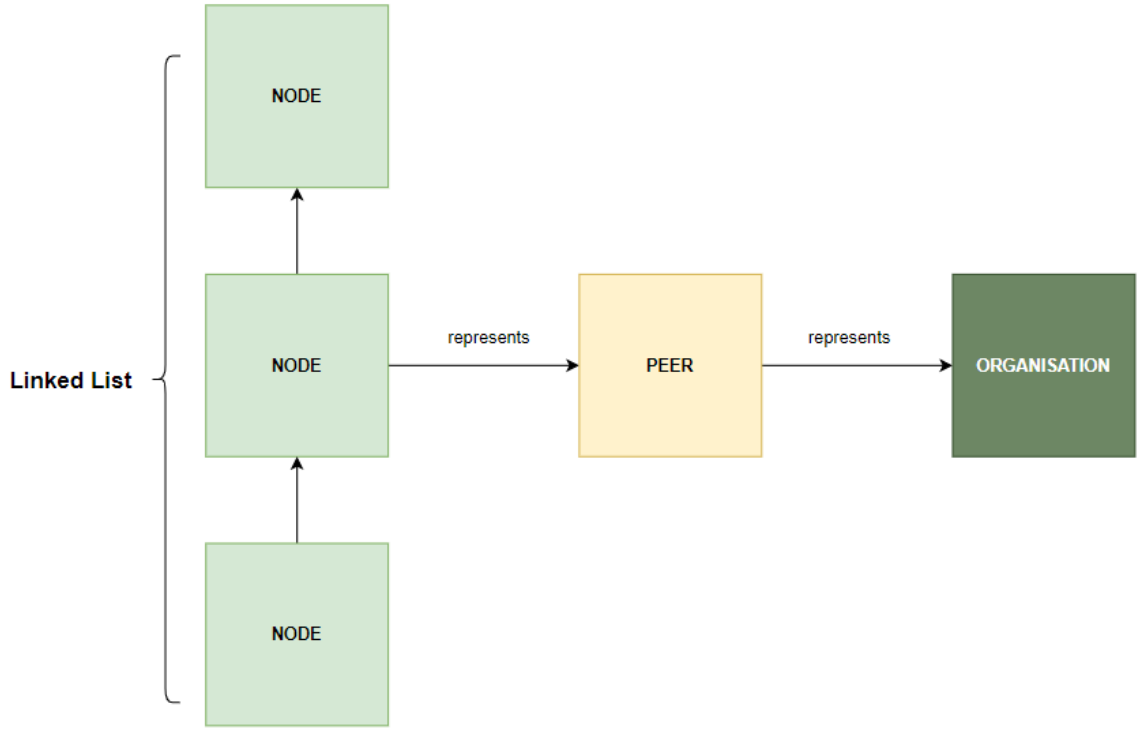


Figure 5.2: Connection between nodes, peers and organization

Figure 5.2 is a simplified conceptualization of our proposed architecture and shows an overall structure of the proposed architecture. Upon close observation, we can see that the system has one linked list as the basic system channel which will be making up the consortium of organizations and companies that will be joining us. We will be allowing these organization HR representatives to join as our peers who will be therefore representing their companies respectively. Any activity by these peers will be documented in the transaction ledger as the activity of the company represented by that peer in the system. Therefore, it is crucial for the companies to assign a

decision-maker as the account owner. Additionally, multilateral transactions can also be possible between the peer companies via multiple private channels which is an important feature of enterprise blockchain according to [48].

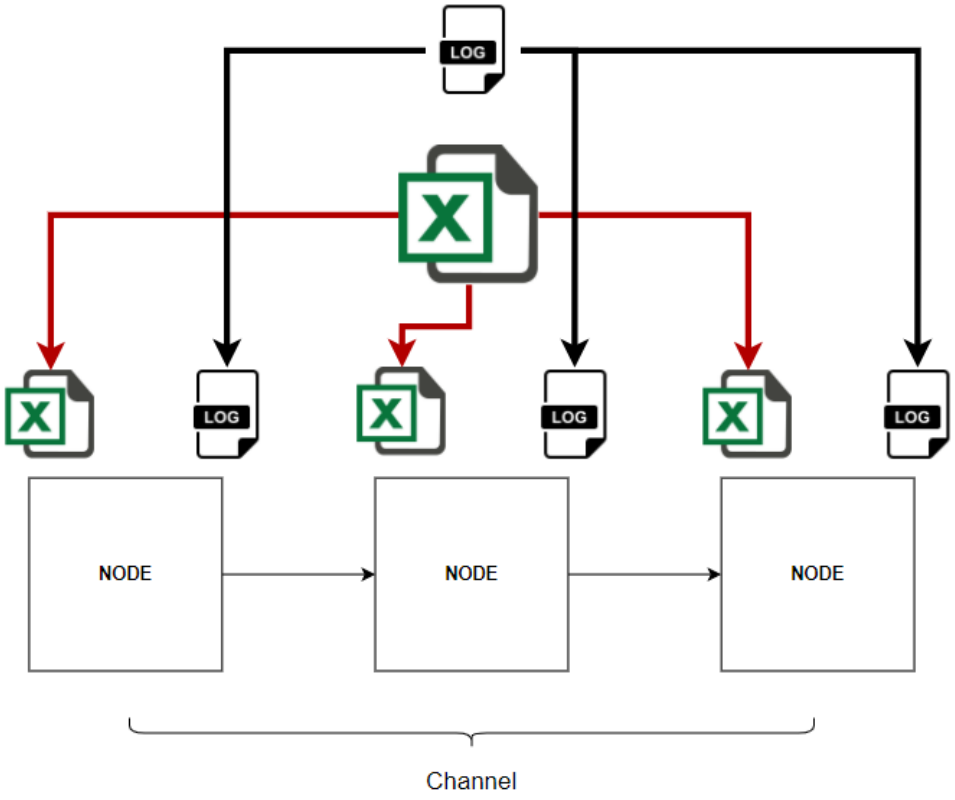


Figure 5.3: Displaying the decentralized distribution of ledgers in the system.

Figure 5.3 shows the file structure after the blocks are created in the blockchain. Each node is assigned to a folder and each node has a copy of the main ledger and the transaction ledger, hence making the decentralised system.

5.1 Permission Protocol to Register a Peer

To register a new company, that is a new peer into the permissioned blockchain consortium, we have created a permissioned protocol using the following methodologies.

5.1.1 Certificate Authority

The certificate authority is the body that will be in charge of generating the digital certificates or signatures that will be then sent to the MSP or Membership service Provider to be verified and then based on that the organization can be allowed in the system channel. The certificate authority in our system is in charge of generating certification IDs which are sent to the MSP for verification upon which the MSP decides whether to allow a peer into the system or not.

5.1.2 Membership Service provider

The membership provider will be responsible for allowing the organization into the consortium. The MSP will be also responsible for overlooking other responsibilities such as the creation of a revocation list for instances where the signature certificate provided by the CA is in any case tampered with or updated for any purposes. Therefore it is obvious that the MSP will stay in constant communication with the Certificate Authority. The MSP will be in charge of giving an OTP number to the admin which will then be emailed to the requester. The requester must enter the OTP to register in the system. As stated in [40], the modular architecture of MSP will allow it to be a plug-and-play structure, our code of the MSP is also such that it is modular in nature and can be greatly optimized with further iterations.

5.1.3 AES Encryption

The digital signature certificate that will be generated by the CA will be stored in the MSP as a form of encrypted data. For this purpose, we are suggesting the use of AES(Advanced Encryption Standard) which will prevent the theft of these crucial data. This is to prevent any other fraudulent parties from using the data to pose as a valid participant and entering the system channel or any other channels.

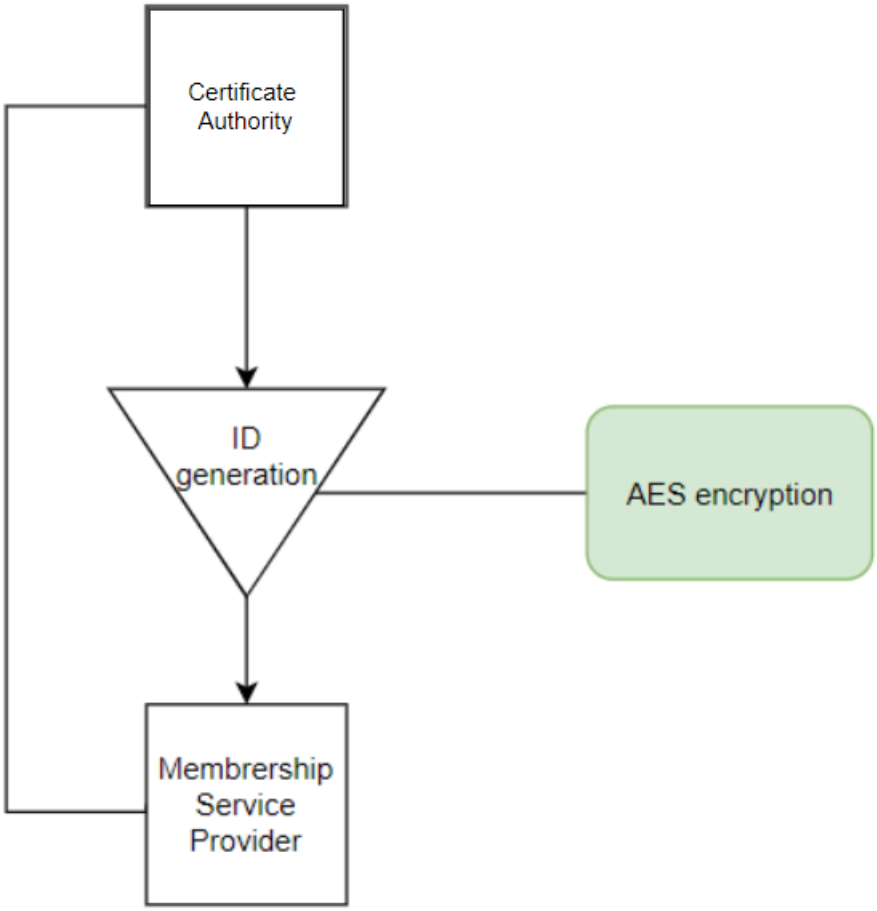


Figure 5.4: Displaying the role of CA and MSP along with AES protocols

The figure 5.4 shows the complete overview of the permission protocol and the order of process of the CA and MSP. It is to be put into special attention that data encryption is crucial in this phase.

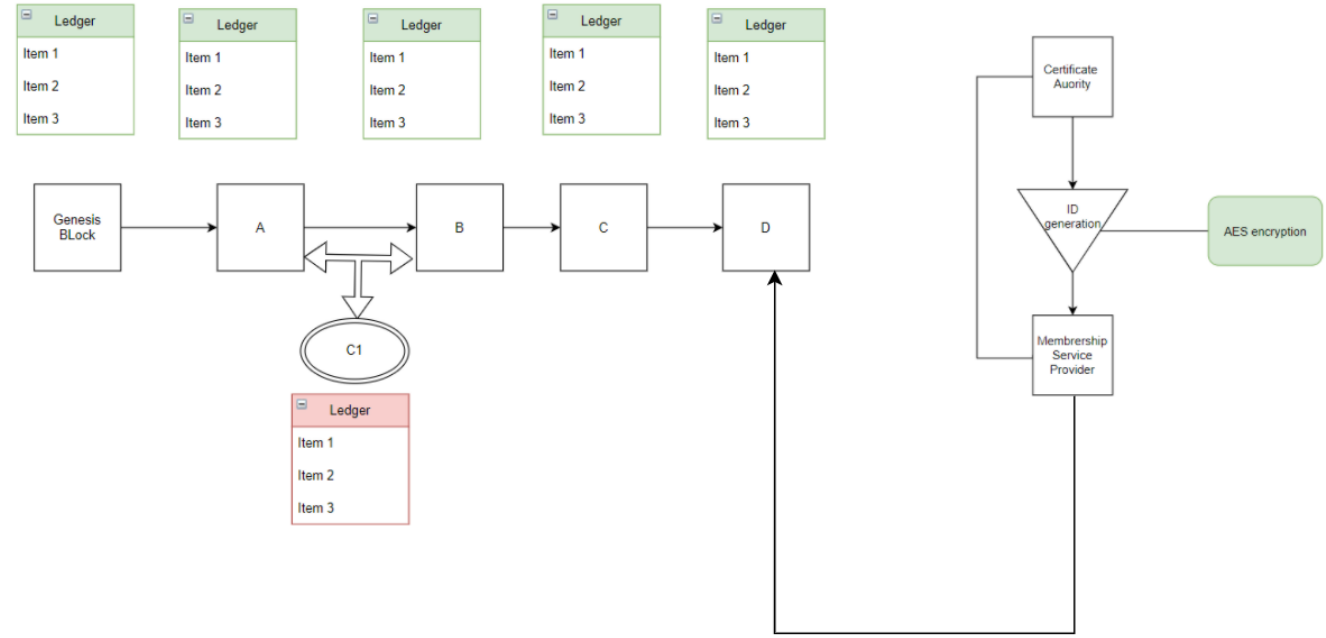


Figure 5.5: Overall structure combined persimmon protocol and structures

Figure 5.5 shows the complete overview of the basic blockchain chain backbone comprising of the permission protocol and the system channel architecture. This shows us that it is absolutely mandatory to be eligible to gain access inside the blockchain.

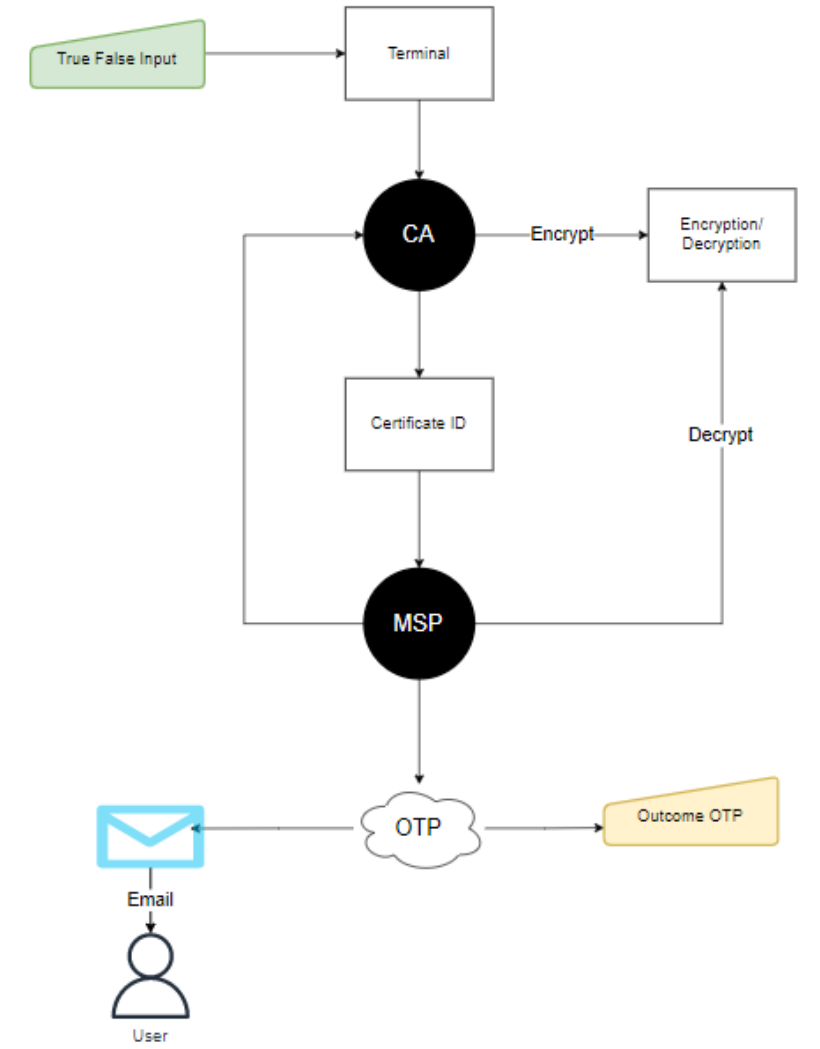


Figure 5.6: The permission protocol

```
Is the data provided valid to process with the certificate ID?True
This is the provided certification ID: 483c
The member validation process was run
```

Figure 5.7: The initial document verification from the admin side

As we can see in the figure 5.6 and 5.7, the system admin of the system first checks all the documents and tries to see if the requirements are matched for the company to be entered in the channel. After the requirements match, the admin allows the system to proceed and send an OTP to the client through email and Whatsapp.

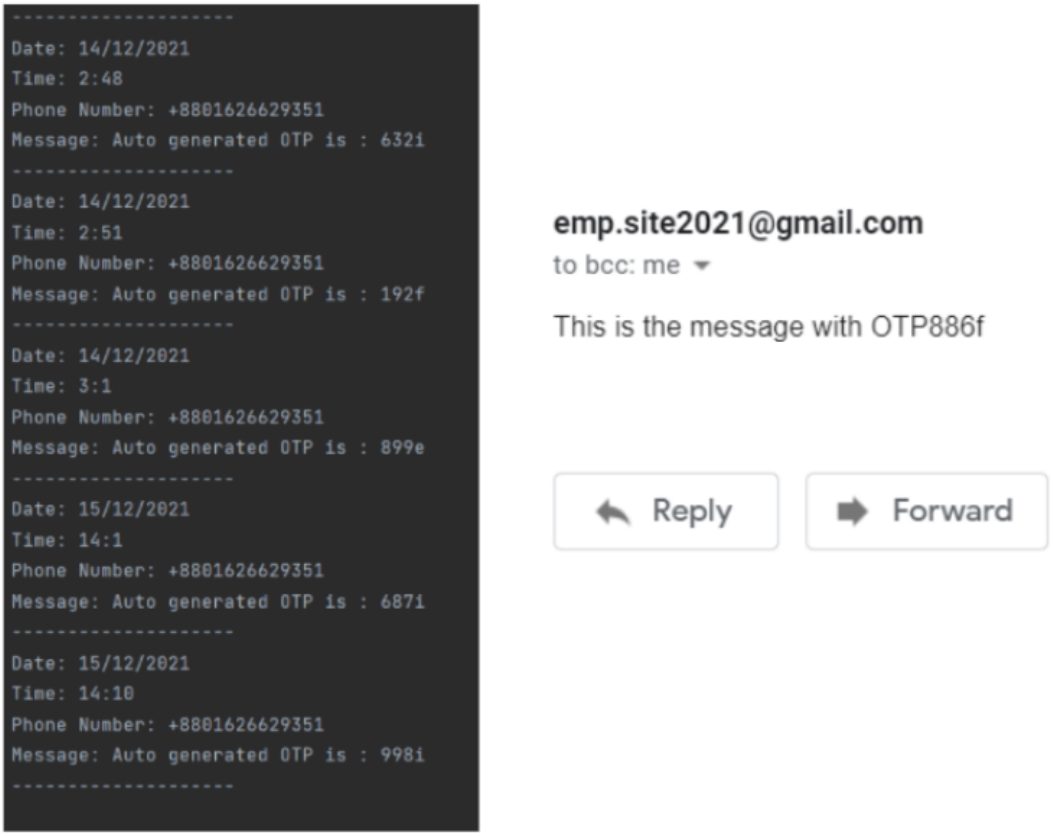


Figure 5.8: OTP being sent via email whatsapp

```
Node is created
socket created
waiting for connection
['Company_A', '886f', 'pass1234']
This is the entered OTP provided by client 886f
client connected with ('127.0.0.1', 10121) Company_A 886f pass1234
This is the client side OTPlist: ['886f']
```

Figure 5.9: Client connection set up and node created.

On the client side, the client receives the OTP and then registers in our system using the OTP provided. On successful registration a node is created in the name of the client. In this case, a node was created in the name of Company A as given by the client. Figure 5.8 and 5.99 shows the results.

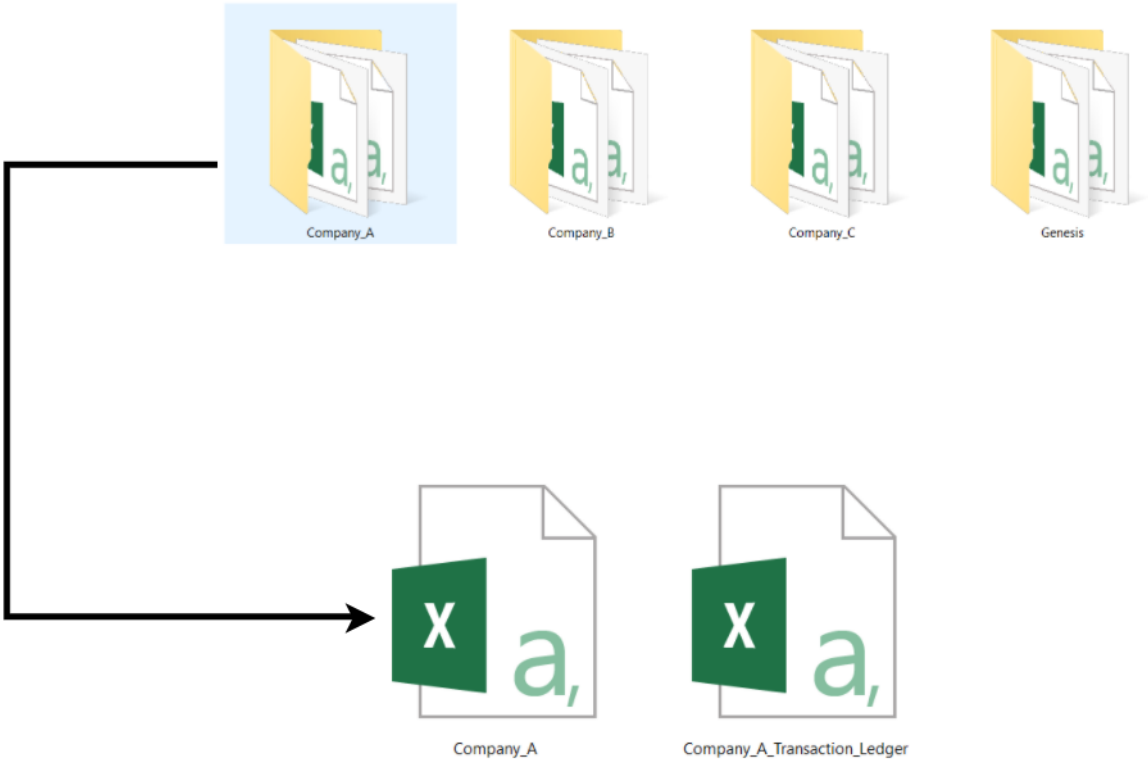


Figure 5.10: The final file structure after 3 companies are inside the chain and contents of each file shown.

The diagram 5.10 shows the creation of multiple nodes in the names of Company A, Company B, Company C respectively. The genesis node is one that is created by default and precedes all the other nodes. Each mode file contains their own copy of data ledger and transaction ledger as shown which makes the data decentralization happen. Realistically this will happen in the client's preferable data-store or cloud data server through additional APIs.

5.2 Joining and Updating files using Pandas and CSV Libraries

According to research work [5], Python has been increasingly gaining in popularity in scientific research in recent years. According to the same source, the 'pandas' library has been under development since 2008 and will one-day lead to scientific research being made a lot more attractive and practical. We have used the 'pandas' and 'csv' libraries to manipulate data files in this research.

Specific template is to be followed for the addition of new datasets as our data format is primarily in the CSV format. The new CSV file, which is to be added to the existing records, is first converted into a data frame. Then the data frame is appended with the dictionary that contains all the previous datasets. Then this new dictionary is converted to a CSV file, which contains all the previous records as

well as the records that were just appended to it. Thus this allows us to maintain a single database with all company records, both previous and new data can be found here.

5.3 Smart Contracts

A smart contract as we know and stated earlier is a computer program or set of programs that execute a certain activity inside the blockchain automatically when a certain trigger is initiated. Like any blockchain, our system is operated with smart contracts. To better explain the flow of methods we can use Figure 5.11.

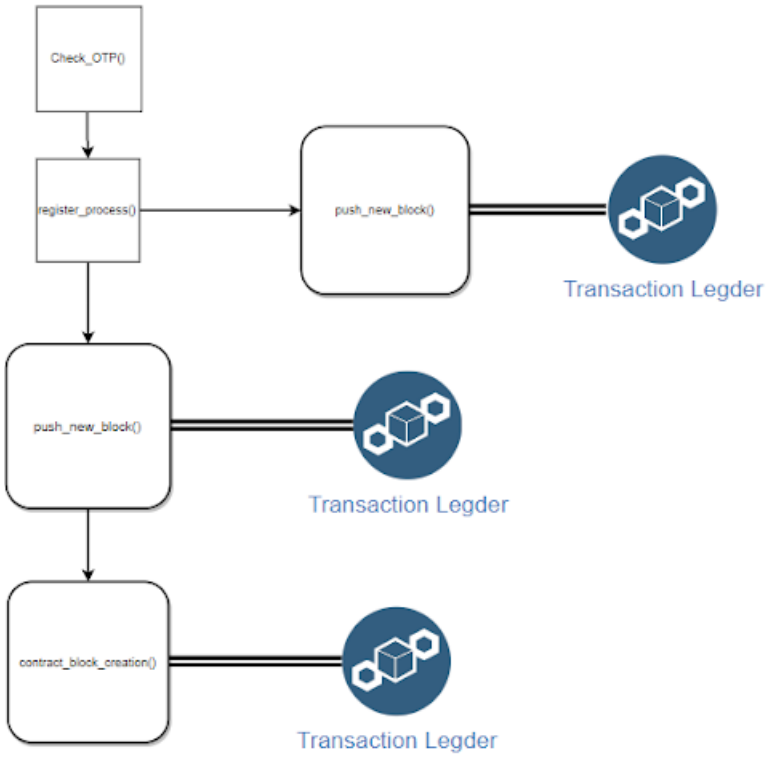


Figure 5.11: The smart contracts shown and connection to transaction ledger

Figure 5.11. is the one use case of the smart contracts which is involved in the registration process. We can see the smart contracts that allow the process to happen as functions or chain codes in the diagram. Figure 5.12 shows the default smart contract list. Figure 5.11 shows the smart contracts involved in the permission protocol. The Check_OTP() function checks the OTP(One Time Pin) provided by the user and then contacts the MSP to check its validity. If valid, the next of functions are run automatically whereby the register_process() runs to register the peer in the system chain. On the successful registration, the push_new_block() method is run is automatically triggered that adds a node in the system. The contract_block_creation() finally gives a decentralized copy of all the data in the other nodes to this newly joined node making the system decentralized. It is also important to notice that every transaction or activity of the smart contracts is noted down in the transaction_ledger to keep the historicity and accountability of the system.

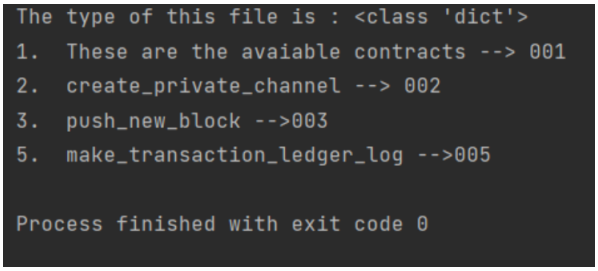


Figure 5.12: Default smart contacts list

5.4 Security and Technical details

5.4.1 Database

Primarily there are two types of data logs or ledgers. The first will be for sharing the organization data in the private channel. The second will be the transaction log. This log will be storing all the transactions in the network to create transparency and accountability. Both the logs will be ledgers that will be immutable in the sense that data cannot be deleted from these ledgers however can be appended via a consensus protocol.

5.4.2 Consensus Protocol

The consensus mechanism is there to make sure that no one peer can automatically update or append any information in the blockchain without the consent of the other peers in the channel. We will be using a simple majority consensus mechanism in a channel whereby one peer will have one vote and a 50 percent plus one voting will determine the consensus of the action.

5.4.3 Encryption

We will be using an advanced encryption standard for encrypting these data.

- The certification authority data,
- The MSP data
- The ID storage for certification authority
- the data stored in the csv files
- The data stored in the transaction logs
- All other databases such as one storing the user credentials
- All data ledgers

Encryption of data ensures security and reliability of all ledgers in the system. Furthermore even in cases where the system is compromised, the integrity of important data such as login credentials can be kept encrypted with near impossible chances of decryption via a malicious party or parties.

5.4.4 Hash Value Calculator

The MSP will be in charge of giving an unique hash value to each block or peer that is created. Each peer in the chain will have the hash value of the previous node and itself. The hash value will be set in such a manner that the hash of the previous block will be used to generate the new hash. This function will be managed by the MSP.

5.4.5 Block Deletion and Revocation

This process will be handled by the MSP. The MSP simply has to revoke the access to the system for that specific peer. Once that specific peer is put in the revocation list then the password for that certain block will be changed by the MSP to a random new password. The organization representative will no longer be able to login to the system and must request the CA for a new certificate ID that will allow it to create a new peer to be added in the private channel. Once the new peer is created then the representative can access it from that node. The last revocated block stays in the channel however is now not accessible by any of the representatives.

5.5 Socket Programming for Decentralization

We have utilized socket programming for our proposed architecture of socket programming for creating the client and server gateway. Our architecture is not centralised as in the client-server model. Rather the system is such that it uses the capacity of socket programming to create a peer-to-peer network that decentralises the processing power of the servers. Decentralised servers will be added as the clients join our system. As stated in [11], the fault tolerance of the system increases with the increase in servers and hence will be the case for our system too.

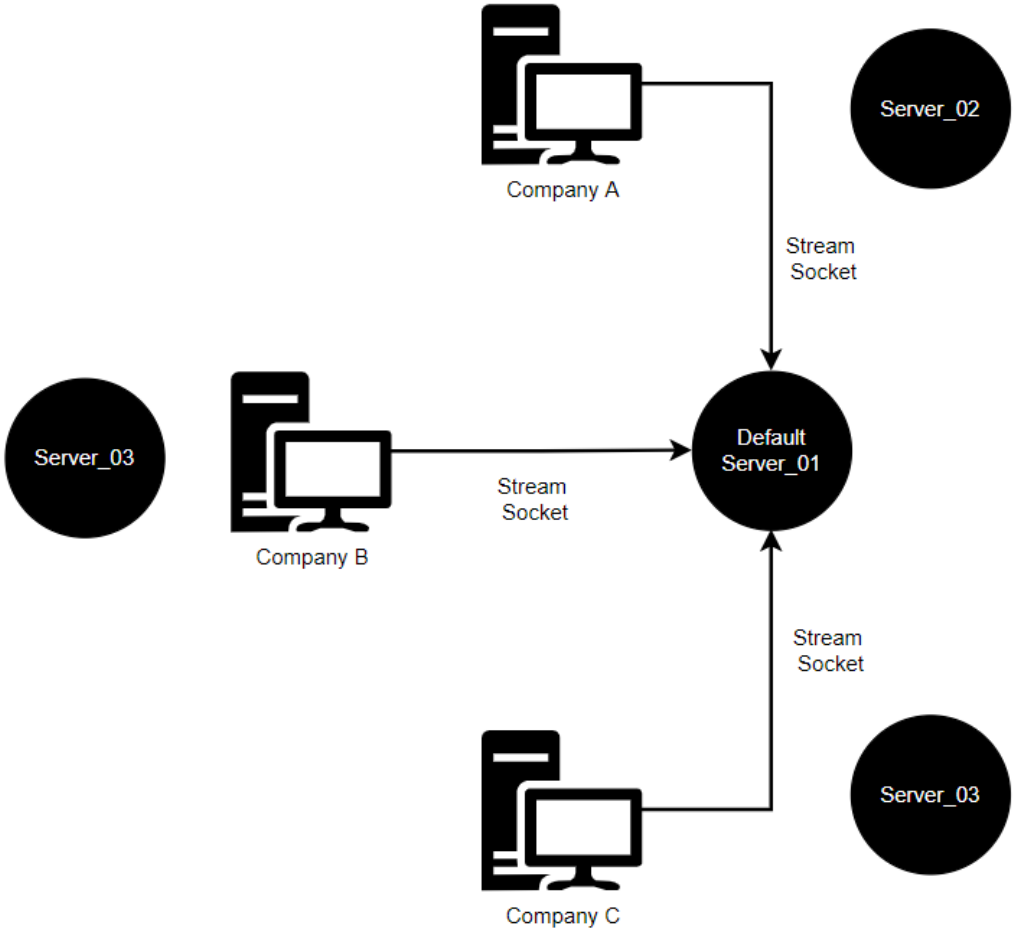


Figure 5.13: Decentralised system before server crash.

As shown in the figure 5.13 above the peers in the blockchain system generally ping a certain server for a certain functionality and interaction. However to achieve processing decentralization there are other servers in the system that can be idle or be involved in some other functions in the system.

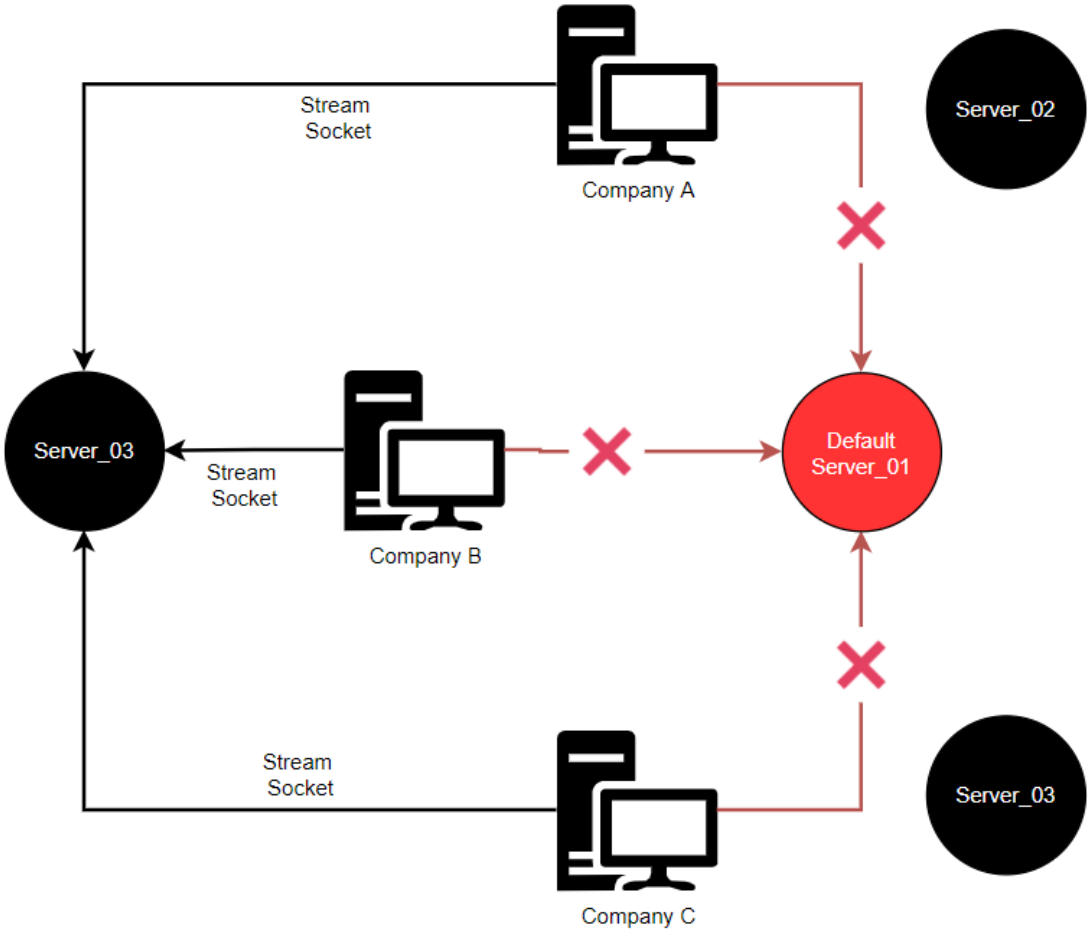


Figure 5.14: Decentralised system after server crash.

As we can see in 5.14, initially the clients in the system are pinging the server_01 for processing the data and storage purposes. However if this server crashes, the second server, server_02 will be allowing the process of the system to continue forward. Since server_02 is independent of server_01, a faulty server_01 will not affect the process of server_02.

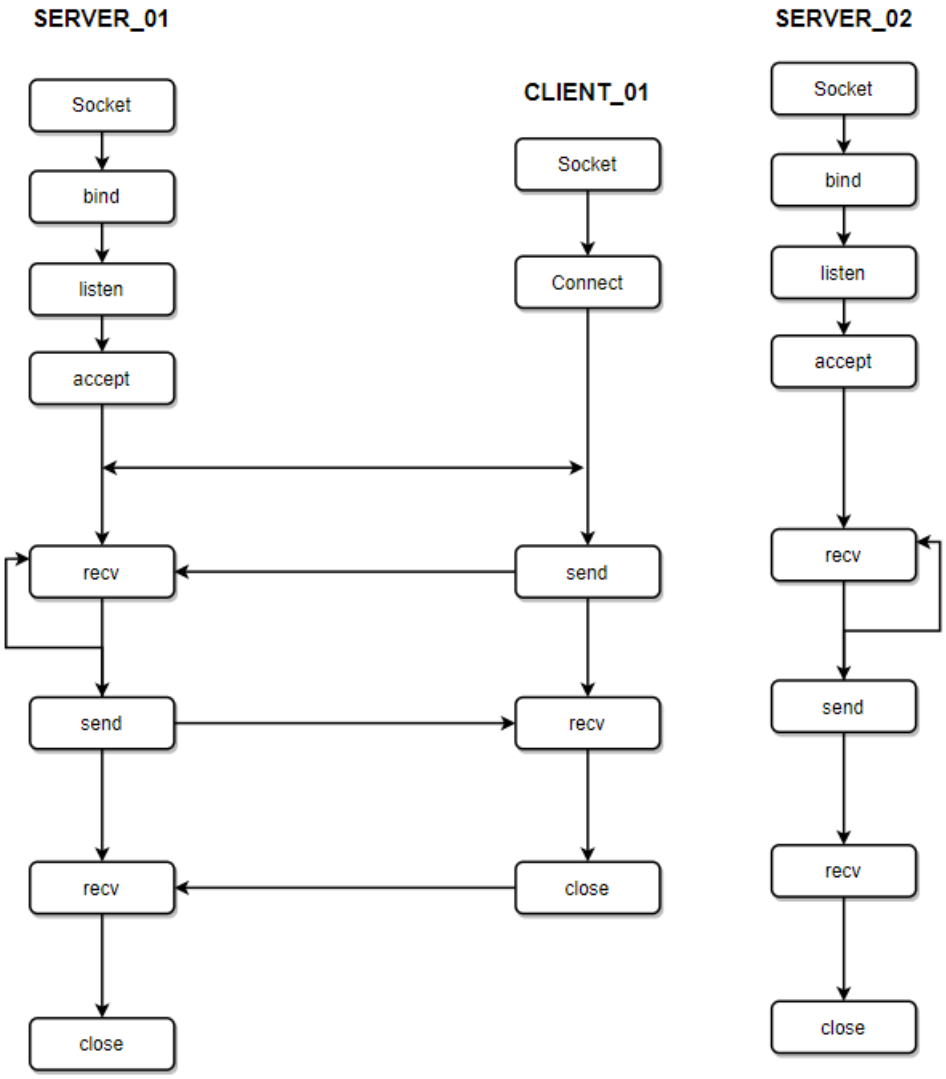


Figure 5.15: Socket connection established with server 01

Figure 5.15 shows the use of socket programming to create a peer to peer network for client-system communication. Here Client_01 is a client script that tries to establish a connection with server_01. The server_02 stays on standby or processes certain parts of the system to be sent to some other servers in the system.

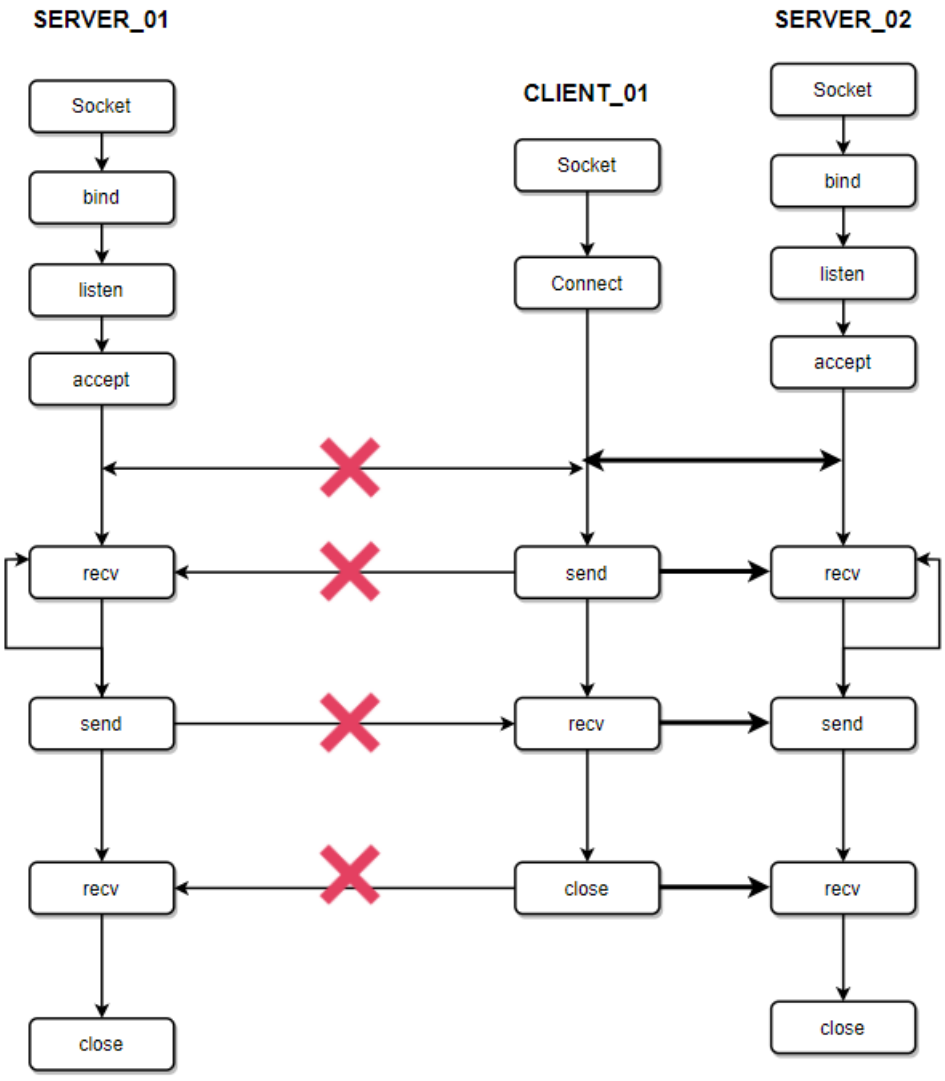


Figure 5.16: Socket connection established with server 02 after client is unable to connect to server 01

Upon the unsuccessful attempt to connect with server_01, shown in figure 5.16, the client can opt to connect to server_02 instead. As mentioned in [4], the use of decentralized servers as shown here allows for decentralised processing and much of the system as it is independent of the arithmetic details can utilize the P2P framework for servers.

5.6 Applications

Through our permissioned blockchain structure, the companies can register themselves and create a consensus in transferring data and information. Since the platform is transparent and non-biased the platform credibility is high and representatives can share the company information at ease. All the data shared is decentralized and hence there is no single point of corruption in data that can terminate the activities of the platform. On successful registration of a peer into the blockchain, all data and transaction ledgers are automatically sent to newly added peer thus ensuring data decentralisation. On addition of new records in the data ledger, newly updated data ledger is sent to all peers in the blockchain while storing the transaction data in transaction ledger which is also shared with all the peers. AES encryption of all data and transaction ledgers also ensures that the data is safe and unbreakable without the right administrative bodies involved [57]. The research work [51] points out the tradeoff between security and complex business logic and recommends the preservation of complex business logic and hence is the case for our proposed blockchain simulation where the business logic is put priority.

This data can therefore be manually verified by permitted companies. Moreover, permitted company can retrieve the data manually for use in the next step which is performance appraisal prediction using machine learning.

According to [26], the multi-client usability of blockchain makes it a fast and efficient system for sharing data that might be needed dynamically for processes to run as is the case for our HR system. Our blockchain structure will ensure transparency as the files and data can be viewed by all the peers in a certain private channel. Enterprise-level blockchain like this will ensure transparency in the field and bring on better standards and best practices to be shared by all the others in the channel.

According to [54], the use of blockchain makes the system much more transparent and reliable and each individual data is no more required to be looked upon by an admin body and rather the smart contracts automatically keep the system in check. Paper [54] shows that there is a system in blockchain that allows trust and transparency to be maintained and collaboration with untrusted pirates can also be done through a different on-chain method. Since our system has a specific focus on employee information, our system should prevent information fraudulence from the employee side. This is possible because the system keeps track of all the activities in the chains and then the activities can be analyzed for any sort of anomaly in the data. Also since machine learning is being utilized in the system so the system can find out any anomaly in the data that is provided in further processing. According to [51], some bottlenecks can occur due to the design of the chain code or smart contracts in the blockchain. The further incorporation of machine learning in this scope will allow for better reasonableness and performance analysis of this feature of the blockchain system.

5.7 Legal, Privacy, and Ethical Issues

Business ethics refers to the standard of morally right and wrong operations in business. Businesses ethics remains to be an important research aspect as our proposed solution is concerned with the enterprises. According to [34], the use of blockchain technology in job recruitment and relevant industry is beneficial and ethically valid from all the four analyzed aspects of utilitarianism, contractarianism, virtue ethics, and from the viewpoint of deontologist groups of thinkers. Here, it is also recommended that free will should be open to individuals in the blockchain system, meaning, that they should have consent and will to join the blockchain system and have control over their data. Therefore we recommend a system that utilizes smart contracts and chain codes to ensure the consent of individuals and stakeholders while designing the architecture to satisfy the ethical concerns of individuals. We also recommend further efforts in research in the ambivalent use of blockchain, specifically private and enterprise-level blockchains, with a specific focus on involving decision-makers at enterprises for better insights in further research.

Data privacy is a pertinent issue. Privacy of sensitive data is a major concern to our recommended system. Privacy and secrecy of confidential and sensitive data such as personal information are very important to be handled with extreme care and stored with vigilant efforts. As stated in [23], data sharing can promote process, data, and decision transparency, finally yielding greater utility and hence increasing the utility of business as also stated in this research, and thus data privacy and security should be a pivotal focus in private permissioned blockchains. Therefore continued trust-building efforts are to be put forward by increasing individuals' trust towards the system as stated in [34] through analyzing the blockchain options and alternatives in the blockchain. Furthermore, extension to proof-of-work other trust-based mechanisms is to be looked into in order to derive the best mechanism to satisfy the concern towards data privacy and build confidence. Therefore we propose a blockchain and prediction and ranking system to take into consideration the issue of privacy breach and consent in sharing employee details not only from the enterprise end but also from the employee end.

Due to the ever-evolving and developing legal frameworks and jurisdiction towards blockchain in every domain, the architecture of the enterprise blockchain must also adapt to changes in the legal system to accommodate rules and legal policies. Different national policies of governments towards blockchain usage have to be taken into special consideration and flexible architecture is to be followed to adapt to the legal needs privacy concerns and specific needs of industries. The detailed analysis of the legal issues of blockchain is well outside the scope of this paper however the legal research of blockchain will open up new avenues to the design feature of private permissioned blockchain. According to [28], all technological infrastructures must fall under the local and regional data protection acts such as the GDRP, and hence the usage of enterprise private blockchains must comply with these rules and regulations. Additional research needs to be done in order to understand the specific usage depending on the context for blockchain in further research. Regulatory framework for blockchain technology at regional and national levels should be developed via further research to ensure, as also stated in [34], insure minimization of negative

externalities and create a level playing field for stakeholders.

Chapter 6

Experiment & Result Analysis

Our research aims to find an system to verify applicant information and retrieve previous employee record from permissioned blockchain based achitecture, to use in predicting their performance appraisal scores and rank them. This section focuses on the experiments carried out to find the best performing machine learning algorithm, develop a ranking algorithm and make a web-based application. The following sections have in-depth discussions of the experiments.

6.1 Candidate Performance Score Prediction using Machine Learning

This section has been broken down into three stages, where in the first stage, we evaluate the performance of five machine learning algorithms using min-max scaling and standard scaling. Here we find standard scaling works better for most of the learning models. We therefore introduce XGB and re-evaluate it with a decision tree and random forest on an unscaled data frame. In the second stage, we remove the feature with the highest correlation and perform RFE to find the best fourteen features, and do result analysis to check this algorithm will be usable. Lastly, we rearrange the dataframe for each learning model by selecting features with the highest correlation and features found from the previous stage. At last, we analyze the data. We have used google colab to clean and scale our datasets and use classification models on them.

6.1.1 Feature Scaling and Machine Learning Algorithm Analysis

In order to find the predicted values of performance rating, two different types of scaling have been used, Minmax Scaling and Standard Scaling, and five different classification models have been used which are Logistic Regression, Random Forest, Decision Tree, Naive Bayes, and Neural Network. Through this, we are able to compare the scaling techniques and the classification models to find the best approach for predicting our candidate performance.

We have scaled our train and test twice, first using Min-Max Scaling and then using Standard Scaling. We have used these two types of scaling to run on five different

classification models to get 10 sets of accuracy results which has been discussed in the results and analysis section.

Comparison & Result Analysis

After scaling our data in two ways and running five classification models, we get 10 sets of results for the accuracy of each model. The comparison is shown in the table below.

Table 6.1: Accuracy Comparison of Classifiers

ML Classifiers	Min-Max Scaled	Standard Scaled
Neural Network	0.98	0.97
Naive Bayes	0.99	0.99
Logistic Regression	0.98	0.99
Decision Tree	1.00	1.00
Random Forest	1.00	1.00

In Table 6.1, we have compared the accuracy values between Min-Max Scaled data and Standard Scaled data for four different classification models which are Naive Bayes, Neural Network, Decision Tree, and Logistic Regression. From the results we can assume that Standard Scaler gives better accuracy results among the two scaling methods.

Confusion Matrix and AUC-ROC curve generation

A confusion matrix sums up the predictions a classification problem produces. It illustrates the ways in which a classification model can be confused when making a prediction. The matrix divides the predicted values into separate sets; True Positive, True Negative, False Positive, and False Negative. False Positive and False Negative indicate an error while making a prediction. Confusion Matrix is used to find F1 score and AUC-ROC curve.

AUC-Roc curve shows the relationship between True Positive Rate (Y-axis) vs False Positive Rate (X-axis,). A higher number of X indicates that there is a higher number of False Positives than True Negatives. A higher value of Y indicates that there is a higher number of True Negative than False Positive.

From the previous stage, we know standard scaling gives higher accuracy for all instances, therefore, in this stage, we have used Standard scaling for our train and test data and plotted the confusion matrix and AUC-ROC curve for all our five classification models. Therefore, we get the following results.

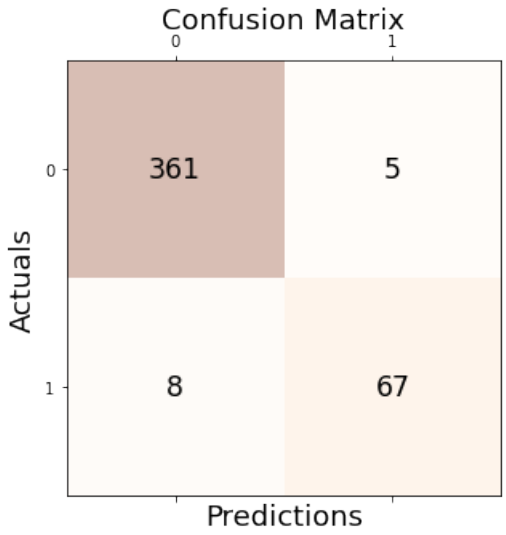


Figure 6.1: Confusion matrix of Neural Network Classifier

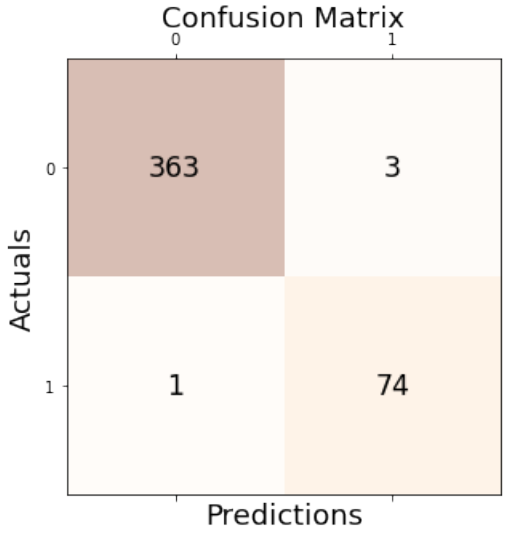


Figure 6.2: Confusion matrix of Naive Bayes Classifier

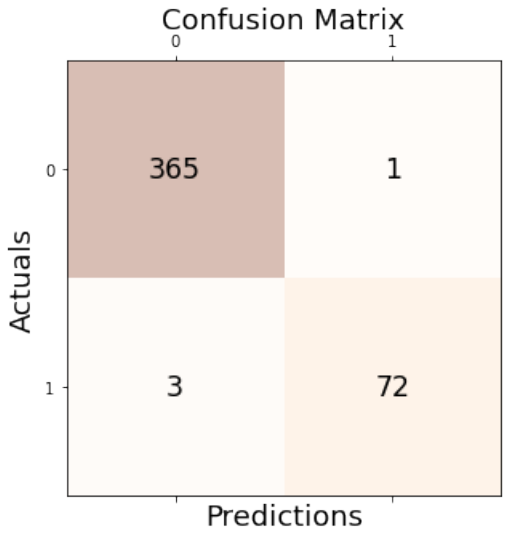


Figure 6.3: Confusion matrix of Logistic Regression

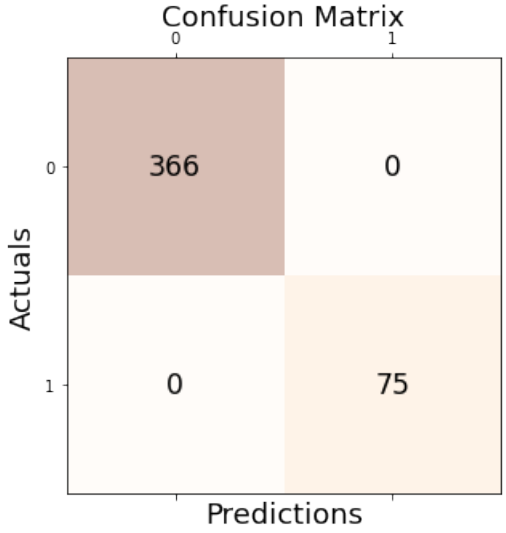


Figure 6.4: Confusion matrix of Decision Tree

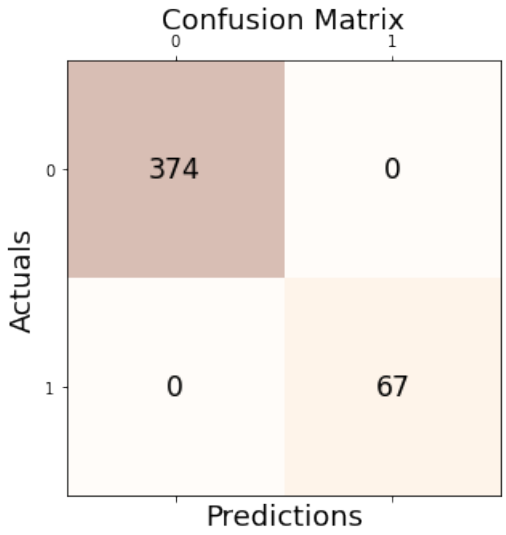


Figure 6.5: Confusion matrix of Random Forest

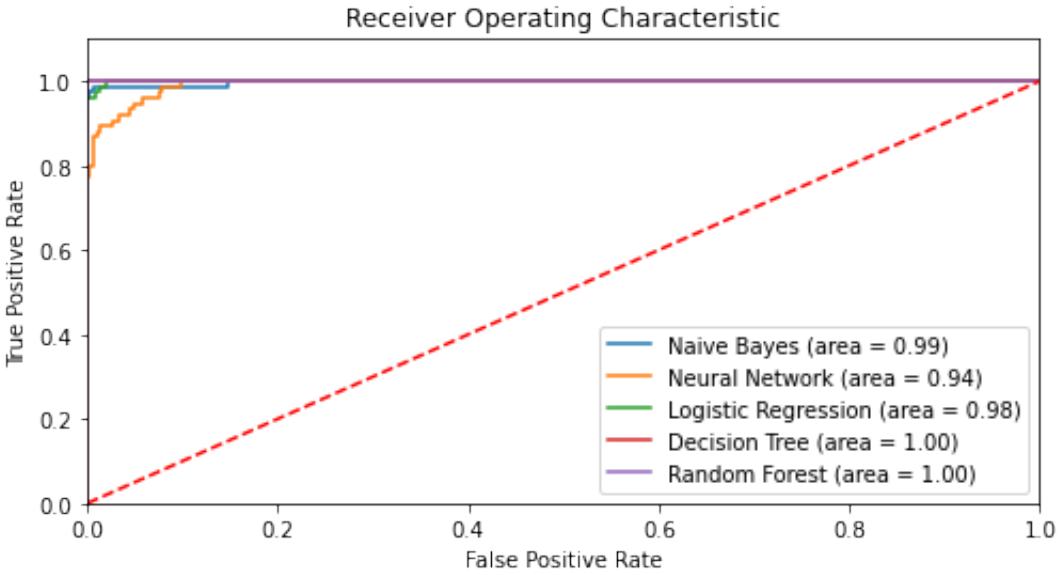


Figure 6.6: AUC-ROC curve for five classification models

Figures 6.1 to 6.5 show the confusion matrix of all the five machine learning models and figure 6.6 shows the AUC-ROC curves generated by each confusion matrix. From the results, we can say only Decision Tree and Random Forest is performing well compared to all other classifiers. We can state this by seeing the confusion matrix There is False Negative for all classifiers except Decision tree and Random Forest, where, the False Negative values are zero. On the other hand, for Naive Bayes, Neural Network Classifiers, and Logistic Regression the False Positive value in the confusion matrix is greater than zero which means that this algorithm is also not suitable for this dataset.

F1 scores Comparison F1 score is a function of Precision and Recall which is calculated from our confusion matrix. Precision describes how many true predictions actually turned out to be true. Recall describes the number of positive predictions the model could identify correctly. The formula for F1 is given below.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall)}$$

We have calculated f1 values for each of our models which have been scaled using standard scaling and this score is together used with Confusion Matrix and AOC-ROC curve to compare the methods.

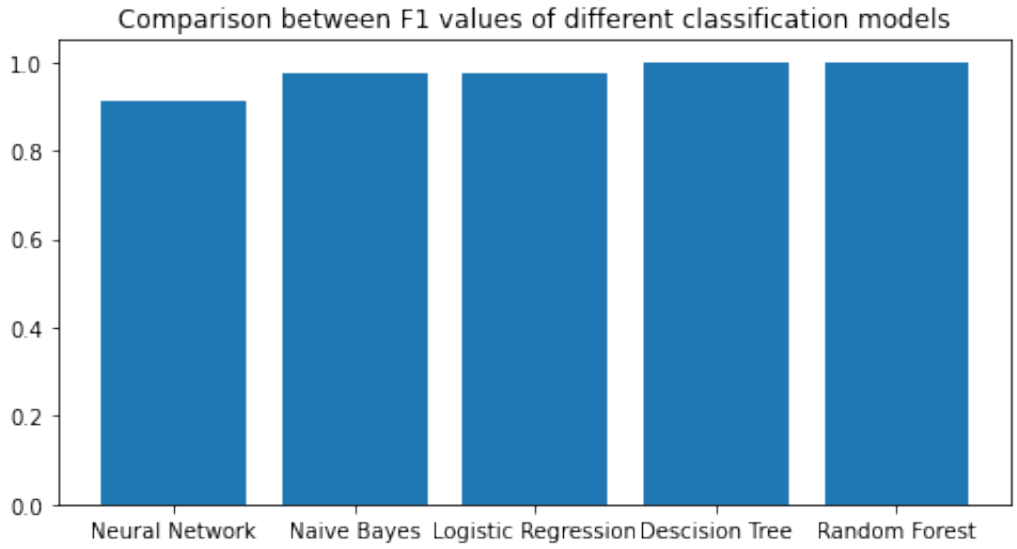


Figure 6.7: Comparison between classifiers in terms of f1 scores

In figure 6.7, we can again see that Decision Tree Classifier and Random Forest are working best in both cases as the f1 score is 1.0 whereas for other methods the results are slightly below 1.0.

Therefore, we can conclude by stating that Decision Tree and Random Forest are working best for this dataset based on their Accuracy score, Confusion Matrix, AUC-ROC curve, and F1 score.

6.1.2 XGBoosting

Since in the previous section, we concluded that Decision Tree and Random Forest worked best when our data set is scaled. Decision tree is a type of supervised machine learning algorithm while Random Forest is a type of ensemble machine learning algorithm. Therefore, now we want to see how another ensemble machine learning model called XGBoost works for our dataset. XGBoost is a scalable machine learning model to boost trees. According to research [53], XGBoost is an ensemble machine learning algorithm where several trees are combined to give the prediction. Ensemble Machine Learning does not require feature scaling due to not being affected by variance in data, therefore from now we will not do feature scaling. In this section, we will analyze performance of Decision Tree, Random Forest, and XGBoosting in unscaled data.

We ran all the decision trees and the two ensemble learning algorithms, namely Random Forest and XGBoost in our unscaled dataset and got the following results shown in table 6.2 and figure 6.8, 6.9 and 6.10.

Table 6.2: Accuracy Score and F1 score of different machine learning algorithms in unscaled data

Machine Learning Model	Accuracy Score	F1 Score
Decision Tree	1.00	1.00
Random Forest	1.00	1.00
XGBoost	1.00	1.00

In Table 6.3 we can see the Accuracy Score and F1 score for all the models are 1.0, which means all models are performing well for this dataframe.

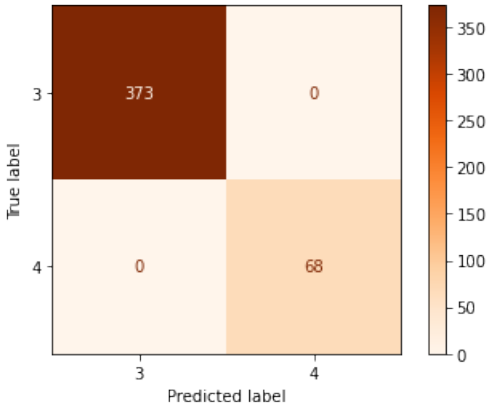


Figure 6.8: Confusion Matrix of Decision Tree on unscaled data

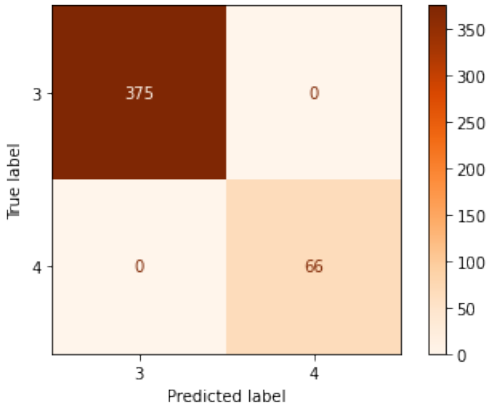


Figure 6.9: Confusion Matrix of Random Forest on unscaled data

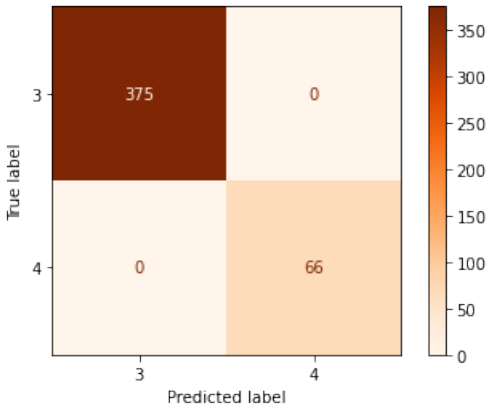


Figure 6.10: Confusion Matrix of XGBoost on unscaled data

From the data, in Figures 6.8, 6.9, and 6.10, we can again see all the confusion matrices are performing well as there are no values in False Negative and False Positive. Therefore, it can be understood that the decision tree and all ensemble learning models work perfectly for the unscaled dataset. However, after we drew a tree using our decision tree model shown in figure 6.11, we saw only the feature percentage_salary_hike is both the root node and decision node.

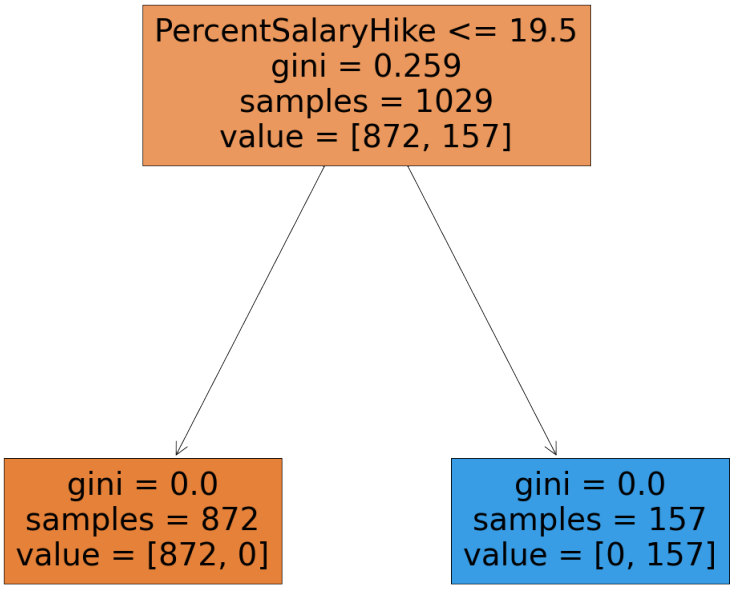


Figure 6.11: Decision Tree for unscaled data

6.1.3 Recursive Feature Elimination on Ensemble Learning Algorithm & Result Analysis

In the previous section, we proved all three machine learning models work well for unscaled data. To further evaluate the ensemble models, in this section we will carry out further feature reduction to re-evaluate the models to find the best performing model for predicting candidate performance rating.

Recursive Machine Learning Algorithm is a feature selection technique that recursively eliminates features after ranking their importance.

During data pre-processing we found percentage_salary_hike has the highest corelation of 0.77 with our label which is performance_rating feature, and this value was very high compared to other correlation values. Therefore, while carrying out RFE with this feature in the data, the algorithm only selects a set of values before or after this feature.

Hence, for this stage we will conduct RFE on data without percentage_salary_hike feature to reduce the impact of a correlation value and rank other features in an unbiased system.

For RFE on decision tree, random forest, and XGB, we started to print 14 best features for each particular model, and the following three tables 6.5, 6.6, and 6.7 are the results which show a list of 14 features that are suitable for each model to work efficiently.

Table 6.3: Features selected after using RFE on Decision Tree

Selected Features
Age
DailyRate
DistanceFromHome
HourlyRate
JobInvolvement
JobSatisfaction
MonthlyIncome
MonthlyRate
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsWithCurrManager

Table 6.4: Features selected after using RFE on Random Forest

Selected Features
Age
DailyRate
DistanceFromHome
HourlyRate
JobRole
MonthlyIncome
MonthlyRate
NumCompaniesWorked
TotalWorkingYears
TrainingTimesLastYear
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

Table 6.5: Features selected after using RFE on XGBoost

Selected Features
BusinessTravel
DailyRate
DistanceFromHome
HourlyRate
JobRole
JobSatisfaction
MonthlyIncome
MonthlyRate
NumCompaniesWorked
TotalWorkingYears
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion

The following table 6.6 shows the accuracy score, F1 score and figure 6.12, 6.13 and 6.14 shows the confusion matrix produced after RFE is done.

Table 6.6: Performance of different machine learning algorithms after RFE run on data without PercentSalaryHike feature

Machine Learning Model	Accuracy Score	F1 Score
Decision Tree	0.71	0.171
Random Forest	0.84	0.00
XGBoost	0.85	0.056

In table 6.8, we can see the result analysis of the decision tree and two ensemble learning models using accuracy score and F1 score. It can be observed that the performance of all three models lower while using them on data without the feature PercentSalaryHike. Among the model although it can be found XGB works best in the situation but it can also be realised that these results are not ideal and predictions must be done in data with PercentSalaryHike feature.

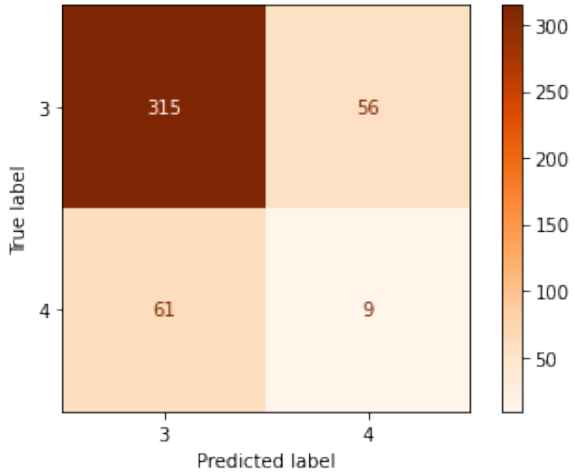


Figure 6.12: Confusion Matrix of Decision Tree after RFE

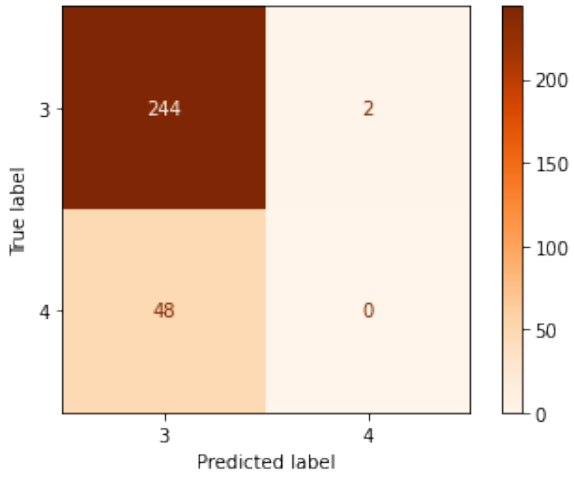


Figure 6.13: Confusion Matrix of Random Forest after RFE

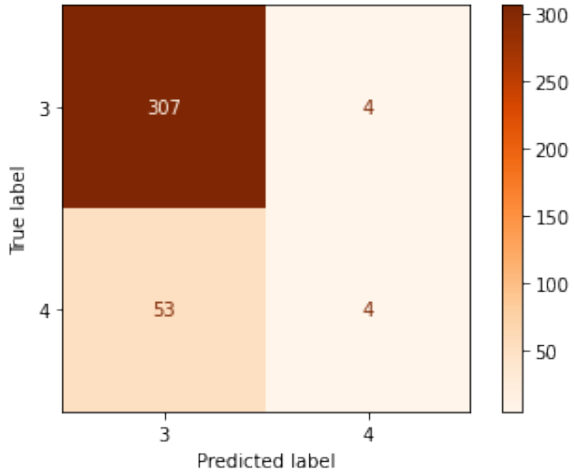


Figure 6.14: Confusion Matrix of XGBoost after RFE

From the results, in Figures 6.12, 6.13, and 6.14, we can see the confusion matrix results are not good enough as there are high values in the False Negative and False Positive sections. Therefore, through this analysis, we can conclude which features have the highest importance after the PercentSalaryHike feature, for each ensemble model. We also learn it is not efficient to use data without the PercentSalaryHike feature as the accuracy scores are lower, the confusion matrix is uneven and F1 scores are very low in Table 6.4.

6.1.4 Final Evaluation

From the analysis of previous section, in this stage we will take the 14 features we got separately which are listed in figure 6.3, 6.4 and 6.5 for decision tree, random forest and XGB respectively along with PercentSalaryHike feature and run them with their respective models.

After running the data with their respective models we get the following results shown in table 6.7 and figure 6.15, 6.16 and 6.17.

Table 6.7: Performance of different ensemble learning algorithms after RFE

Machine Learning Model	Accuracy Score	F1 Score
Decision Tree	1.00	1.00
Random Forest	1.00	1.00
XGBoost	1.00	1.00

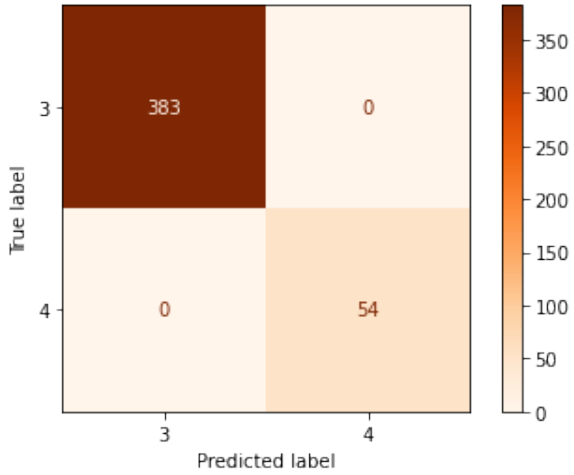


Figure 6.15: Confusion Matrix of Decision Tree on final dataframe

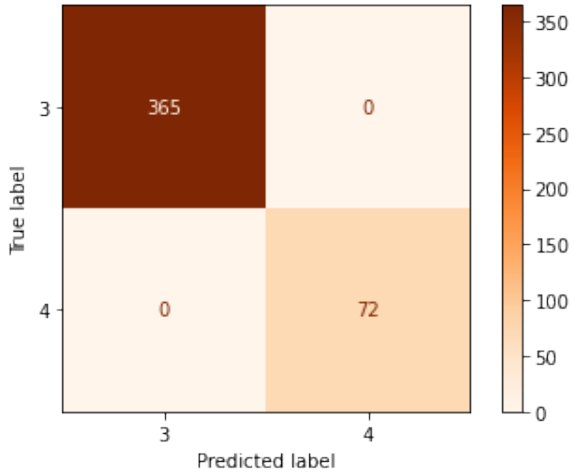


Figure 6.16: Confusion Matrix of Random Forest on final dataframe

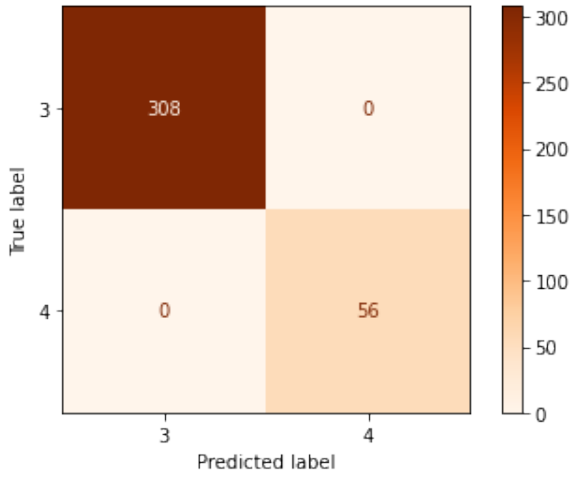


Figure 6.17: Confusion Matrix of XGBoost on final dataframe XGBoost

From the above results we can conclude, decision tree and the two ensemble algorithms work well for our proposed system. We can see in Table 6.5 that the accuracy scores and F1 scores is 1 for all three ensemble models. Meanwhile in figure 6.15, 6.16 and 6.17, all confusion matrix has 0 data in False Nagtive and False Positive. In conclusion, decision tree and all the ensemble learning models are working well.

6.2 Pickling Models

Pickle is a Python library for serializing and deserializing object structures. This process is therefore used to transform a Python object into a byte stream in order to save it to a file/database, keep the program state between sessions, or send data over the network. When unpickling a byte stream, the pickle module first makes a copy of the original object, then fills it with the right data.

For our research, we wanted to make an app to demonstrate how data can be predicted and ranked, therefore to make the app we pickled our models from.ipynb files and ran it in pycharm with streamlit library. Therefore, our decision tree and ensemble models are pickled from google colab while and later un-pickled in pycharm for being used in Streamlit application.

6.3 Candidate Ranking using Topsis

In [13] we found the mathematical formulation for implementing TOPSIS for ranking based on similarity scores. In TOPSIS there are a number of steps to find the final score. In our MCDM.ipynb file, we have applied the following steps to calculate the scores for each employee using the TOPSIS method.

Step 1. The topsis method starts with the dataset and the weight of each criterion in the divided dataset. The weights are pre-decided and put as an input to the algorithm. The dataset values are then used to calculate the normalized ratings.

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{i\,j}^2}} \tag{6.1}$$

The normalization of the values are then initiated. The normalization process is done using the formula above. Normalization ensures a common scaling of the data in the data frame.

Step 2. The normalized values are then fed to find out the weighted normalized rating using the formula given. This step takes into consideration the relative importance of each step according to the weight values. The weights are multiplied by the normalized values found previously.

$$V_{ij} = w_j n_{ij} \ for \ i = 1, ..., m; \ j = 1, ..., n. \tag{6.2}$$

Step 3. The maximum and the minimum values are then found out to find out the positive ideal being the maximum value and the negative ideal being the negative value. The positive and the negative ideal solutions will be used to determine the standard of each applicant via the distance from the positive, the negative ideal solutions.

$$A^+ = (v_1^+, v_2^+, ..., v_n^+) = ((_i max v_{ij} \mid j \in I), (_i min v_{ij} \mid j \in J)) \tag{6.3}$$

$$A^- = (v_1^-, v_2^-, ..., v_n^-) = ((_i min v_{ij} \mid j \in I), (_i max v_{ij} \mid j \in J)) \tag{6.4}$$

Step 4. Next, we calculate the separation measures using the formula given below.

$$d_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^+)^2}, \ i = 1, 2, ..., m. \tag{6.5}$$

$$d_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^-)^2}, \ i = 1, 2, ..., m. \tag{6.6}$$

Step 5. Then we calculate the relative similarities of each value to the ideal solutions. The similarities give us the information of the candidate's position from the

positive ideal.

$$R_i = \frac{d_i^-}{d_i^- + d_i^+} \tag{6.7}$$

Step 6. The data is then used to make a ranking for all the alternatives that give us the final scoring. Then we find out the similarity score of each value in the dataset. This gives us an idea of the overall performance scoring of the employees. We then sort the values in the data frame in a descending order, whereby the most suitable candidate with highest score value can be seen at the top and the least suitable candidate with lowest score is at the bottom Hence, we have ranked the employees based on their scores for the next step of the recruitment process.

Results

After applying the TOPSIS algorithm and calculating the similarity score for each column, we get the following results.

Table 6.8: Scores for first five employee using Topsis method

Employee ID	Score
ID1	0.484038
ID2	0.290065
ID3	0.472685
ID4	0.431198
ID5	0.48178

Table 6.8 shows the first five results after implementing topsis, where the Employee ID column shows the ID of the employee while Score shows the score we found using the Topsis model.

Next, we decode the values and the actual names of the columns of the dataset are returned to the data frame. We convert the employee id into a list for the next manipulation. The data of the similarity score is added to the data frame as a column value. We make a final data frame named 'employee' which contains the score column along with all other features with their original string values to make it easier for the HR manager to understand the details of the employee. Next, we sorted the values in the data frame in a descending order, whereby the best performing employee with the highest score value can be seen at the bottom and the worst-performing employee at the bottom. Hence, we have ranked the employees based on their scores for the next step of the recruitment process

Table 6.9: Sorted Topsis Score of applicants

Employee ID	Score
ID596	0.720468
ID291	0.717584
ID1278	0.692787
ID1304	0.689587
ID545	0.68944

Table 6.9 shows how table 6.8 changes after we sort the vales of score from highest to lowest. Therefore ranking the most suitable candidate first and follows the pattern in decreasing order of scores. Here, we can see the top five candidates are ranked.

6.4 Prediction and Ranking implementation in Streamlit Application

To visualize our results we have chosen Streamlit. Streamlit is an opensource python library to creating machine learning web applications. We have created two sections in our streamlit app, one left section is used to upload csv file that consists of candidate data while the right section displays the uploaded csv file and has buttons to make prediction either using decision tree, random forest or XGBoost. There is also another button to rank the candidates using MCDM model. Figure 6.18 shows the the UI of our streamlit app.

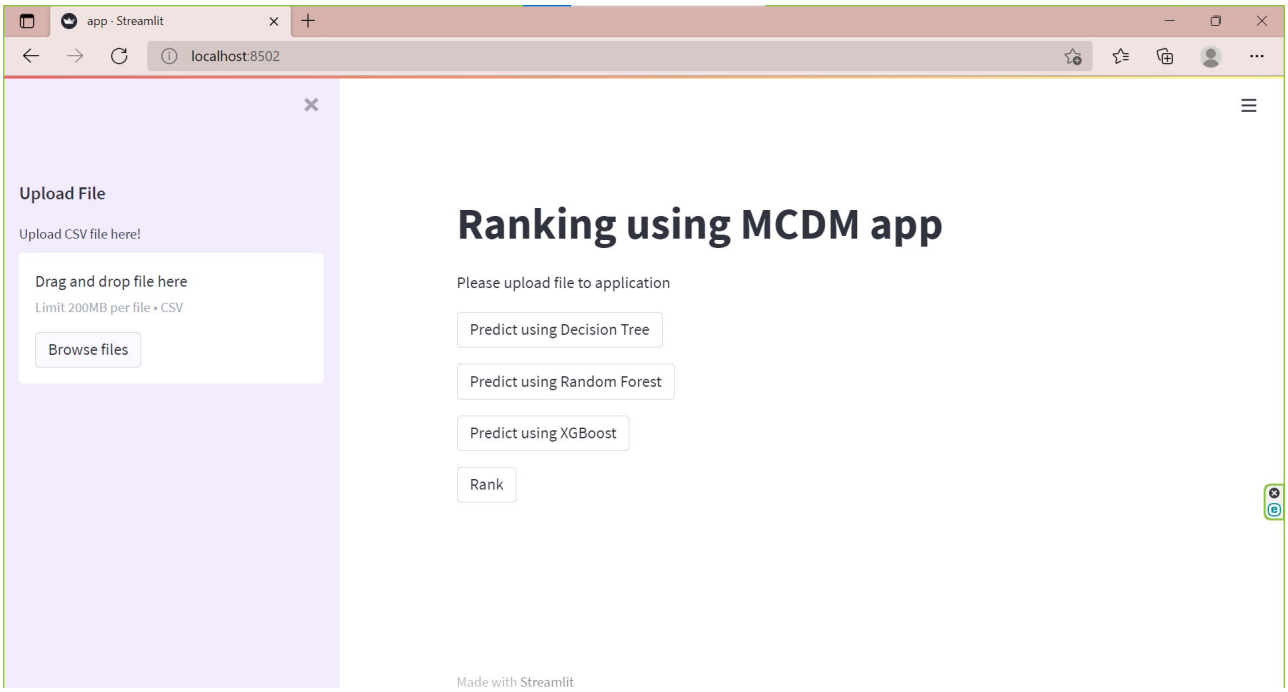
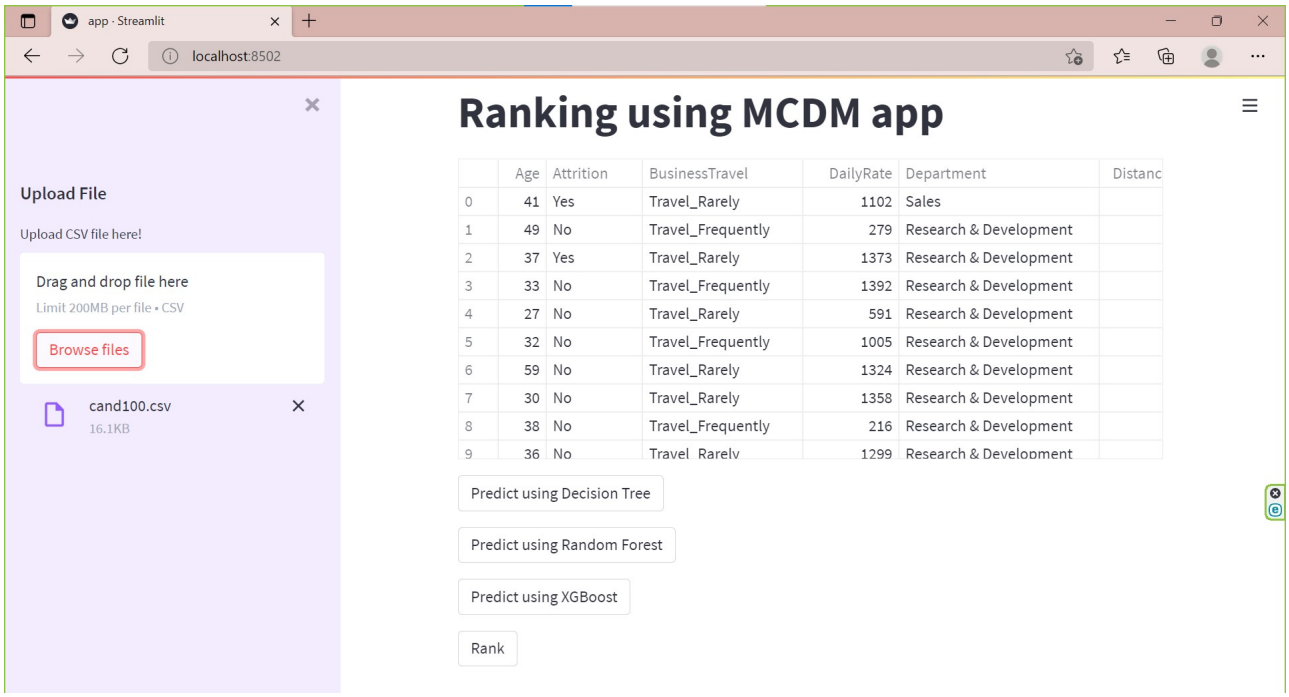


Figure 6.18: Overview of Streamlit App

6.4.1 Upload and Display CSV file

Figure 6.19 shows what happens when we upload file in our Streamlit app. There is a browse file button where we click and upload a CSV file. In this case we have uploaded a CSV file consisting of 100 candidates information. After uploading, the contents in this file is displayed on the other section.



	Age	Attrition	BusinessTravel	DailyRate	Department	Distanc
0	41	Yes	Travel_Rarely	1102	Sales	
1	49	No	Travel_Frequently	279	Research & Development	
2	37	Yes	Travel_Rarely	1373	Research & Development	
3	33	No	Travel_Frequently	1392	Research & Development	
4	27	No	Travel_Rarely	591	Research & Development	
5	32	No	Travel_Frequently	1005	Research & Development	
6	59	No	Travel_Rarely	1324	Research & Development	
7	30	No	Travel_Rarely	1358	Research & Development	
8	38	No	Travel_Frequently	216	Research & Development	
9	36	No	Travel_Rarely	1299	Research & Development	

Figure 6.19: Streamlit app view of uploaded dataframe

6.4.2 Prediction

When we click on the button, 'Prediction using Decision Tree', the app predicts the PerformanceRating values for the dataset and displays the predicted values below. It also updates the PerformanceRating values in ranking section so that ranking can be done with the predicted scores we found using one of the ensemble learning algorithm. The same process repeats with 'Prediction using Random Forest' and 'Prediction using XGBoost' button where the only change is random forest classifier and XGB models are used respectively. Figure 6.20, 6.21 and 6.22 shows the predictions by each model.

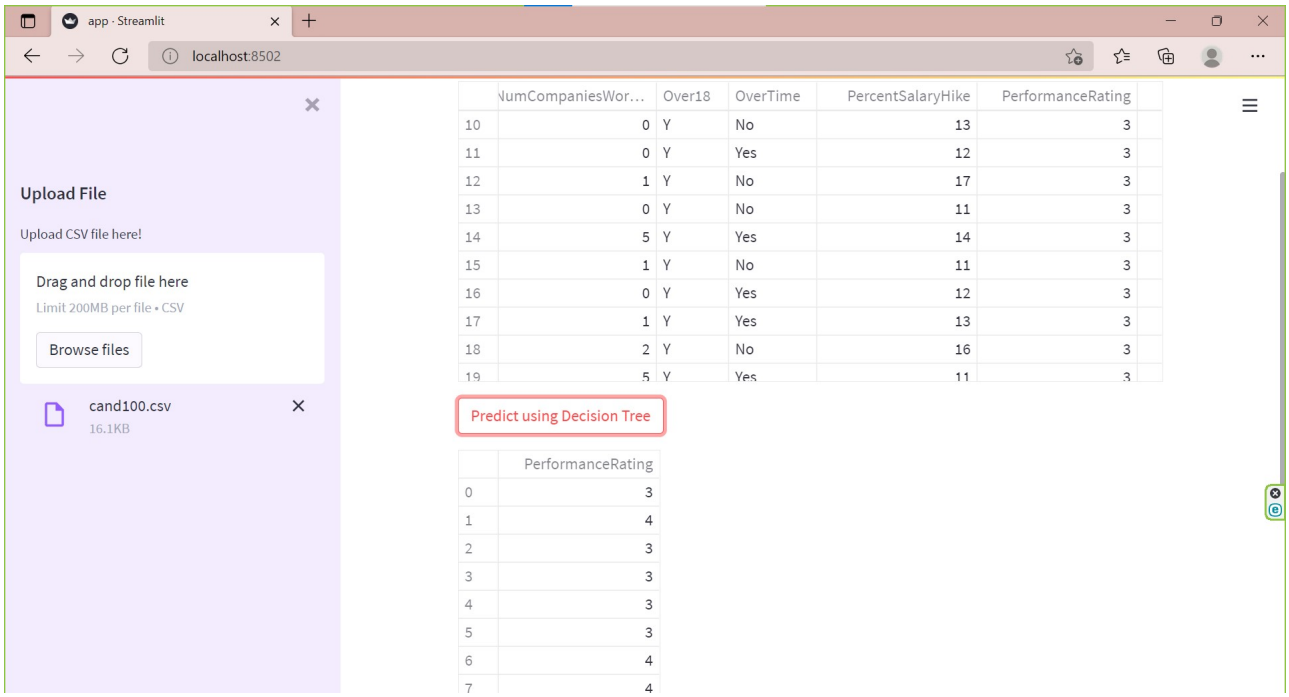


Figure 6.20: Prediction using Decision Tree

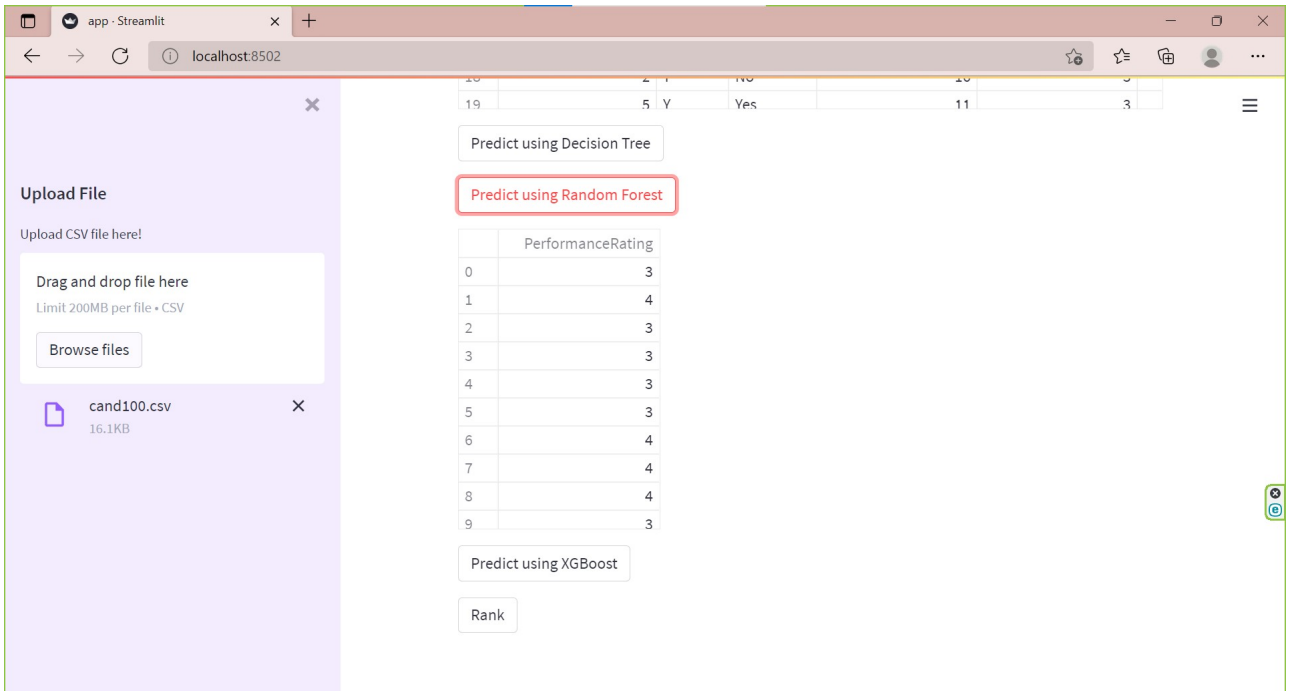


Figure 6.21: Prediction using Random Forest

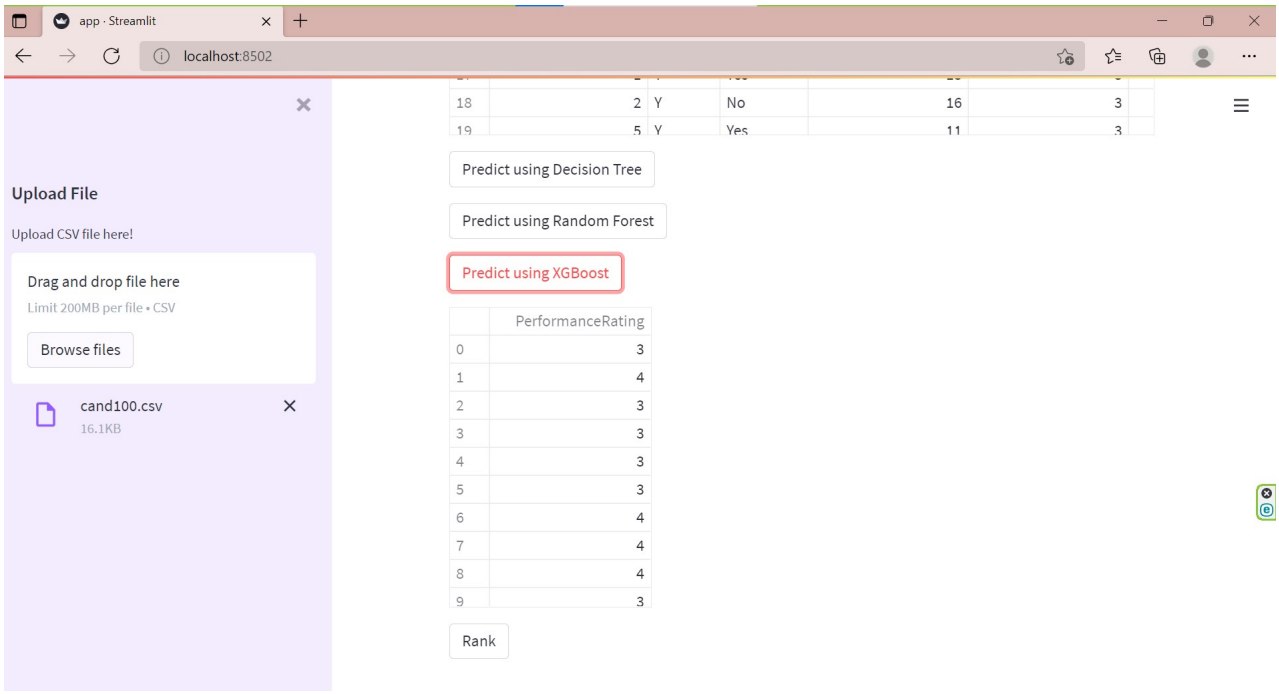


Figure 6.22: Prediction using XGBoost

6.4.3 Ranking

The uploaded data gets updated with PerformanceRating values each time PerformanceRating is predicted. When ranking button is pressed, the updates data is analysed using Topsis algorithm of MCDM model to rank the candidates based on 10 features.

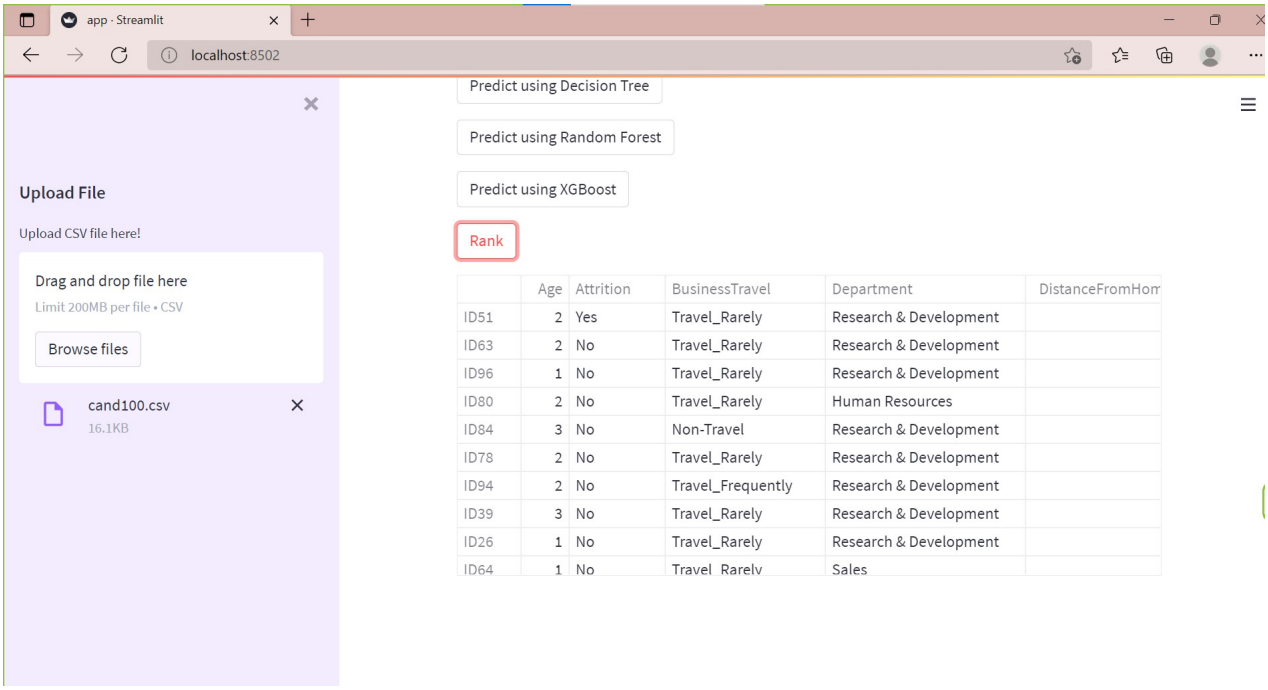


Figure 6.23: Ranking Candidates in Streamlit app

Chapter 7

Conclusion & Future Work

In our research, we used python to create a robust blockchain architecture to store an employee's previous work records which therefore can be used in verification step of recruitment. The performance appraisal data retrieved from blockchain framework along with the information about a candidate can be used in ensemble learning framework to predict future performance appraisal and thus be used in ranking candidates in applicant pool.Lastly, we created a streamlit web application that will collaboratively predict and rank job applicants.

The dataset we used was created by IBM data engineers and is largely used for human resource-related prediction. However, due to discrepancies between correlations among features, this data set fails to replicate reality. Yet we could successfully analyze our data to predict using different machine learning models and thus create our application to automate candidate selection. We would like to improve our dataset in the future and use original data from multiple companies if possible. Hence, we would like to apply our developed methodology in an original dataset.

As part of further improvements for our blockchain architecture, we intend to intrude on the front-end structure of our blockchain. We plan to make the decentralized application web-based via which the company representatives can access the system and the decentralized servers. One other utility improvement can be the use of machine learning algorithms for automatic fraud detection in the blockchain that will allow the blockchain to be more efficient and reliable. Moreover, we want to use encrypted code to fetch employee record therefore, preventing human intervention.

In the future, we want to connect our blockchain architecture with our Streamlit app that will include automatic verification and data retrieval method. The verified and retrieved employee past record will automatically enter our prediction model and ranking algorithm. Thus creating a comprehensive applicant selection app for job roles.

Bibliography

[1] C.-L. Hwang and K. Yoon, "Methods for multiple attribute decision making," in *Multiple attribute decision making*, Springer, 1981, pp. 58–191.

[2] C.-Y. J. Peng *et al.*, "An introduction to logistic regression analysis and reporting 96 j," *Educ. Res*, vol. 3, no. 10, 2002.

[3] D. Zhang and L. Zhou, "Ieee transactions on systems man and cybernetics-part c: Applications and reviews," *513 Discovering Golden Nuggets: Data Mining in Financial Application*, vol. 34, no. 4, 2004.

[4] N. A. Hamid, "A lightweight framework for peer-to-peer programming," *Journal of Computing Sciences in Colleges*, vol. 22, no. 5, pp. 98–104, 2007.

[5] W. McKinney *et al.*, "Pandas: A foundational python library for data analysis and statistics," *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.

[6] C. Bai, D. Dhavale, and J. Sarkis, "Integrating fuzzy c-means and topsis for performance evaluation: An application and comparative analysis," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4186–4196, 2014.

[7] H. Hota, L. Sharma, and S. Pavani, "Fuzzy topsis method applied for ranking of teacher in higher education," in *Intelligent Computing, Networking, and Informatics*, Springer, 2014, pp. 1225–1232.

[8] L. M. Gladence, M. Karthi, and V. M. Anu, "A statistical comparison of logistic regression and different bayes classification methods for machine learning," *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 14, pp. 5947–5953, 2015.

[9] T. Chen *et al.*, "Guestrin, c.: Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, 2016, pp. 785–794.

[10] D. Jesse Yli-Huumo, "Where is current research on blockchain technology," *A systematic review*, 2016.

[11] M. H. Tran, S. V. U. Ha, *et al.*, "Decentralized online social network using peer-to-peer technology," *REV Journal on Electronics and Communications*, vol. 5, no. 1-2, 2016.

[12] A. Abdullah, "Advanced encryption standard (aes) algorithm to encrypt and decrypt data," *Cryptography and Network Security*, vol. 16, pp. 1–11, 2017.

[13] S. A. A. Alrababah, K. H. Gan, and T.-P. Tan, "Comparative analysis of mcdm methods for product aspect ranking: Topsis and vikor," in *2017 8th International Conference on Information and Communication Systems (ICICS)*, IEEE, 2017, pp. 76–81.

[14] S. Tikhomirov, "Ethereum: State of knowledge and research perspectives," in *International Symposium on Foundations and Practice of Security*, Springer, 2017, pp. 206–221.

[15] X. Wang, L. Feng, H. Zhang, C. Lyu, L. Wang, and Y. You, "Human resource information management model based on blockchain technology," in *2017 IEEE symposium on service-oriented system engineering (SOSE)*, IEEE, 2017, pp. 168–173.

[16] X. Zhao, Z. Chen, X. Chen, Y. Wang, and C. Tang, "The dao attack paradoxes in propositional logic," in *2017 4th International Conference on Systems and Informatics (ICSAI)*, IEEE, 2017, pp. 1743–1746.

[17] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An overview of blockchain technology: Architecture, consensus, and future trends," in *2017 IEEE international congress on big data (BigData congress)*, IEEE, 2017, pp. 557–564.

[18] V. Garg, S. Srivastav, and A. Gupta, "Application of artificial intelligence for sustaining green human resource management," in *2018 International Conference on Automation and Computational Engineering (ICACE)*, IEEE, 2018, pp. 113–116.

[19] J. Golosova and A. Romanovs, "The advantages and disadvantages of the blockchain technology," in *2018 IEEE 6th workshop on advances in information, electronic and electrical engineering (AIEEE)*, IEEE, 2018, pp. 1–6.

[20] A. Y. Gromov, T. A. Petrovskaia, A. A. Suslina, N. I. Khizriyeva, and M. A. Stepanov, "Human resources intelligent selection algorithm with improvement of data validity," in *2018 7th Mediterranean Conference on Embedded Computing (MECO)*, IEEE, 2018, pp. 1–4.

[21] A. Mohamed, W. Bagawathinathan, U. Iqbal, S. Shamrath, and A. Jayakody, "Smart talents recruiter-resume ranking and recommendation system," in *2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, IEEE, 2018, pp. 1–5.

[22] P. Sarda, M. J. M. Chowdhury, A. Colman, M. A. Kabir, and J. Han, "Blockchain for fraud prevention: A work-history fraud prevention system," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, IEEE, 2018, pp. 1858–1863.

[23] S. Schwerin, "Blockchain and privacy protection in the case of the european general data protection regulation (gdpr): A delphi study," *The Journal of the British Blockchain Association*, vol. 1, no. 1, p. 3554, 2018.

[24] D. Vujičić, D. Jagodić, and S. Ranđić, "Blockchain technology, bitcoin, and ethereum: A brief overview," in *2018 17th international symposium infoteh-jahorina (infoteh)*, IEEE, 2018, pp. 1–6.

[25] S. Amin, N. Jayakar, S. Sunny, P. Babu, M. Kiruthika, and A. Gurjar, "Web application for screening resume," in *2019 International Conference on Nascent Technologies in Engineering (ICNTE)*, IEEE, 2019, pp. 1–7.

[26] F. Knirsch, A. Unterweger, and D. Engel, "Implementing a blockchain from scratch: Why, how, and what we learned," *EURASIP Journal on Information Security*, vol. 2019, no. 1, pp. 1–14, 2019.

[27] P. Misra and A. S. Yadav, "Impact of preprocessing methods on healthcare predictions," in *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, 2019.

[28] F. Rizal Batubara, J. Ubacht, and M. Janssen, "Unraveling transparency and accountability in blockchain," in *Proceedings of the 20th Annual International Conference on Digital Government Research*, 2019, pp. 204–213.

[29] P. Sanz Bayón, "Key legal issues surrounding smart contract applications," *KLRI Journal of Law and Legislation*, vol. 9, no. 1, pp. 63–91, 2019.

[30] T. Tarusov and O. Mitrofanova, "Risk assessment in human resource management using predictive staff turnover analysis," in *2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)*, IEEE, 2019, pp. 194–198.

[31] A. P. Wibawa, A. C. Kurniawan, D. M. P. Murti, *et al.*, "Naıve bayes classifier for journal quartile classification.," *Int. J. Recent Contributions Eng. Sci. IT*, vol. 7, no. 2, pp. 91–99, 2019.

[32] A. F. Zulfikar, D. Supriyadi, Y. Heryadi, *et al.*, "Comparison performance of decision tree classification model for spam filtering with or without the recursive feature elimination (rfe) approach," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, 2019, pp. 311–316.

[33] M. Dabbagh, M. Kakavand, M. Tahir, and A. Amphawan, "Performance analysis of blockchain platforms: Empirical evaluation of hyperledger fabric and ethereum," in *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, IEEE, 2020, pp. 1–6.

[34] C. Dierksmeier and P. Seele, "Blockchain and business ethics," *Business Ethics: A European Review*, vol. 29, no. 2, pp. 348–359, 2020.

[35] H. Hewage, K. Hettiarachchi, K. Jayarathna, K. Hasintha, A. Senarathne, and J. Wijekoon, "Smart human resource management system to maximize productivity," in *2020 International Computer Symposium (ICS)*, IEEE, 2020, pp. 479–484.

[36] S. Hua, S. Zhang, B. Pi, J. Sun, K. Yamashita, and Y. Nomura, "Reasonableness discussion and analysis for hyperledger fabric configuration," in *2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, IEEE, 2020, pp. 1–3.

[37] L. T. Khrais, "Comparison study of blockchain technology and iota technology," in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, IEEE, 2020, pp. 42–47.

[38] T.-H. Kim, G. Kumar, R. Saha, *et al.*, "A privacy preserving distributed ledger framework for global human resource record management: The blockchain aspect," *IEEE access*, vol. 8, pp. 96 455–96 467, 2020.

[39] ——, "A privacy preserving distributed ledger framework for global human resource record management: The blockchain aspect," *IEEE access*, vol. 8, pp. 96 455–96 467, 2020.

[40] M. S. Krstić and L. J. Krstić, "Hyperledger frameworks with a special focus on hyperledger fabric," *Vojnotehnički glasnik*, vol. 68, no. 3, pp. 639–663, 2020.

[41] M. Moniruzzaman, F. Chowdhury, and M. S. Ferdous, "Examining usability issues in blockchain-based cryptocurrency wallets," in *International Conference on Cyber Security and Computer Science*, Springer, 2020, pp. 631–643.

[42] V. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, 2020, pp. 729–735.

[43] D. J. M. Reddy, S. Regella, and S. R. Seelam, "Recruitment prediction using machine learning," in *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, IEEE, 2020, pp. 1–4.

[44] D. Salah, M. H. Ahmed, and K. ElDahshan, "Blockchain applications in human resources management: Opportunities and challenges," *Proceedings of the Evaluation and Assessment in Software Engineering*, pp. 383–389, 2020.

[45] M. N. Shahid, "A cross-disciplinary review of blockchain research trends and methodologies: Topic modeling approach," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[46] E. B. Sifah, H. Xia, C. N. A. Cobblah, Q. Xia, J. Gao, and X. Du, "Bempas: A decentralized employee performance assessment system based on blockchain for smart city governance," *IEEE Access*, vol. 8, pp. 99 528–99 539, 2020.

[47] M. Vashistha and F. A. Barbhuiya, "Document management system using blockchain and inter planetary file system," in *Proceedings of the 2nd ACM International Symposium on Blockchain and Secure Critical Infrastructure*, 2020, pp. 212–213.

[48] L. Wei, J. Wu, and C. Long, "A hierarchical trust management architecture based on blockchain for crossover service," in *Proceedings of the 2020 The 2nd International Conference on Blockchain Technology*, 2020, pp. 155–159.

[49] S. Zhang, S. Hua, B. Pi, J. Sun, K. Yamashita, and Y. Nomura, "Performance diagnosis and optimization for hyperledger fabric," in *2020 2nd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*, IEEE, 2020, pp. 210–211.

[50] ——, "Performance diagnosis and optimization for hyperledger fabric," in *2020 2nd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*, IEEE, 2020, pp. 210–211.

[51] D. Khan, L. T. Jung, and M. A. Hashmani, "Systematic literature review of challenges in blockchain scalability," *Applied Sciences*, vol. 11, no. 20, p. 9372, 2021.

[52] S. N. Khan, F. Loukil, C. Ghedira-Guegan, E. Benkhelifa, and A. Bani-Hani, "Blockchain smart contracts: Applications, challenges, and future trends," *Peer-to-peer Networking and Applications*, pp. 1–25, 2021.

[53] G. Mariammal, A. Suruliandi, S. Raja, and E. Poongothai, "Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modified recursive feature elimination technique with various classifiers," *IEEE Transactions on Computational Social Systems*, 2021.

[54] A. Sharma and M. Kalra, "A blockchain based approach for improving transparency and traceability in silk production and marketing," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1998, 2021, p. 012 013.

[55] A. Sharma, A. S. Chudhey, and M. Singh, "Divorce case prediction using machine learning algorithms," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, IEEE, 2021, pp. 214–219.

[56] Z. Wang, H. Jin, W. Dai, K.-K. R. Choo, and D. Zou, "Ethereum smart contract security research: Survey and future research opportunities," *Frontiers of Computer Science*, vol. 15, no. 2, pp. 1–18, 2021.

[57] Y. Zhou, X. Song, and M. Zhou, "Supply chain fraud prediction based on xgboost method," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, IEEE, 2021, pp. 539–542.