Authorization Statement

I Dareen Ali Mohammed AbuRabi'e authorize Isra University to provide hard copies or soft copies of my thesis to libraries, institutions or individuals upon their request.

Name: Dareen Ali Mohammed AbuRabi'e

Signature: 

Data: 13-9-2021

إقرار تفويض

أنا دارين علي محمد ابو ربيع ، أفوض جامعة الاسراء للدراسات العليا بتزويد نسخ من رسالتي ورقياً وإلكترونياً للمكتبات أو المنظمات أو الهيئات والمؤسسات المعنية بالأبحاث والدراسات العليا عند طلبها.

الاسم : دارين علي محمد ابو ربيع 

التاريخ: 13-9-2021

Employing Artificial Intelligence And Data Mining for Smart Staff Recruitment

Prepared by

Dareen Ali Mohammad AbuRabi'e

Supervised by

Prof.Mohammad Al-Fayuomi

Co_Supervised by

Dr. Ahmed Bani Mustafa

A Thesis

Submitted to Faculty of Information Technology as a Partial Fulfillment of the Requirement for Master Degree in Software Engineering

August 2021

جامعة الإسراء
ISRA UNIVERSITY

This Thesis (Employing Artificial Intelligence & Data Mining for Smart Staff Recruitment)

Examination Committee	**Signature**
Dr. Mohammad Al-Fayuomi Prof. of Software Engineering	
Dr. Mohammad Saraireh Prof. of Computer Engineering	
Dr. Ahmed Bani Mustafa Assis.Prof. of Data Science and Software Engineering	
Dr. Saleh Abu-Soud Prof. of Computer Science (Princess Sumaya University for Technology)	

DEDICATION

I thank God who gave me the opportunity to gain more knowledge and facilitated all things for me ….

To My beloved family, the symbol of love and giving, and who supported me throughout my entire life, especially my mom for her constant prayers ….

To my friends who encouraged and supported me ….

Finally,

To all my loved ones, without whose love, encouragement, and support, this and

very much more, would have never been accomplished

I dedicate this success to all people in my life who have touched my heart, all those who believed in me. My gratitude is beyond words.....

ACKNOWLEDGEMENT

All thanks and appreciation go to the supervisors: Prof. Mohamed El-Fayoumi and Dr. Ahmed Bani Mustafa, for their support and guidance.

I would like also to thank everyone who provided me with help during the writing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	Natural Language Process
DoS	Department of Statistics
KNN	K-Nearest Neighbours
AI	Artificial Intelligence
ML	Machine Learning
DM	Data Mining
ANN	Artificial Neural Network
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive
SVM	Support Vector Machine
NB	Naïve Bayes
HR	Human resources
CRISP-DM	Cross-Industry Standard Process for Data Mining

Employing Artificial Intelligence And Data Mining for Smart Staff Recruitment

Prepared by

Dareen Ali Mohammad AbuRabi'e

Supervised by

Prof.Mohammad Al-Fayuomi

Co_Supervised by

Dr. Ahmed Bani Mustafa

Abstract

Recruiting staff is one of the most difficult and important decisions to be made by the management. Hiring the wrong candidate would lead to losing valuable potential employees for that may lead to wasting organization resources, profit, and reputation. It may also expose the employer to troubles and may lead to legal procedures. In this work, the researcher proposes an intelligent approach for staff recruitment that employs machine learning, data mining, text mining and natural language processing (NLP) for performing smart staff recruitment. This work aims at enabling employers to utilize artificial intelligence techniques to perform unbiased, efficient, and smart automated recruitment of the best candidates which would help the organization to guarantee growth and prosperity. The proposed approach involves employing data mining for finding the most important predictors of successful staff performance using the organization's historical data. A job specification is then automatically generated. It includes recruitment criteria based on the identified predictors. Text mining and natural language processing are then applied to match the candidate's CV to the job specification to screen and shortlist candidates. The proposed system was applied to a dataset that was acquired from the Jordanian Department of Statistics (JDOS) which consists of profiles of 529 employees that contain 19 features. The dataset was used for constructing 27 models that were generated in three experiments and used nine machine learning algorithms. The best performance was achieved using the K-Nearest Neighbours (KNN) which scored 91% classification accuracy, Random Forest with 89% classification accuracy, and Random Committee 86%. The results were excellent and were also better than most of the results that were reported in similar studies. As for the results of CVs matching, the performance achieved was 80% using the random forests algorithm.

CHAPTER 1: INTRODUCTION

An Organization's vision and goals are usually achieved by good employees. Most managers consider that competent staffs are the organization’s most precious resource. The employee's ability and competence to accomplish tasks with efficiency and quality help to achieve sustainable growth for most organizations, which makes selecting the right employee a very crucial task. One of the most important factors for successful organizations is recruiting the right employees who are both competent and successful. However, many organizations still depend on the traditional approach for selecting employees, which involves the manual screening of all applicants’ CVs to create a shortlist of candidates who are then invited to interview.

The manual approach typically leads to the overwhelming effort exerted by human resources department (HR) and recruitment panels with job applications and CVs which consume a lot of their time and also the time of job applicants and candidates. In addition, the traditional method causes errors and bias which might be driven by personal relationships and might be influenced by the recruiting panel’s technical background, experience, preferences, and personality. It might also be influenced by their ethnicity, race, religion, gender, origin, and nationality.

Moreover, recruitment of the wrong or incompetent staff ia costly as it will cause damage to the employer's business, reputation, profit and may also lead to legal procedures.

Replacing and sacking staff also may cause even more problems and be of a high cost as employees are usually protected under the law and the terms of their contracts. Therefore, it is necessary to focus on selecting the right staff from the beginning.

To ensure the success of any company, it is necessary to focus on defining a clear and transparent mechanism for employee’s recruitment that match candidates’ profile to the

vacant job specifications. This mechanism would both avoid bias and ensure recruiting the right and potentially successful employees.

This thesis suggests utilizing the fourth industrial revolution for proposing an automated artificial intelligence approach that uses data mining, machine learning, text mining, and natural language processing to tackle the issues of the traditional recruitment process. This model would save effort and time, avoid bias and ensure the recruitment of the right candidates based on the organization's historical HR data that include the profiles and the performance indicators of all previous and existing staff who have similar job titles.

The proposed model consists of two steps: (1) the first is the useof data mining and machine learning to identify the most important factors that contribute to the success of the existing employees by analyzing their employment profiles and performance indicators. These factors are then used for deciding on an unbiased set of criteria that can be used for writing the requirements which are then used as a base for creating and announcing job specifications; (2) The second component uses text mining and (NLP) to screen resumes and select the best potential candidates by assessing their possible fulfillment of the job specifications and selection criteria.

This thesis will contribute to the development of human resource management in general and to improving staff selection and recruitment process, which could save cost, time, and effort and would also avoid bias that which might affect efficiency and validity of the recruitment process.

1.1 Research hypothesis

This thesis is designed to assess the hypothesis that:

" Data mining of the profiles of existing staff can be used for identifying the performance predictors of successful employees and for matching and hiring the best candidates ".

1.2 Research Question

In this research, the question is: can anyone avoid bias in resume screening and recruitment process to enable the employer to hire the best candidate(s) who both fits the vacant job specifications and is/ are expected to excel in the job that they are recruited for?

1.3 Research Aim and Objective

This thesis aims to enable employers to unbiasedly select and recruit the right candidates for the vacant position who are predicted to be both competent and successful to guarantee growth and prosperity for the employer business and improving the accuracy of writing the job specification, and improve the methods of selecting the appropriate employee for the advertised job specification, more specifically.

To achieve this aim, the researcher will:

- enhance the method of writing a job specification.
- set automated methods to select the best candidates in a short time, and set evaluation criteria to evaluate all resumes in the same way without bias.
- Reinforce accuracy and performance for matching resume with the job specification.

1.4 Research Outlines

This thesis is organized into seven chapters. Chapter 1 provides an introduction to the work covering its context, motivation, essence, and a brief overview of its applied methodology. Chapter 2 provides relevant that are related to the thesis domain and a critical analysis of the related work. Chapter 3 presents what is data mining, Chapter 4 presents what is artificial intelligence and machine learning. Chapter 5 presents what is classification. And presents the natural language processing in chapter 6. Chapter 7 presents general view of hr recruitment. In Chapter 8 proposed methodology, and chapter 9 presents dataset. Chapter 10 presents the results of the work that is applied using three experiments, while Chapter 11 provides a discussion of the obtained results. Finally, in Chapter 12, we provide a conclusion for the current research and recommendations for future work.

CHAPTER 2: LITERATURE REVIEW

This chapter introduces a set of literature scientific researches on the previous studies that are most relevant to our scientific research and that has been collected from many scientific sources and studied and summarized which how to extract the most important variables that affect employee performance and determine the appropriate CV.

In the project management community, studying and analyzing the requirements is important, in this research the important studying is to determine the correct job specification and matching the right resumes with it, so job requirements analysis is to come up with a realistic specification of the job and the specifications that should be in the employee. Job specification refers to a written description, in which each of the job requirements is described, by describing basic information about the job and including the job title, the general definition of the job, detailed data about the tasks, duties, and detailed tools used in that job. After preparing the job description, the job will be announced, so that many candidates apply to fill this job vacancy, after which the process of filtering the candidates who match the job specification will take place.

Regarding the job specification as shown in the work of (Wyse 2019) is a textual narration of each required job and contains tasks and an accurate specification of duties. the way of specification varies from one company to another, but the most important thing in determining qualifications and skills and characteristics, that It must be available in an applicant of vacancy. And After announcing jobs and publishing job specifications, many applicants flow and sometimes exceed 100 applicants, and to form a coordinated team able to meet the requirements at the right time at an affordable cost, it is necessary to choose the right person for the project in terms of experience, skills, or ability to learn.

The book of (Coverdill and Finlay 2017) discuss the job specifications is very important because include the requirement of the job, so if it is not clear or wrong lead to recruited

not suitable employees, job specifications summarized what the company needs from requirements of educational, skill and so else.

As what relates to research in the previous work of human resource requirements and create job specification and the mechanism of recruiting staff accurately and deal with data and choosing a feature as a broad topic, we found a set of research work that presented methods for the most appropriate human resource allocation process of recruitment that we can mention here, Below is a summary of the latest research for last 10 years.

In reference to (Lee and Han 2008), their paper depicts the most recent skills and requirements for programmers/ analysts of the Information Systems sector. The data are collected from 500 websites where advertisers expect that programmers/ analysts are able to perform the roles of computer program writers besides the technical tasks. The findings of this study illustrate that the top-priority skills of programmers/ analysts are development, software, social skills, and business. However, the skills of architecture, network, hardware, management, and problem-solving are less considered.

Rodriguez & Chavez (Rodriguez and Chavez 2019) conduct a study to explore selected information from some resumes after adopting a clustering algorithm to match the profile of the applicants against the requirements of the vacancies. The main objectives of the study can be summarized as follows: To provide an additional training data set from candidates and employers, conduct tests of the clustering model to examine reliability and performance of the job matching system, and perform an end-user software evaluation. This simplifies the employers' mission to find appropriate candidates for a particular job in different companies and make more informed decisions. The data is collected from the 2,283 candidates' resumes in the Philippines. To analyze data collected, analytic software known as WEKA is used. Findings of the study display that the main step to determine

the best fit between applicants and job description is establishing the job and company profiles. It is concluded that the attributes Job title, work experience, educational attainment, and civil status have the same rank and significance, whereas age and gender are less significant attributes

The work proposed by (Tsai, Moskowitz and Lee 2003) , to select appropriate human resources and determine the relationship between tasks and human resources and identify appropriate human resources to constitute the project team, using a diagram of critical resource (CRD), and the Parameter design implementation using the Taguchi method. The tasks are difficult to predict and cannot be controlled, but resources can be controlled. This method has achieved good results, but still needs to improve a balance between accuracy computational and efficiency, measures the duration and quality, and applies this method to multiple projects.

As for identifying features, The work in (Liu and Yu 2005) Presented a survey to help researchers how to use algorithms for clustering and classification, the ways for selecting important attributes:

- Approach based on the filter: depend on data intrinsic characteristics.
- Approach based on the wrapper: depend on training a learner through a procession of selection.
- Hybrid model: depend on combine filter and wrapper.

But the filter is faster and the best from the computational side.

But The work of (Salappa, Doumpos and Zopounidis 2007) presented analyzing efficiency and performance of FS algorithms, SVM proved effective in terms of efficiency in identifying features; however, the author recommended that relying on one algorithm is not preferred.

And The work in (Khoshgoftaar, Gao and Seliya 2010) about pre-process of data by feature selection to select attributes that are important in the dataset and take data sample to fix imbalance data, the author compared between following:

1. Modelling and feature selection from the original data.
2. Modelling from selecting the feature from the sample and the original data.
3. Modelling from the sample and selecting the feature from the original data.
4. Modelling and choose features from the sample.

The result is the best performance is to choose the feature from the sample but to predict the model there are similarities in both cases.

The work in (Chien and Chen 2008) presented an approach to analyzing applicants' data to predict their performance, using association rules and the decision tree By exploring the rules linking the demographic characteristics of applicants (profile attributes) and performance that aim to build a hiring strategy, the recommendation this work recommended is using different variables such as school rate, address, and the number of certificates approved and recommended to use the technique of the neural network to detect hidden relationships between variables.

The work (Strohmeier and Piazza 2015) discussed in his book about the role of AI in HR management and put forward various ideas, including the use of artificial neural networks in turnover prediction, also, his book summarized and discussed the function of search engines in searching for candidates for jobs vacant, and discussed, based on the exploration and foundation.

The approach of (Courtin and Tomasena 2016) presented trace-based for analyzing activity to define the work teams according to the competencies, this work was a case study of defining a Java programming workgroup based on the competence of team members using the textual analysis tool and graphical analysis tools. This work

recommends implementing a recommender system to create appropriate links and focus on the accuracy of competencies.

The work in (Chen, Zhang and Niu 2018) propose a novel algorithm for extracting CV data by split texts of CV into several text blocks, punctuation index, words, design syntax information for the new sentence to each row in cv, to get organizational data to select information, the system gets CVS then built ontology features of candidates, in the second step the ontology represented job requirement, the last step the system maps first step into the second step and select suitable candidates, Precision= 0.92, Recall=0.75, *F*-1=0.82, the author recommended in future to use other algorithms of classification and improve recall rate.

The work in (Paredes-Valverde, et al. 2018) presents the ontologies to build a decision support system for selecting the best employees to work on a new project. Ontologies discovered the semantic relatedness between new and previous software projects the requirement specification. so the system can, suggests those employees who participated in a similar project. the approach proved its effectiveness by evaluating a software development organization. The model can recommend relevant HR who participated in a similar project. But there are some projects that the system does not suggest, based on the results, the accuracy based on increasing the number of similar projects, and the more the number of similar projects, the more accurate the system recovery for the most suitable employees, Some contributions focused on insufficient variables in identifying the best CVs and ignoring other important variables, and the results of other contributions were that the accuracy of the chosen CVs was not 100%, and the system ignored the appropriate CVs. Like as this work. The percentage of accuracy is 0.7382, a percentage that is inaccurate because it depends on the proportion of similar projects. Therefore, there is still an urgent need to suggest designs whose results are close to 100%, to reduce error.

The work of (Yahiaoui, et al. 2019) presented a tool that helps discover the right skilled people from User profile database (SQL DB Server) based on activity analysis, Design their own competencies repository, and update the database. by retrieving the Web information through a competence repository, a web scraping tool, and the Gephi of graphical analysis tool, to extract appropriate information, the work focused to developing process of evaluation.

The model that was proposed in (Ajoudanian and Abadeh 2019) is a novel fuzzy C-means(CF) method clustering and using correlation such as PCC applied on GitHub data for selecting similar projects on GitHub, is used to recommend appropriate human resources that participated in similar projects by utilizing a logarithm of subgraph detection to determine initial centers of the clustering method and to use correlation measures. The precision was 0.853. This method can be enhanced by using various dense clustering and parameter methods to obtain new results and analyze them.

The approach proposed by (Georgiou, Gouras and Nikolaou 2019) for building a new gamified assessment method to select employees by the methodology of the situational judgment test (SJT) for predicting relevant behaviors and selecting employees. SJT is a gamified evaluation to determine soft skill. It is Evaluates four main soft skills which are: flexiblilit, Making a decision, Adaptation and Resilience. The limitation of this approach is the sample size was 321 applicants to take the test. The number should be morewith gamified assessment methods.

The work of (Necula and Strîmbei 2019) presented an approach to employ Semantic Web technologies and SPARQL queries and data science to obtain semantic data to determine the skills by applying KNN, classification via regression, and random forest, NB, SVM. It used the decision tree algorithms to résumé data described previously, with terms from the ontologies. The researcher used the real website to get data with HTML

online format. In the future, authors recommend predicting the team's success rate based on the variables chosen.

The approach proposed by (Acikgoz 2019) developed a model to present the relationship between job search activities and applicant characteristics, organizational characteristics, and activity of recruitment. The model acknowledges that the hiring process is not linear and does not end when applying for a job, but rather progresses dynamically until job seekers reach their goals. The model combines factors at the individual level with the organizational level to attain the goal of employment. The work recommends testing the model on several levels.

The work of (Brandão, Silva and dos Santos 2019) explores the online recruitment tool, it is a quick method to reach the audience, attract the most efficient applicants to provid information on select candidates and recruitment processes.

The recommendation in this approach for developing online recruitment and deep understanding of candidate files and the changes that are being made to them.

The book of (Bondielli and Marcelloni 2019) presented Applicant Tracking Systems (ATSs), a data-driven model for selecting a candidate automatically, using text mining and N|LP and building a semantic representation, an entropy to measure weight for relevance. This model is based on a controlled strategy and on data. However, this approach still needs to be improved and to conduct more experiments and suggestion relying on a strategy that is not controlled supervision.

The work of (Roy, Chowdhary and Bhatia 2020) presented a system for recommending a resume, to categorize the correct categories according to job specifications using k-NN and cosine similarity. The approach developed an automated resume recommendation system that used a Linear SVM Machine learning classifier to extract relevant resumes depending on similarity with an accuracy of 78.53%, and 0.3899 with Random Forest. This approach used n-grams to apply Text classification, NER, and NLP and used

distance-metric classification. The dataset in this work is in Excel format from Kaggle. They recommended using PDF format in the future and to enhance the model by using deep learning such as Neural Network.

The work of (El Mohadab, Bouikhalene and Safi 2020) proposed a system that used neural language processing and Decision Tree classifier with filter String to Word Vector, to predict information of researcher. The classifier presents good results Precision is 0.943, Recall is 0.971 and F-measure is 0.957. The work used a dataset of scientific researchers from the Postgresql database that contains all the records of doctoral students at the university.

The approach proposed by (Buil, Catalán and Martínez 2020) explores the results of applicants to a recruitment tool through drawing on TAM and SDT that focuses on a functional term in gamified recruitment processes. How applicants' replay on the recruitment processes is more suitable and HR managers learn how they can use these tools, by using the partial small squares structural equation model.

The authors recommended using tools to understand the effect gamified in the recruitment processes. The registration to participate was through a website by a questionnaire in a competition for selecting new graduate talent. The work contributes to gamified recruitment literature using partial small squares for structural equation modeling, but it has limitations such as its focus on autonomous motivation, through there are many types of motivation.

Obviously, the previous works did not focus to generating job specifications in any way. They also didn't the use PDF CVS files and random forests with NLP to Classify the right CVS. This means that the current work will contribute process methods that were not used before. Below is table 1 which summarizes of previous works from newest to oldest:

Table 1: Summary of Related Works

Related work	Input	Techniques	Output
Tsai, Moskowitz, & Lee, 2003	available candidates	- using a critical resource diagram (CRD) to determine the relationship between tasks and human resources, - implemented the Parameter design using the Taguchi method to identify appropriate human resources.	constitute the project team
(Lee and Han 2008)	The data were collected from 500 websites	Content analysis	the top-priority skills of programmers/ analysts are development, software, social skills, and business. However, the skills of architecture, network, hardware, management, and problem-solving are less considered.
Chien & Chen, 2008	employee profiles	using association rules and decision tree	talents are the most suitable
Strohmeier & Piazza, 2015	Search by search engines	neural networks, use knowledge-based search engines	Turnover prediction employee
Paredes-Valverde, del Pilar Salas-Zárate, Colomo-Palacios, Gómez-Berbís, & Valencia-García, 2018	-SRS documents - employees' profiles	ontologies	- select the best employees to participate in a new software project. - suggest the employees who participated in a similar project - accuracy is 0.7382%
Chen, Zhang, & Niu, 2018	Cv file	-propose a novel algorithm for extracting CV data. - split texts of CV into several text blocks, punctuation index, words, design syntax information for the new sentence to each row in cv. -build ontology features of candidates	select suitable candidates, Precision= 0.92%, Recall=0.75%, *F*-1=0.82%
Rodriguez and Chavez 2019	The data is collected from the 2,283 candidates' resumes in the Philippines	adopting a clustering algorithm	It concluded that the attributes of Job title, work experience, educational attainment, and civil status have the same rank and significance, whereas age and gender are less significant attributes
Georgiou, Gouras, & Nikolaou, 2019		a new gamified assessment method to select employees by methodology of the situational judgment test (SJT). SJT to gamified Evaluation to determine skill soft,	prediction relevant behaviours and select employee
Necula & Strîmbei, 2019	data from Indeed website html online format	Semantic Web technologies, SPARQL queries, résumés data described with ontologies terms	identify skills suitable for job, with accuracy

Ajoudanian & Abadeh, 2019	GitHub data, GHTorrent data set in a MySQL database.	novel fuzzy C-means clustering(CF), by utilizing an algorithm of sub-graph detection in determining initial centers of the clustering method and the use of correlation measures	recommend appropriate human resources that participated in similar projects with a precision of 0.853
Yahiaoui, Courtin, Maret, & Tabourot, 2019	User profile database (SQL DB Server)	tool that helps discover the right skilled people by retrieving information from the Web by the Gephi graphical analysis tools, a competence repositor, and a web scraping tool	extract the right skilled people
Roy, howdhary, & Bhatia, 2020	CVS Excel format from Kaggle	Using Linear SVM Machine learning classifier, n-grams to apply Text classification, NER, and NLP and used distance-metric classification	similarity with accuracy of 78.53% with Linear Support Vector Machine Classifier , and 0.3899 with Random Forest
ElMohadab, Bouikhalene, & Safi, 2020	dataset of scientific researchers from the Postgresql database that contains all the records of a doctoral student at the university	Using neural language processing and Decision Tree classifier	Precision is 0.943, Recall is 0.971 and F-measure is 0.957.
Buil, Catalán, & Martínez, 2020	Participation was through a website by a questionnaire to participate in a competition and to select new graduate talent	Tools of gamified recruitment by drawing on TAM and SDT and trial support that focuses on a functional term in gamified recruitment processes method, by Using the partial small squares structural equation modelling.	To participate in a competition and to select new graduate talent

CHAPTER 3: Data Mining(DM)

DM is the process of extracting implicit and potentially useful knowledge from data, which involves the exploration and analysis of the target dataset, to find patterns and meaningful knowledge that is hidden in data using machined learning algorithms that are used to uncover hidden trends, patterns, and relationships. DM aims at performing either descriptive or predictive tasks which can be classified into five main tasks: (1) Classification; (2) Clustering; (3) Regression; (4) Association; and (5) Features Extraction. (Chung and Gray 1999).

The DM process involves performing many sequential and iterative phases that differ according to the process model applied. The popular Cross-Industry Standard Process for Data Mining (CRISP-DM) process model consists of six phases (Chapman, et al. 2000), while newer process models such as MeKDDaM consist of eleven phases (Banimustafa and Hardy 2020). In this work. CRISP-DM was used due to its simplicity and generic applicability, and figure 1 illustrate the DM steps :

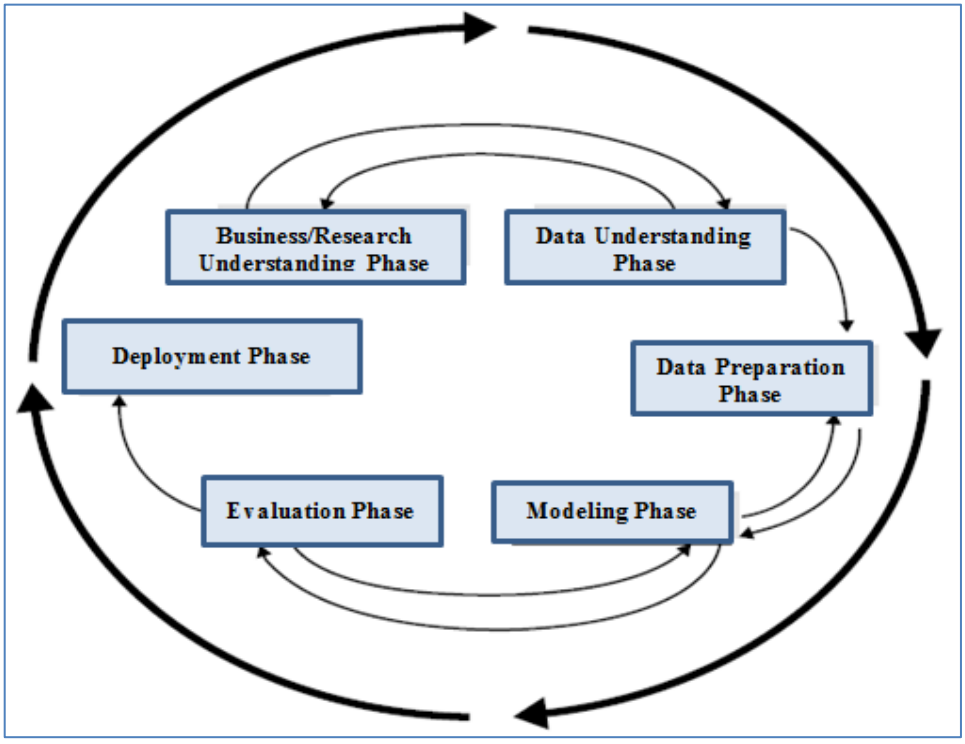


Figure 1: Data mining phases

The CRISP-DM consists of six phases that occur in a cyclical process:

1. **Business Understanding Phase:**

Understand the business or research, what is the project goal, and what are its requirements and for exploration problems (Chapman, et al. 2000).

2. **Data Understanding Phase:**

The data that is being worked on may be incomprehensible to people who are not experts in the field or who work in Data mining, so the data should be well understood and quality. After collecting data, recognizing data, and evaluating it is preferable to take a sample that has multi-pattern for exploration and analysis it by statistical analysis. Data exploration provides insight into the data that will be used for building the data mining model (BaniMustafa and Hardy 2011).

3. **Data preparation phase:**

Preparing and Pre-processing the data to be used, determining the variables to be analyzed, and is an important stage for obtaining a correct analysis.it includes below steps such as:

- **Dataset Pre-Processing**: This approach inputs the staff profiles from historical data. Data processing is an important step for building a good machine learning model. The raw data had some problems that must solve it, such as missing data, outliers, selected features.
- **Exploration:** This part explores the data set and analyzes it to understand it and extract knowledge from it. has been used Descriptive Statistics which helps to describe the characteristics of JDoS.
- **Balanced datasets**: An unbalanced data set is a problem that affects classification. It is data, in which the distribution of categories is uneven, in which the number of one category is much higher than another.

Processing unbalanced data is very important to help build models with high accuracy, and if the distribution is balanced, the result is better. There are a lot of methods for solving the problem of unbalance in the data, such as under sample, over sample, and re-sampling.

Resampling techniques use random states, with replacing Sample, by adding samples from the lower category, and remove samples from the higher category, to be more balanced. In other words, resample method create subsample randomly by replacement (Charte, et al. 2015)

- **Feature selection:** The most important step for building a ML model is Feature selection, Sometimes the dataset contains unimportant information, which causes an increase in training time and needs more storage capacity. so when choosing important features, it contributes to the speed of training the model and improves its performance, and help to solve problems of classification, and better the accuracy of the model (Gholami, Pourpanah and Wang 2020).

4. **Modeling phase:**

   A mining model is created by applying an algorithm to data, but it is more than an algorithm or a metadata container: it is a set of data, statistics, and patterns that can be applied to new data to generate predictions and make inferences about relationships.

5. **Evaluation phase:**

This step will determine the degree to which the resulting model meets the business requirements. Any model built and trained needs to be evaluated to know the accuracy and evaluation of the effectiveness and quality of the model, and determining the criterion for choosing to measure the performance of the classifier is one of the

important things in evaluating performance. There are several criteria important measures for evaluation such as error rate, ROC curve, precision and recall (Flach 2003) .

- **Cross-Validation:** It is a tool or technique for evaluating the model that was built for a dataset to use in the future dataset, by dividing the data into a training part, which is the part that the algorithm is applied to, and the other part is the test part through which the error rate is calculated (Wong 2015).

  In k-fold cross validation, the dataset divided randomly, into k with equal sub size datasets.one of the sub-datasets(k) is used to test the model, and other sub-datasets(k), k-1 are used to train data.

  From the k datasets, a single sub dataset is kept to validate data for testing the model, and the remaining k-1 sub-datasets are used as training data. then repeat the folds k times(cross-validation) with each of the k sub-datasets used once as validation data. The k effects from the folds can then be averaged (or otherwise combined) to provide a single estimation.

  The benefit of this method is that each observation is used for both training and validation, and every comment is used for validation precisely once.

- **Confusion Matrix:** It is an array of two dimensions, (actual and expected), which shows the performance of the algorithm. The row represents cases in an expected class, The column represents cases in the actual class the Instances was Correctly Classified is the sum of instances, which the classifier correctly categorized, and Incorrectly Classified Instances is the sum of instances, which the classifier classified it incorrectly categorized, in other words it is indicate the numbers of predictions of the model and if it classified the classes correctly or incorrectly (Ruuska, et al. 2018).

The matrix helps discover if the model has wrong classified, and show number of True Positive (TP) which a model positive prediction and it's true, False Negative(FN) which a model negative prediction and it's false, False Positive (FP) which a model positive prediction and it's false, True Negative (TN) which a model negative prediction and it's true, P = positive, N = Negative, As shown in Figure 2 :

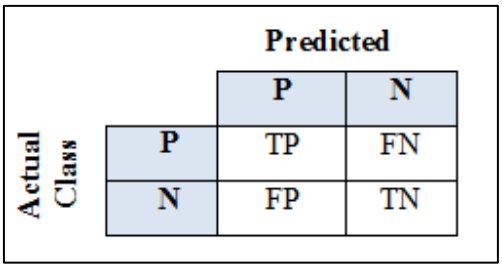


Figure 2: Confusion Matrix
(Luque, et al. 2019)

- **Classification Rate (accuracy):** It is considered the most important measure to evaluate performance classifiers, it represents the number of successfully categorized data to the total number of categorizations (García, Luengo and Herrera 2015), it calculates based on the following formula Equation 1 (Luque, et al. 2019)**:**

$$(T\,P+\,T\,N)\,/\,(T\,P+\,FN+\,T\,N+\,FP) \qquad (1)$$

- **Precision:** It represents how close the measured values are to other values, it calculates based on the following formula equation 2 (García, Luengo and Herrera 2015):

$$TP/\,(TP\,+\,FP) \qquad (2)$$

- **ROC :** The receiver operating characteristics, It is a numerical and optical scale to assess the performance of a workbook to ensure data quality, as it provides a summary of behavior Predictor, is a representation that true positive

rate (TPR) against false positive rate (FPR) through drawing. TPR known recall or sensitivity, or probability, and FPR known specificity or probability of false alarm equations 3,4 (Verma 2019) :

$$TPR = TP / (TP + FN) \quad (3)$$

$$FPR = FP / (FP + TN) \quad (4)$$

The learner tries to choose a sample in which there is a high percentage of pros from the test cases. In the vertical axis drawing, the ratio of the number of positive cases from the sample relative to the total positive number, while the horizontal axis represents the ratio of the number of negative cases from the sample relative to the total negative number (Flach 2003).The graph shows the rating for the test and each of the rating points for the curve predicts the accuracy of the test (Witten and Frank 2002), (Vuk and Curk 2006) .

- **Recall:** It's a measure How much did we correctly expect from all the positive categories in the test dataset (Verma 2019), And it is calculated with equation 5:

$$Recall = TP / (TP + FN) \quad (5)$$

6. Deployment phase:

The concept of deployment in data science refers to the application of a model for prediction using a new data. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data science process, or extracting data in another part (Chapman, et al. 2000).

CHAPTER 4: ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Artificial Intelligence (AI) is how to make a computer think like a human being, and the ability to memorize, it is the characteristics of hardware and software and their ability to simulate the human mind, work to deduce the various problems, and find appropriate solutions based on previous experiences. in 1940 s, the term first appeared the first time and in 1956, the first workshop on artificial intelligence research was held at Dartmouth College, and the field of artificial intelligence research was founded as an academic discipline. (Turing and Haugeland 1950).

Machine learning (ML) is a specific field within the AI domain. It is applied of AI concepts and within the intersection of statistics and computer science, it is a science to teach a computer and programs to act like humans, through learning from data and train and develop smart software and systems, where the system is trained and response is expected after the process of displaying examples of input and output behavior by providing them with information and data and their interaction, thus increasing their ability to learn with time better, that's mean learn from data and predict and make decisions, it depends on algorithms and data availability (Does machine learning really work? 1997).

Machine learning methods can be applied in various fields: health education, economics, human resources, manufacturing, marketing, etc. such as robotics education, vehicle control automatically, computer vision, speech recognition, word, and natural language processing, Neuroscience research, etc. (Jordan and Mitchell 2015).

There are several learning algorithms to teach a model and it has four types as below:

- Supervised learning: also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes

accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids overfitting or underfitting. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and more (Jiang, Gradus and Rosellini 2020).

- Unsupervised learning: uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, image and pattern recognition. It's also used to reduce the number of features in a model through the process of dimensionality reduction; principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, probabilistic clustering methods, and more (Gentleman and Carey 2008).
- Reinforcement learning: is a behavioral machine learning model that is similar to supervised learning, but the algorithm isn't trained using sample data. This model learns as it goes by using trial and error. A sequence of successful outcomes will be reinforced to develop the best recommendation or policy for a given problem (Kaelbling, Littman and Moore 1996).

  The 2 that won the Jeopardy! challenge in 2011 makes a good example. The system used reinforcement learning to decide whether to attempt an answer (or

question, as it were), which square to select on the board, and how much to wager—especially on daily doubles.

- Semi-supervised learning: Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of having not enough labeled data (or not being able to afford to label enough data) to train a supervised learning algorithm (Chapelle, Scholkopf and Zien 2009).

CHAPTER 5: CLASSIFICATION

Classification is the act of assigning objects to one of several predefined categories, and used as a tool for predicting values by analyzing data to extract models that accurately describe important categories and classifications of data that help in understanding the data, its use in several applications including performance level prediction and so on, by building a classification model that predicts specific features that describe the category that is explored and clarify the category to which the element belongs to be explored, assign values of class for new instances (Kesavaraj and Sukumaran 2013).

Each record contains target variable information, and the goal of the classification model is to generate a classification for the same target variable, but for new records that are not in the database, this is the work of the classification algorithm, it trains the classification on existing data and then selects the classification of the new data (Larose and Larose 2014).

The process of Classification, Construct model (describe classes which determined previously).

A good classifications models depend on the quality of the training data, prediction is based on past and future information, current information consider a set of features but future information Are set of result, it hasn't category values, (Taymouri, et al. 2021). The process of Classification As shown in Figure 3:

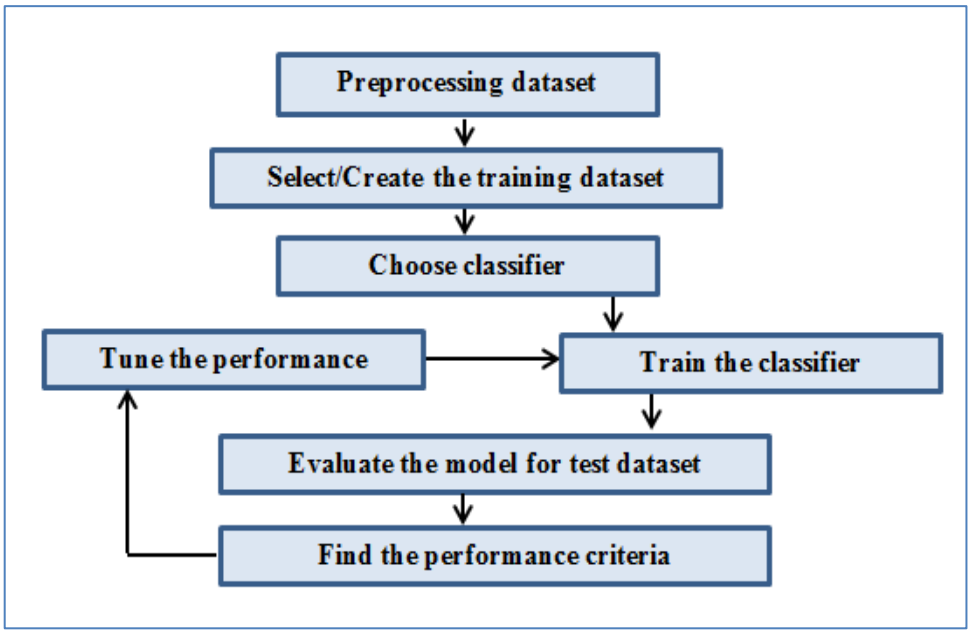


Figure 3: Classification Process (Punjani and Atkotiya 2018)

There are many classification techniques, including the following:

Support Vector Machine (SVM): SVM Suggested and developed by Boser, Guyon, and Vapnik in 1992 (Farhat and Vapnik 1992), it is a supervised ML model for regression and classification and Used for pattern recognition and analysis, it is a method for prediction, based on separate the data into two parts, based on finding the hyperplane or a line separating the two parts, by a quadratic equation, it contains constraint equality and inequality, Accordingly, it is done that draw the hyperplane and the data closest to the line or hyperplane called SVM, and we consider the most important elements in the data set (Meshram, et al. 2020) (Cervantes, et al. 2020).

Random forest: This algorithm was proposed by Tim in 1995 (Ho 1995), and developed by Leo Breman. It is used in classification to build predictive model, Randomness is the basis in this model since the trees will be completely different, it's consists of decisions trees which uncorrelated to each other, while training, By dividing the data, and within each section the model build prediction model.

The tree is more accurate than single trees and its reliable classifiers such as SVM, the importance of this model stems that it does not require expertise or preparation of large

data, it can train data that contain the missing value, unbalanced data, and categorical data Contrary to SVM model (Belgiu and Drăguţ 2016).

K Nearest Neighbor (KNN): KNN is an approach for classifying data, Suggested and developed by Evelyn Fix and Joseph Lawson Hodges Jr in 1951 (Fix and Hodges 1951). KNN is a non-parametric algorithm that means it hasn't any assumptions for data distribution, it separate data points into many categories for the prediction of new point, to classify the record in the dataset, the record was retrieved his k nearest neighbors that select by voting, KNN bias to k value, the best method to select k is run algorithm much time with different k and select the best performance (Deng, et al. 2016).

Naïve Bayes (NB): NB Suggested by Thomas Bayes, and published by his friend Richard Price after Thomas died in 1763 (Bayes 1763). This technique is based on the statistical concept, and it was given this name because it considers assumptions independence, that is, there is independence between the Attributes Features, The technique is highly efficient in the prediction process, especially in big data and needs some knowledge for training (Bellhouse 2004).

J48: It's a classifier algorithm was proposed in 1993 by Ross Quinlan (Quinlan 2014). To build decision trees by the C4.5 algorithm. J48 contain a graph it depends on the divide and conquers method to assign classes of independent instances and test attributes, C4.5 algorithm is consist of four rule, first one is decision rule generator (C4.5), second is production rule generator (C4.5rules), third is Decision tree interpreter (consulter), and finally Production Rule interpreter (consultr), (Quinlan 2014).

CHAPTER 6: NATURAL LANGUAGE PROCESSING

NLP is one of the areas of AI that demonstrates the interaction between human language and the computer, that is, programming the computer to be able to analyze and process natural language data, so that the computer can understand documents, extract and classify information. In the year 1950s It was the beginning of NLP. IN 1960s developed NLP systems (Swanson 1960). NLP is the human talked languages (Chopra, Prashar and Sain 2013).

Here are some important terms used:

- Tokenization: It is a method for dividing document to tokens (phrases, words), that's mean convert text to words and put spaces between words. In English, Documents and texts are tokenized into words by using spaces and removed some words like (a the, was, is) because it is not important meaning (Vijayarani, Janani and others 2016).
- Stop Words: Removing redundant and words such as (a, the, was, is) are important pre-processing for improving performance and accuracy and reduce the time of training.

Steps of processing NLP (Nagarhalli, Vaze and Rana 2021) **:**

1. **Analysis of Morphological and Lexical:** The language dictionary consists of vocabulary, expressions, and words Morphology describes and analyzes a word, but Syntactic splits the text into words.
2. **Syntactic Analysis:** Called Parser also, Study the structure of sentences, it works on word analysis and it makes sure that the input is consistent with the rules of the programming language. To depict the structured grammar of the sentence, also the tokens are extracted in a tree form, after making sure that it is by the rules of the language.

3. **Semantic Analysis:** It is called an Intermediate code generator, using to access a language that the device understands, the outputs of the previous stage tokens are used, and this stage is responsible for making sure that the codes are free from any errors. Its dictionary for exacting from context, In short, analyzes and studies the meaning of the word and the meaning of the language.
4. **Analysis of Pragmatic:** Understand and analyzes the nature of language and its uses, not concerned with the structure of language. Figure 4, illustrated Steps of processing NLP:

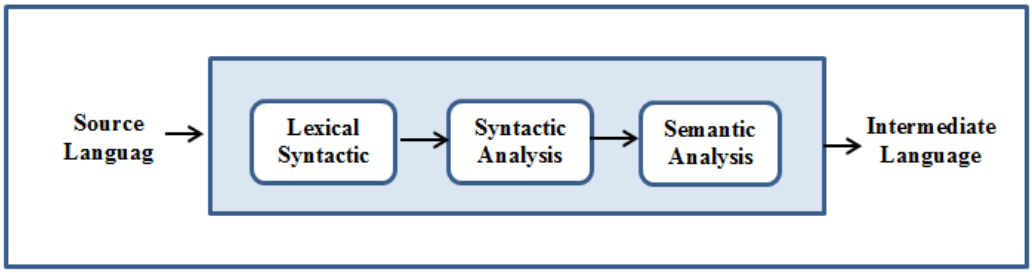


Figure 4: Steps of processing NLP

CHAPTER 7: HUMAN RESOURCES RECRUITMENT

Recruitment is a key responsibility of the HR department. While HR works in many areas including employee engagement, employee development, statutory compliance, data management, and many others, one of the key areas of focus for HR is Recruitment

Recruitment refers to the process of identifying, attracting, interviewing, selecting, hiring, and onboarding employees. In other words, it involves everything from the identification of a staffing need to filling it. Regardless, recruitment typically works in conjunction with, or as a part of Human Resources (LaBerge 1962).

The recruitment process is a set of actions taken by the organization to attract candidates for work who are distinguished, qualified, and able to contribute to achieving the goals of the organization (Visa, Einolander and Vanharanta 2015).

The roles of Human Resources in recruitment are:

1. Identify the hiring need
2. Devise a recruitment plan
3. Write a job spesificatiom
4. Advertise the position
5. Recruit the position
6. Review applications
7. Interviews
8. Applicant assessment
9. Background aheck
10. Decision

The process of appointing people requires a prior and careful study of the job spesificatiom, i.e. the requirements and standards that must be met by the person who will occupy the job (Laurim, et al. 2021).

CHAPTER 8: PROPOSED APPROACH

This chapter presents the methodology and how it is implemented. Tools, package, and libraries relevant to a designing model, will all be explaind in detail.

This methodology is present in figure 5 illustrating how to achieve the declared goal of the research. It consists of four main steps to enhance and enable employers to unbiasedly select and recruit the right candidates for the vacant position. The candidate for that position should be competent and successful to guarantee growth and prosperity the employer's business.

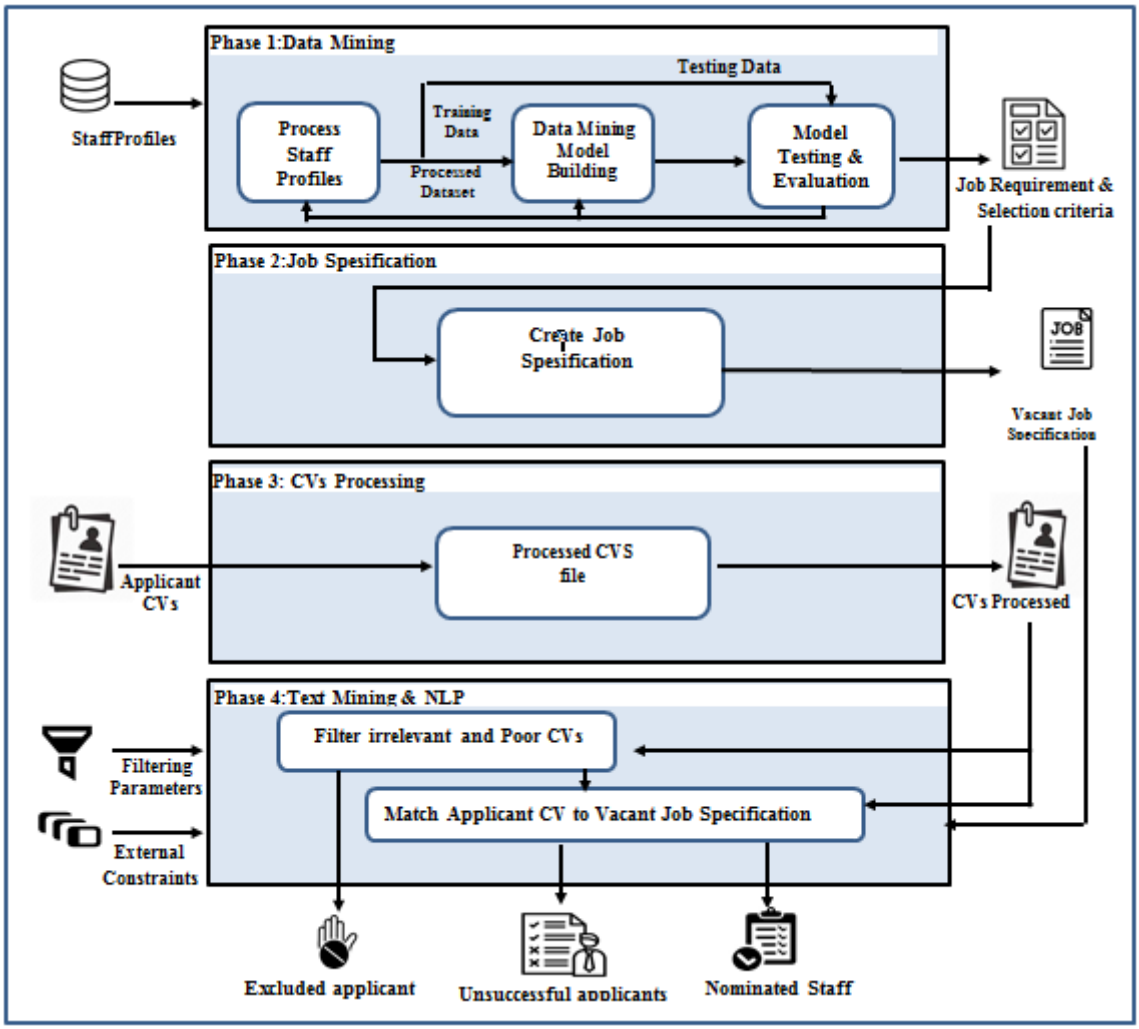


Figure 5: Proposed Approach

First Phase: Data Mining: Mining the existing JDoS staff profiles to identify the most important factors which must be considered and included in job specification for the vacant position. The data mining algorithm achieves this task by ranking and extracting the most important characteristics in the staff profile that contribute more to the prediction of the highest performance of existing staff, particularly those with similar or relevant job titles to the vacant job.

Second Phase: Job Spesification: Job requirement characteristics, and the selection criteria that were predicted to create the job specifications for the are fed to the this phase which involves utilizing these characteristics in preparing the job specification document and its involved selection criteria.

Third Phase: CVs Processing: At this phase, job seekers to submit their CVs, may be in different formats. The CVS are processed by converting them to PDF formats, to be deal with at a later stage.

Final Phase: Text Mining and NLP: In the final phase,After preparing the job specification and its requirements, and candidate submissions, comes the stage of Matching job Specification with Resumes and filtering suitable CVs that closely match the required job description.

The CVs are analyzed, and matched to job specification and selection criteria document using Python. The analysis involves excluding irrelevant and poor CVs and perform text mining and natural language processing algorithms to cut down the number of considered, and to create a shortlist that includes the recommended and the most competent candidates who fit job specification,foreseen to success in their possible employment, and exclude unsuccessful applications and Applicants who do not meet the company's legal requirements.

This phase has 3 elements illustrates shows in Figure 6:

- First element is a large dataset for ML. It is like a word warehouse containing words that are expected to be written in different ways. For example, a bachelor's degree can be written "Bachelor", and this helps to find similar or relevant words in a more accurate way.
- Second element is a dictionary for each job that contains the requirements of each job in detail.
- The third element is a repository of all CVs.

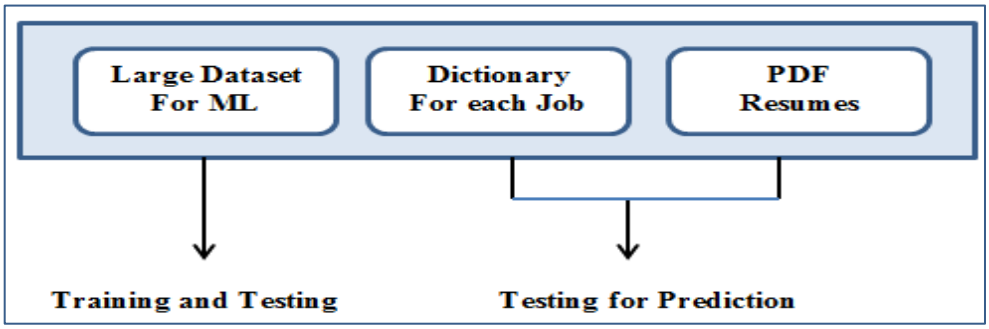


Figure 6 : Elements for Matching PDF Resumes with Job specification

The detailed process of algorithm for matching resumes and job specifications. is as follows; First, each resume is processed as described in Figure 7 described. Second, some noises are removed from the text, like space, line, special characters, after determining the list of stop words, a comparision is mode between words and job title **.**

```
Load PDF File then
For loop
For each pdf
Remove special characters, space, line, numeric,
Words= (word 1, word 2, word 3, word n)
Stop word :( word 1, word 2, word 3, word n)
If job= programmer then
For each word in stop word
If Word= Stop word then
Append list ()
Else
Continue
Stop word
Resumes
```

Figure 7: Algorithm for matching between resumes and job spesification

In the following, Data Flow Figure 8 describe process of Matching PDF Resumes with the job specification:

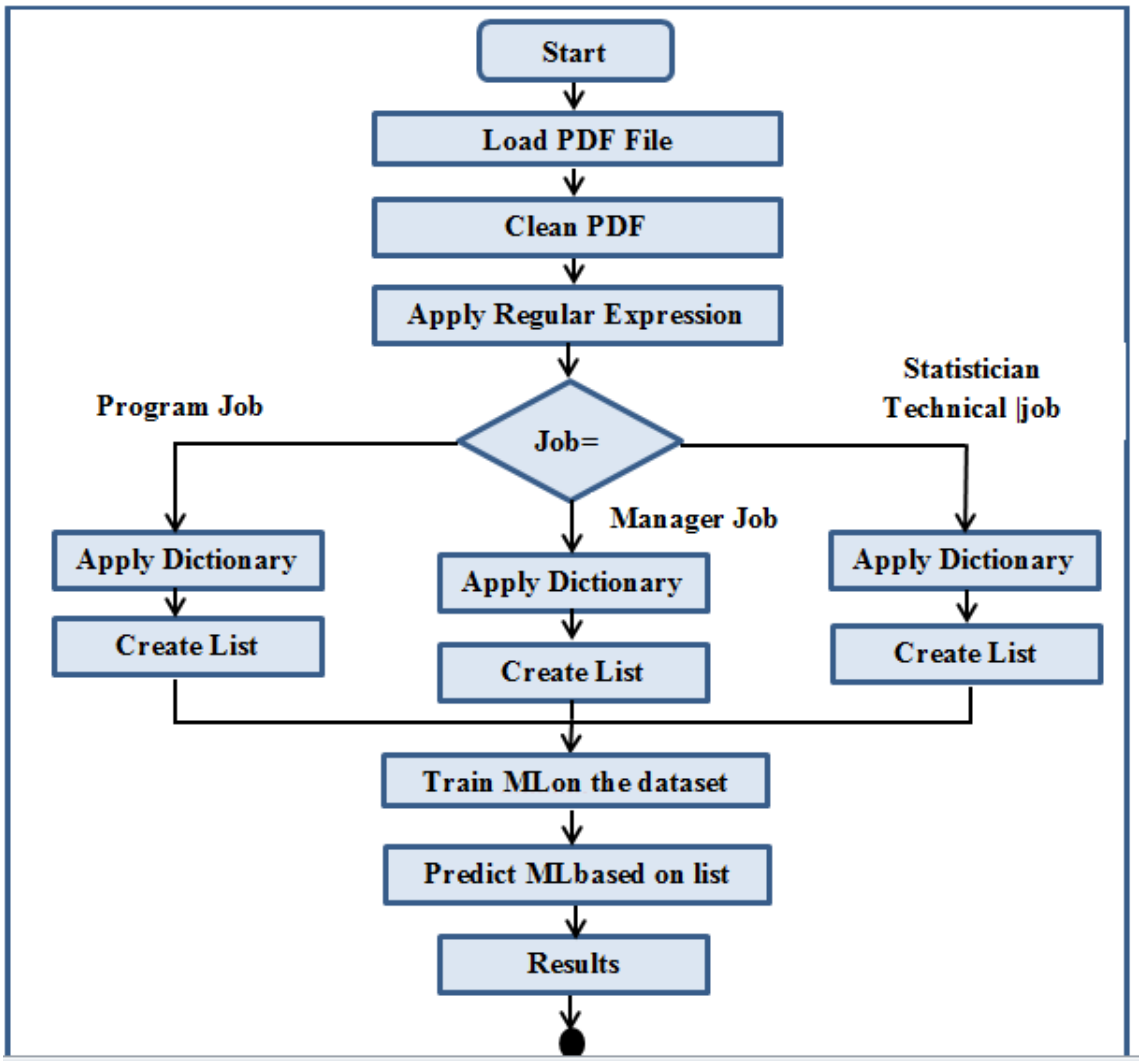


Figure 8: Flow chart for Matching PDF Resumes with job specification

CHAPTER 9: DATASETS

In this chapter, the writer reviews the data set used in the thesis in detail. Two sets of data were used: (1) The first was used to predict job requirements and selection criteria. (2) The second represents the set of candidates CVS.

8.1 First Dataset:

The dataset used to predict job requirements and selection criteria were obtained from the Department of Statistics in Jordan (JDoS); it is a Jordanian government department specialized in statistical work.

JDoS has 529 employees. The HR has a database that contains all info of employees that was extracted from an Oracle HR database, and aggregated in a single dataset. Table 2 illustrates the HR database attributes of JDoS:

Table 2: JDoS dataset Attribuites

	Attributes	Data Type	Description
1	Birthday	date	Birthday of employee
2	Gender	numeric	Gender of employee
3	Age	numeric	Age of employee
4	Marital Status	numeric	Marital Status of employee
5	No of Children	numeric	Number of children the employee has.
6	Address	text	Governorate in which the employee lives.
7	Qualification	text	last academic qualification obtained
8	Specialization	text	Specialization of employee
9	Average	numeric	Average of the last Academic qualification degree obtained
10	Graduation Year	numeric	Year of graduation
11	University	text	The university from which the employee graduated
12	University Type	text	Type of university
13	Directorates	text	The directorate in which the employee works
14	Job Title	text	job the employee works in
15	Salary	numeric	Salary of employee
16	Date of Hiring	date	Hiring date of employee
17	Experience	numeric	Years the employee spent in the job
18	Annual Performance evaluation	numeric	Percentage of assessment

Data processing is an important step for building a good machine learning model. After assembling target data, they must be prepared and cleaned. Cleaning removes observations including noise and missing data. Figures 9,10 present screen shots of JDoS dataset :

ser	GENDER	AGE	MARITAL	CHILDREN	ADDRESS	QUALIFICATION_	SPECIALIZATION_D	AVERAGE	GRADUATION
1	male	60	single	0	Amman	bachelor's	Agriculture	very good	1991
2	female	60	married	0	Irbid	bachelor's	computer science	good	1981
3	male	59	married	4	Amman	PHD	business management	excellent	2015
4	male	59	Widower	2	Amman	bachelor's	Economy	Acceptable	1991
5	male	59	married	4	Balqa	master's	Economy	very good	1990
6	male	59	married	4	Balqa	bachelor's	Economy	Acceptable	1990
7	female	58	married	4	Zarqa	bachelor's	mathematics	very good	2002

Figure 9: JDoS dataset screen shot 1

ser	UNIVERSITY	University Type	DIRECTORATE	JOB_TITLE	EXPERINCE	PERFORMANCE
1	Cyril and Methodius,Jugoslavija	outside Jordan	Agriculture dept	statistician technical	14	very good
2	Al al-Bayt	government	IT dept	programmee	12	Excellent
3	Azhari leader / Sudan	outside Jordan	Manager	Manager	33	Excellent
4	Yarmouk	government	Population dept	statistician technical	13	Excellent
5	Sindh/ Pakistan	outside Jordan	Manager	Manager	31	very good
6	Petra university	private	Economic dept	statistician technical	11	Excellent
7	The Niles/ Sudan	outside Jordan	Gender dept	statistician technical	11	Excellent

Figure 10 JDoS dataset screen shot 2

Data preparation phase: The row data have been processed as follows:

- Some employees were found without marital status values, the missing values were processed by updating data by HR.
- Attribute values of some employee's categories have null values such as university, specialization, and average, etc.., so the employees who have a high school qualification were excluded. The processed dataset contain 431 employees.

- Birthday and hiring date where specified features were considered unnecessary attributes, because age and experience attributes give information, but other features are relevant.
- Figure 11 Represents heat map which clarifies negative and positive correlations. It is usually correlate variables that affect credibility of results, which are usually removed. For example, hiring date because the researcuer can derive that years of experience. There is a high correlation between graduation year and the average which can be explained by market inflation in recent years compared to the 80s and 90s. The researcher found a high correlation between age and sex that this can be explained by the difference of retirement ages between males and females, according to the national insurance retirement age. However, we decided to keep both to avoid possible loss of information since males and females both have the option of early retirement 10 years earlier than the typical retirement age (60 for males and 50 for females).

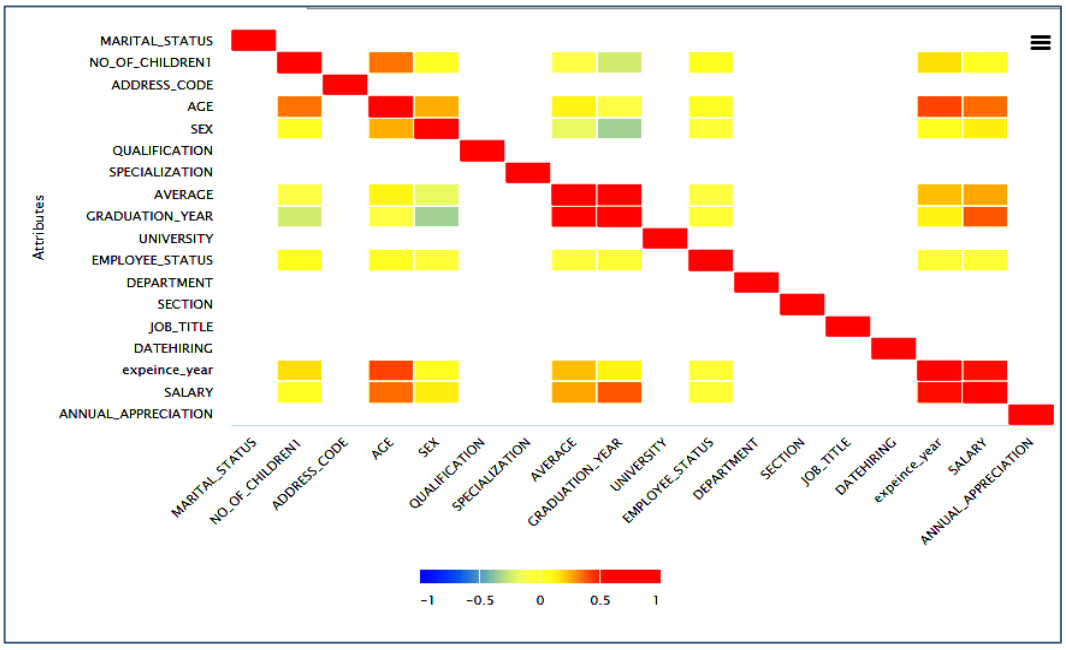


Figure 11: Heat Map to clarify correlation betweem features

- Resampling techniques were used to balance data, using random states, with replacing Sample, by adding samples from the lower category and removing samples from the higher category.

  In other words, resample method creats subsample randomly through replacement (Charte, et al. 2015).

- The data were explored by performing statistical analysis of each attribute, using SPSS tools:

1. Table 3 shows the percentage of descriptive statistical analysis of employees gender. It was noted that the number of females is slightly greater than that of males. The percentage of females was 50.8% but that of males was 49.2%. Data were balanced not much difference was found in numbers. There was no gender discrimination in jobs.

Table 3: descriptive statistical analysis of gender

Gender	Frequency	Percent
1. males	212	49.2
2. females	219	50.8

2. Table 4 shows the percentage of descriptive statistical analysis of ages of employees. It was noted that the highest percentage was that for ages from 35 to 44 which reted 49.2%, followed by age group 45 to 54 with 39.9%. The percentage was 5% for the age group from 55 to 61, and 9.5 for the group from 25 to 35. The researcher noticed that youngest age employees are those of 31 years of age because appointments are nominated by Civil Service Commission which adopts a competitive system. The oldest age was 60 years as age of retirement referred to.

Table 4: descriptive statistical analysis of age

Age	Frequency	Percent
1. from 25 to 34 years	23	5.3
2. From 35 to 44 years	225	52.2
3. From 45 to 54 years	161	37.4
4. From 55 to 60 years	22	5.1
5. >60 years	23	5.3

3. Table 5 shows the percentage of descriptive statistical analysis of marital Status data. It was noted that the highest percentage of employees was that of the married which was 86.3%. This is because most age groups of employees are 35 or above. So it is expected that the ages for this group are married, while the percentages of unmarried people were 12.8% and those of widows were 9.9%.

Table 5: descriptive statistical analysis of marital Status

marital Status	Frequency	Percent
single	55	12.8
married	372	86.3
Widower/ widow	4	.9

4. Table 6 shows the percentage of descriptive statistical analysis of number of employees children. It was noted that the highest percentage was 44.9% for those with 4 children, followed by 26.1% for 3 children, and 14.6 for two children. While the percentage of those who have one child was 9.6%, It was also noted that the percentage age of married employees with no children was 4.8%. Those with 5 children percentage of 1.9 %, while those who had 6 or 7 children the percentage was 1%.

Table 6: descriptive statistical analysis of No of Children

No of Children	Frequency	Percent
0	18	4.8
1	26	6.9
2	55	14.6
3	98	26.1
4	169	44.9
5	7	1.9
6	2	.5
7	1	.3

5. Table 7 shows the percentage of descriptive statistical analysis of employees addresses.It was noted that the largest percentage of employed people residing in the capital Amman was 86.5 %, followed by Irbid governorate 5.3 %, followed by Zarqa governorate 3.9 %. Balqa governorate was 1.2 %, but the percentage in Jarash, Karak, Madaba, Mafraq, Tafeleh, Was 1 %.

   It also noted that the highest proportion of employees live in Amman due to JDoS location.

Table 7: descriptive statistical analysis of Address

Address	Frequency	Percent
Ajloun	3	.7
Amman	373	86.5
Balqa	5	1.2
Irbid	23	5.3
Jarash	1	.2
KaraK	2	.5
Madaba	4	.9
Mafraq	2	.5
Tafeleh	1	.2
Zarqa	17	3.9

6. Table 8 shows the percentage of descriptive statistical analysis employees qualifications . It was noted the most of the employees in the dataset, hold a bachelor's degree with a percentage of 82.8%, 8.6% hold intermediate diploma, while and 6 have master's degree. The lowest percentages for the higher diploma degrees were 1.4 and PH.D 1.2.

Table 8 descriptive statistical analysis of qualification

Qualification	Frequency	Percent
1. Bachelor's	357	82.8
2. Higher diploma	5	1.2
3. Master's	27	6.3
4.PH/D	5	1.2
5. Intermediate diploma	37	8.6

7. Table 9 shows the percentages of descriptive statistical analysis average of university. It was noted that most of the employees in the dataset got a good

university average; that their percentage was 38.3 %, while 32.9 % of the employees got acceptable average, 23.4 % got very good rating, 2.6 % got excellent rating, but 2.8 % got weak rating.

Table 9: descriptive statistical analysis of university average

Values	Frequency	Percent
1. excellent	11	2.6
2. very good	101	23.4
3. good	165	38.3
4. acceptable	142	32.9
5. weak	12	2.8

8. Table 10 shows the percentages of descriptive statistical analysis of the University Type. It was noted that most employees who graduated from government universities, their percentage was 82.8 %. As for those who graduated from private universities, their percentage was 1.2 %, and 6.3 % was from that for students who graduated universities outside Jordan.

Table 10: descriptive statistical analysis of University Type

Values	Frequency	Percent
1. government	357	82.8
2. private	5	1.2
3. outside Jordan	27	6.3

9. Table 11 shows the percentages of descriptive statistical analysis of years of experience. It was noted that the percentages of employee's with an experience from 6 to 10 years was 39.4 %, followed by 29.2 % for those with experience of 11 to 14 years, was 11.8% from 15 to 19 years, 10.4 % while those with experience of 1 to 5 years, and 9 % for those with experience of 19 years or more.

Table 11: descriptive statistical analysis of Experience years

Values	Frequency	Percent
1. from 1 to 5 years	45	10.4
2. from 6 to 10 years	170	39.4
3. from 11 to 14 years	126	29.2
4. from 15 to 19 years	51	11.8
5. > 19 years	39	9.0

10. Table 12 shows the percentages of descriptive statistical analysis of the employee Annual Performance. It was noted that annual performance percentage was excellent with a percentages of 65.9 %, the percentages of employees with very good Performance was 33.4 %, and 0.7 % was for good performance.

Table 12: descriptive statistical analysis of Annual Performance

Values	Frequency	Percent
Excellent	284	65.9
Very good	144	33.4
Good	3	.7

By examining the data, it was found that the most functional levels in the JDoS data set were: programmer, manager, and technical statistician. So the dataset was divided into three categories, according to job title assigned for preparing specifications:

1. **Programmer dataset**

The data of those who obtained an excellent annual estimate was analyzed to select the values and characteristics that should be available in any new employee.

The gender attribute was analyzed and was noted that the percentage of female programmers who obtained excellent was 51.9% compared to 18.5 % for males. Gender will be overlooked to avoid racial discrimination.

In any job, a male or female can work, except jobs that require males due to the nature of work. Table 13 illustrates gender and performance rate analysis:

Table 13: Number of programmers by performance rate and gender

Gender	Annual Performance evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1 Male	5	18.5%	2	7.4%
2 Female	14	51.9%	6	22.2%

Age attribute in table 14 was analyzed. It was noted that the percentage of programmer's ages who obtained excellent rating was 35 to 44, age categories percentage was 55.6%,

compared to 11.1 % for those of 25 to 34 years old ,and the percentage of 45 to 54 years old categories was 3.7%, the lowest percentage of employees.

Table 14: Number of programmers by performance rate and age

Age	Annual Performance evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1 from 25 to 34 years	2	7.4%	0	0.0%
2 From 35 to 44 years	16	59.3%	4	14.8%
3 From 45 to 54 years	1	3.7%	4	14.8%

Table 15 shows that marital status analysis results the proportion of programmers with excellent rating were married with a percentage marital status 51.9%, while the percentage of single was 18.5%.

Table 15: Number of programmers by performance rate and marital status

Marital status	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1 single	5	18.5%	2	7.4%
2 married	14	51.9%	5	18.5%
3 Widower/ Widow	0	0.0%	1	3.7%

Table 16 shows the percentage of programmers who obtained excellent rating, and live in Amman where JDoS are located; their percentage is 55.6%, which is the highes, followed by 14.8 % for those who live in Irbid.

Table 16: Number of programmers by performance rate and address

Address	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
Amman	15	55.6%	6	22.2%
Irbid	4	14.8%	1	3.7%
Zarqa	0	0.0%	1	3.7%

Table 17 presents academic qualification of programmers who obtained excellent rating, the largest percentage was 55.6% for those holding a bachelor's degree, followed by 11.1% for those who had high diploma, and 3.7 % for those who had a master's degree.

Table 17: Number of programmers by performance rate and qualification

Address	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
Bachelor	15	55.6%	4	14.8%
Higher Diploma	0	0.0%	2	7.4%
Intermediate Diploma	3	11.1%	2	7.4%
Master	1	3.7%	0	0.0%

Table 18 analyzes the average of university data. It was noted that the highest percentage of programmers who got excellent performance rating was 40.7%, 25.9% with an acceptable rate, and 3.7% with good rate.

Table 18: Number of programmers by performance rate and Average

Average	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
excellent	0	0.0%	2	7.4%
very good	1	3.7%	2	7.4%
good	11	40.7%	4	14.8%
acceptable	7	25.9%	0	0.0%

It was also noted that 37.0% of programmers with excellent annual evaluation got academic qualification from a government universitys, while 33.3% of them graduated from private ones. The difference is not big as table 19 shows:

Table 19: Number of programmers by performance rate and University Type

University Type	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
government	10	37.0%	5	18.5%
private	9	33.3%	3	11.1%

It was also noted that the largest proportion of programmers with annual excellent rating was those with 11-14 years of experience with 28.0%, followed by those with 1-5 years 20.0%, but the percentage was equal for those with 6-10 years of experience those with

15 to 19 years of experience got 8.0%, and 4 % for those with more than 19 years of experience, as shown in the table 20.

Table 20: Number Of Programmers By Performance Rate And Experience

Experience	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1. from 1 to 5 years	5	20.0%	0	0.0%
2. from 6 to 10 years	2	8.0%	3	12.0%
3. from 11 to 14 years	7	28.0%	0	0.0%
4. from 15 to 19 years	2	8.0%	1	4.0%
5. > 19 years	1	4.0%	4	16.0%

2. Manager dataset

Data attributes of managers with excellent annual ratings were analysed. It was noted through the analysis of gender data that the number of males occupying a managerial position is greater than that of females, where the percentage of males was 66.7%, while that of females was 25.5%. Table 21 illustrates gender and performance rate analysis:

Table 21: Number of Manager by performance rate and gender

Gender	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1 Male	34	66.7%	3	5.9%
2 Female	13	25.5%	1	2.0%

Age attribute presenred in table 22 was analyzed. It was noted that the percentage of managers of 35 to 44 years old who obtained an excellent rating was 54.9 %, compared to 21.6 % for the category of 55 to 60 years ,and the category from 35 to 44 years was 15.7% which is the lowest percentage among employees.

Table 22: Number of Managers by performance rate and age

Age	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
2 From 35 to 44 years	8	15.7%	0	0.0%
3 From 45 to 54 years	28	54.9%	3	5.9%
4. From 55 to 60 years	11	21.6%	1	2.0%

Table 23 presenrs marital status analysis. It was noted that the largest proportion of managers with excellent rating were married with 84.3 %, 5.9 % of them were single, and 2 % were Widower/ Widow.

Table 23: Number of Managers by performance rate and marital status

Marital status	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1 single	3	5.9%	0	0.0%
2 married	43	84.3%	4	7.8%
3 Widower/ Widow	1	2.0%	0	0.0%

The table 24 shows the percentage of managers who obtained an excellent rating, and live in Amman where JDoS are located is 84.3%, which is the highest percentage. followed by 5.9 % for those who live in Irbid, and 2 % for those in Zarqa.

Table 24: Number of managers by performance rate and address

Adress	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
Amman	43	84.3%	4	7.8%
Irbid	3	5.9%	0	0.0%
Zarqa	1	2.0%	0	0.0%

As for the academic qualification of managers who obtained excellent rating, the highest percentage was 62.7% for those holding a bachelor's degree, followed by 19.6% with master degree, 7.8% with Ph.D degree, and 2.0 % for Intermediate diploma degree as presented in table 25:

Table 25: Number of managers by performance rate and qualification

Adress	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
Bachelor	32	62.7%	2	3.9%
Intermediate Diploma	1	2.0%	1	2.0%
Master	10	19.6%	1	2.0%
PhD	4	7.8%	0	0.0%

Table 26 a is analyzes average university data. It was noted that the highest percentage of managers of those who got excellent performance with good rating was of 45.1 %, 23.5% with acceptable rate, 9.8 % with very good rate, and 5.9% with excellent rate.

Table 26: Number of managers by performance rate and Average

Average	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
excellent	3	5.9%	0	0.0%
very good	5	9.8%	0	0.0%
good	23	45.1%	2	3.9%
acceptable	12	23.5%	1	2.0%
weak	4	7.8%	1	2.0%

Table 27 analyzes university type from which managers graduated from it. It was also noted that 51.0% of managers got excellent annual evaluation with academic qualification from government university, 39.2% of programmers graduated from universities Ouside Jordan , and 2.0% graduated from private universities.

Table 27: Number of managers by performance rate and University Type

University Type	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
government	26	51.0%	3	5.9%
private	1	2.0%	0	0.0%
Out of Jordan	20	39.2%	1	2.0%

It was also noted that the largest proportion of managers with annual excellent rating was of those 11 to 14 years of experience and those with 15 to 19 years of experience 25.5%, followed by those of more than 19 years of experience 37.3 % as shown in table 28.

Table 28: Number of Manager by performance rate and Experience

Experience	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
2. from 6 to 10 years	2	3.9%	0	0.0%
3. from 11 to 14 years	13	25.5%	1	2.0%
4. from 15 to 19 years	13	25.5%	1	2.0%
5. > 19 years	19	37.3%	2	3.9%

3. Technical statistician dataset

Data attributes of technical statistician with excellent annual ratings were analysed. It was noted through analysis of gender data that the number of males occupying a technical statistical position is close to that of females, as the percentage of males was 30.7%, while that of females was 29.3%. Table 29 illustrates gender and performance rate analysis:

Table 29: Number of technical statistician by performance rate and gender

Gender	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1 Male	83	30.7%	43	15.9%
2 Female	79	29.3%	65	24.1%

Age attribute was analyzed, It was noted that the percentage of technical statistician of 35 to 44 years of age is 31.1%, compared to 24.1% for those of 45 to 54 years of age ,and percentage of 55 to 60 years old is 3 % which is the lowest, and 1.9 % for those 25 to 34 years as shown in table 30.

Table 30: Number of technical statistician by performance rate and age

Age	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1. from 25 to 34 years	5	1.9%	3	1.1%
2 From 35 to 44 years	84	31.1%	64	23.7%
3 From 45 to 54 years	65	24.1%	36	13.3%
4. From 55 to 60 years	8	3.0%	5	1.9%

Table 31 shows marital status analysis. It was noted that the largest proportion of managers with excellent rating was of the married with 51.5%, 8.1% of them were single, and 0.4 % were Widower/Widow .

Table 31: Number of technical statistician by performance rate and marital status

Marital status	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1 single	22	8.1%	15	5.6%
2 married	139	51.5%	92	34.1%
3 Widower/Widow	1	0.4%	1	0.4%

Table 32 shows the percentage of technical statistician who obtained excellent rating, and live in Amman where JDoS is located. Their percentage is 51.5%, which is the highest, followed by 3.7 % who live in Irbid..

Table 32: Number of technical statistician by performance rate and address

Address	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
amman	139	51.5%	92	34.1%
Irbid	10	3.7%	4	1.5%
zarqa	5	1.9%	10	3.7%
balqa	3	1.1%	1	0.4%
Jarash	1	0.4%	0	0.0%
karak	2	0.7%	0	0.0%
madaba	1	0.4%	1	0.4%
tafeleh	1	0.4%	0	0.0%

As for the academic qualification of technical statistician who obtained excellent rating, the highest percentage was 52.2 % for those holding a bachelor's degree, followed by 5.2 % for those of Intermediate Diploma, and 2.2% for those with master's degree as shown in table 33.

Table 33: Number of technical statistician by performance rate and qualification

Qualification	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
Bachelor	141	52.2%	100	37.0%
Intermediate Diploma	1	0.4%	0	0.0%
Intermediate Diploma	14	5.2%	4	1.5%
Master	6	2.2%	3	1.1%
Ph.D	0	0.0%	1	0.4%

Table 34 a is analyzes average university data. It was noted that the highest percentage of technical statistician who got excellent performance rating got an acceptable rate of 23.3%, 18.9% with good rate, and 12.6 % a very good rate.

Table 34: Number of technical statistician by performance rate and Average

Average	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
excellent	4	1.5%	0	0.0%
very good	34	12.6%	21	7.8%
good	51	18.9%	31	11.5%
Acceptable	63	23.3%	55	20.4%
weak	10	3.7%	1	0.4%

It was also noted that 39.3% of technical statistician with excellent annual evaluation have an academic qualification from a government university, 13.3% of programmers graduated from private universities, and 7.4% from Out of Jordan universities as show in table 35.

Table 35: Number of technical statistician by performance rate and University Type

University Type	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
government	106	39.3%	81	30.0%
private	36	13.3%	20	7.4%
Out of Jordan	20	7.4%	7	2.6%

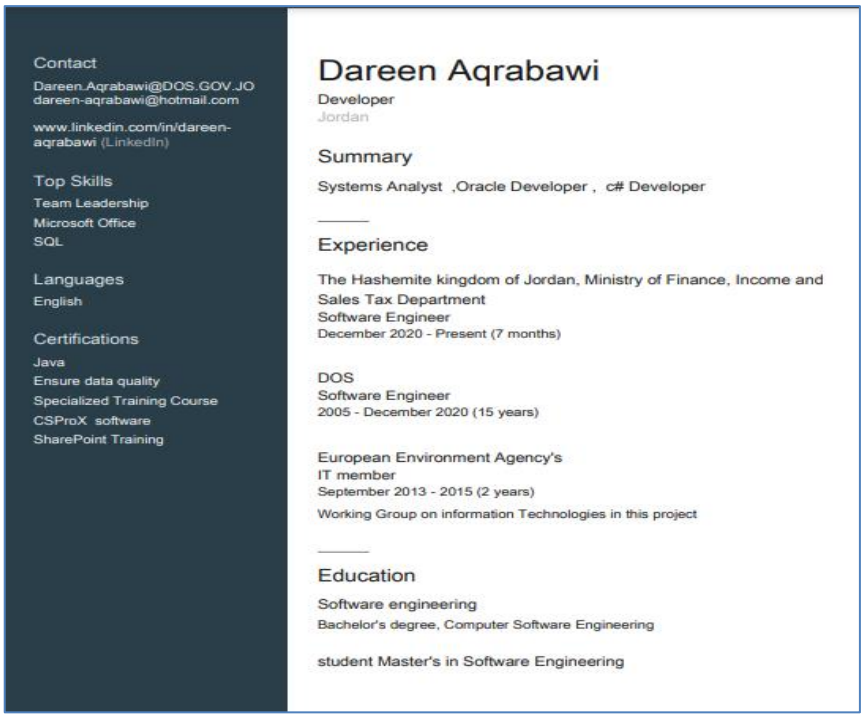
It was also noted that the largest proportion of technical statistician with an annual excellent rating was that of 11-14 years of experience with 37.8 %, followed by those with 6 to 10 years of experience with 12.2 %, and 5.9 %, 15 to 19 years, and 3.0 % for those with more than 19 years of experience, as shown in the table 36.

Table 36: Number of technical statistician by performance rate and Experience

Experience	Annual Performance Evaluation			
	1 Excellent		2 very good	
	Count	Total %	Count	Total %
1. from 25 to 34 years	3	1.1%	7	2.6%
2. from 6 to 10 years	33	12.2%	30	11.1%
3. from 11 to 14 years	102	37.8%	61	22.6%
4. from 15 to 19 years	16	5.9%	7	2.6%
5. > 19 years	8	3.0%	3	1.1%

8.2 Second Dataset:

The approach used a set of CVs that was downloaded in different ways, from multi-sources such as job sites which contain many of thosw searching for jobs like LinkedIn and websites. Figures 12, 13 are illustrated CVS of candidates.



Contact

Dareen.Aqrabawi@DOS.GOV.JO
dareen-aqrabawi@hotmail.com

www.linkedin.com/in/dareen-aqrabawi (LinkedIn)

Top Skills

Team Leadership
Microsoft Office
SQL

Languages

English

Certifications

Java
Ensure data quality
Specialized Training Course
CSProX software
SharePoint Training

Dareen Aqrabawi

Developer
Jordan

Summary

Systems Analyst ,Oracle Developer , c# Developer

Experience

The Hashemite kingdom of Jordan, Ministry of Finance, Income and Sales Tax Department
Software Engineer
December 2020 - Present (7 months)

DOS
Software Engineer
2005 - December 2020 (15 years)

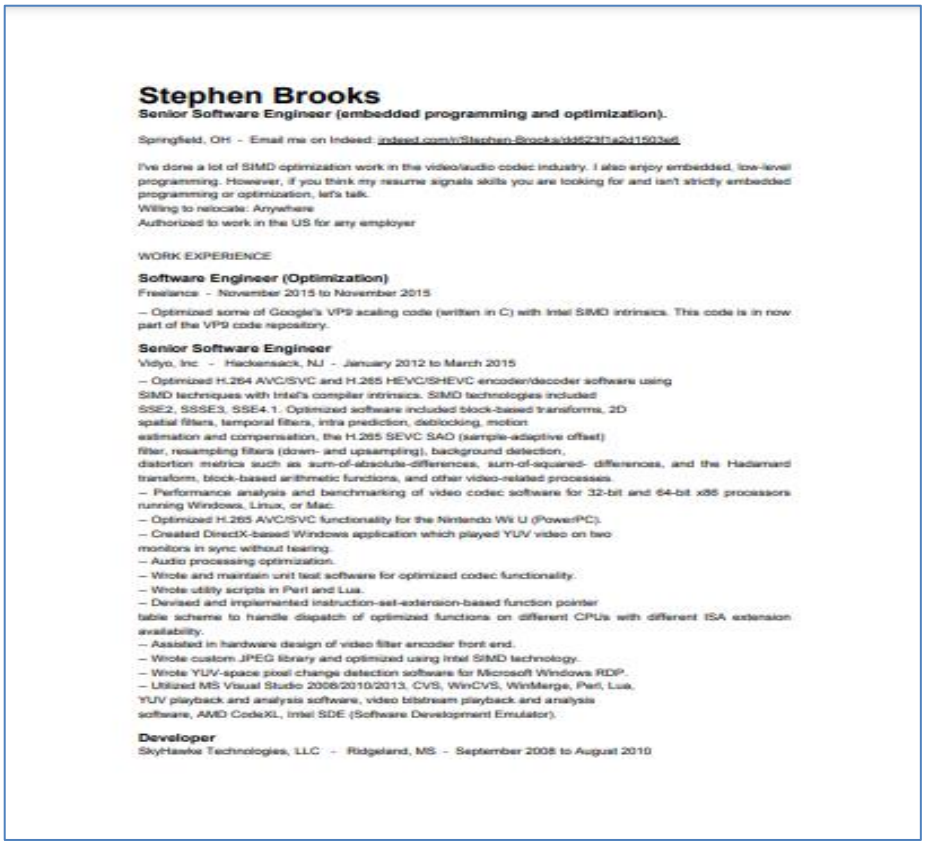
European Environment Agency's
IT member
September 2013 - 2015 (2 years)
Working Group on information Technologies in this project

Education

Software engineering
Bachelor's degree, Computer Software Engineering

student Master's in Software Engineering

Figure 12: Example 1 of cv



Stephen Brooks

Senior Software Engineer (embedded programming and optimization).

Springfield, OH - Email me on Indeed: indeed.com/r/Stephen-Brooks/dd623f1a2d1593e6

I've done a lot of SIMD optimization work in the video/audio codec industry. I also enjoy embedded, low-level programming. However, if you think my resume signals skills you are looking for and isn't strictly embedded programming or optimization, let's talk.
Willing to relocate: Anywhere
Authorized to work in the US for any employer

WORK EXPERIENCE

Software Engineer (Optimization)
Freelance - November 2015 to November 2015

-- Optimized some of Google's VP9 scaling code (written in C) with Intel SIMD intrinsics. This code is in now part of the VP9 code repository.

Senior Software Engineer
Vidyo, Inc - Hackensack, NJ - January 2012 to March 2015

-- Optimized H.264 AVC/SVC and H.265 HEVC/SHEVC encoder/decoder software using SIMD techniques with Intel's compiler intrinsics. SIMD technologies included SSE2, SSSE3, SSE4.1. Optimized software included block-based transforms, 2D spatial filters, temporal filters, intra prediction, deblocking, motion estimation and compensation, the H.265 SEVC SAO (sample-adaptive offset) filter, resampling filters (down- and upsampling), background detection, distortion metrics such as sum-of-absolute-differences, sum-of-squared- differences, and the Hadamard transform, block-based arithmetic functions, and other video-related processes.
-- Performance analysis and benchmarking of video codec software for 32-bit and 64-bit x86 processors running Windows, Linux, or Mac.
-- Optimized H.265 AVC/SVC functionality for the Nintendo Wii U (PowerPC).
-- Created DirectX-based Windows application which played YUV video on two monitors in sync without tearing.
-- Audio processing optimization.
-- Wrote and maintain unit test software for optimized codec functionality.
-- Wrote utility scripts in Perl and Lua.
-- Devised and implemented instruction-set-extension-based function pointer table scheme to handle dispatch of optimized functions on different CPUs with different ISA extension availability.
-- Assisted in hardware design of video filter encoder front end.
-- Wrote custom JPEG library and optimized using Intel SIMD technology.
-- Wrote YUV-space pixel change detection software for Microsoft Windows RDP.
-- Utilized MS Visual Studio 2008/2010/2013, CVS, WinCVS, WinMerge, Perl, Lua, YUV playback and analysis software, video bitstream playback and analysis software, AMD CodeXL, Intel SDE (Software Development Emulator).

Developer
SkyHawke Technologies, LLC - Ridgeland, MS - September 2008 to August 2010

Figure 13: Example 2 of cv

CHAPTER 10: RESULTS

Three machine learning models were built to classify each category of data, and to extract the most important factors that affect performance of the employee in this category. Three dataseta were used: programmer dataset, manager dataset, and technical statistician dataset, which help in the experiment of writing 3 job specifications for three job titles.

Many classifiers were used. The experiment and the best classifier were determined to building a machine learning model based on such data for all job specification categories chosen for the programmer, manager, and statistician experiment. The result was that the model, which used k-Nearest Neighbors, had the highest performance and accuracy compared to other models of the programmer's dataset, As for manager's data set, results of the model that used k-Nearest Neighbors were equal to the results of the model that used the random forest, but the technical statistician dataset was the highest results of the models that used k-Nearest Neighbors, as shown in Table 37:

Table 37: Results of algorithms classification

algorithm	Job	Average	Accuracy	PECISSION	RECALL	ROC
k-Nearest Neighbours	programmer	91%	93%	.93	.93	.91
	manager		94%	.93	.94	.86
	statistician		86%	.86	.86	.86
Random Forest	programmer	89%	88.8%	.893	.889	.98
	manager		94%	.945	.94	.78
	statistician		85%	.85	.85	.90
Random Committee	programmer	86%	81%	.82	.82	.91
	manager		94%	.93	.94	.94
	statistician		48%	.84	.84	.91
Random Tree	programmer	85%	85%	.86	.85	.86
	manager		88%	.89	.88	.81
	statistician		81%	.81	.81	.78
Neural Network	programmer	85%	85%	.86	.85	.96
	manager		88%	.88	.88	.75
	statistician		81%	.81	.82	.79
Decision Tree J48	programmer	85%	89%	.89	.889	.89
	manager		88%	.89	.88	.82
	statistician		79%	.79	.79	.77
Logistic Regression	programmer	80%	85%	.86	.85	.98
	manager		90%	.92	.90	.89
	statistician		66%	.64	.66	.66
SVM	programmer	79%	85%	.86	.85	.86
	manager		88%	.89	.88	.57
	statistician		63%	.57	.63	.51
Naïve Bayes	programmer	67%	70%	.70	.70	.79
	manager		82%	.83	.82	.80
	statistician		50%	.62	.50	.59

Based on these classifiers (k-Nearest Neighbors, Random Forest), the researcher experimented with many ranker algorithms to select the most important features in order to adopt them as requirements for writing job specification document for each experience. The following, Table 38 summarizes the results of experiments to select the most important features, using many algorithms to be adopted in writing job specification requirements for the programmer.

Concerning the job specification for manager position, the following Table, 39 summarizes the most important features in the experiments that used many algorithms to describe the requirements for a manager job.

Finally, the job specification for a technical statistician position, the following table, 40, describes experiment that used several algorithms and summarizes the most important features to describe requirements of this job.

Table 38: The most important features by ranker algorithms **using programmer's dataset**

	Symmetric aluncer Attribute Eval Ranker	information gain ranking	gain ratii ranker	oner attribute Eval	svm attribute Eval	cfs subset Eval\best first	relief attribute Eval	psosearch/ cfsSubsetE	Greedy Stepwise/ cfsSubsetE
1	University type	University type	qualification	University type	University type	qualification	qualification	qualification	qualification
2	specification	specification	experience	experience	Department	University type	experience	specification	university
3	experience	job title	experience	job title	job title		marital status		
4	qualification	experience	University type	experience	qualification		job title		
5	average	average	specification	age	average		Department		
6	marital status	age	marital status	specification	Department		age		
7	age	job title	Department	qualification	specification		Department		

Table 39: The most important features by ranker algorithms using the manager JDoS archived dataset

	Symmetric aluncer Attribute Eval Ranker	information gain ranking	gain ratii ranker	oner attribute Eval	svm attribute Eval	cfs subset Eval\best first	relief attribute Eval	psosearch/ cfsSubsetE	Greedy Stepwise/ cfsSubsetE
1	specification	qualification	qualification	qualification	qualification	specification	specification	qualification	marital status
2	Job type	Job type	Job type	gender	specification	Job type	qualification	Job type	Job type
3	experience	experience	university	average	specification		University type	experience	
4				specification	University type		experience		
5				age	age		age		
6				experience	average		average		
7					experience		Address		

Table 40: The most important features by ranker algorithms using technical statistician JDoS archived dataset

	Symmetric aluncer Attribute Eval Ranker	information gain ranking	gain ratii ranker	oner attribute Eval	svm attribute Eval	cfs subset Eval\best first	relief attribute Eval	psosearch/ cfsSubsetE	Greedy Stepwise/ cfsSubsetE
1	**Address code**	**Address code**	**Address code**	**Job type**	**Job type**	**Address**	**specification**	**qualification**	**marital status**
2	**Job type**	**Job type**	**qualification**	**gender**	**graduation year**	**Job type**	**qualification**	**qualification**	**Job type**
3			**age**	**average**	**specification**	**age**	**gender**		
4				**specification**	**University type**		**experience**		
5				**age**	**age**		**age**		
6				**experience**	**average**		**average**		
7					**experience**		**Address code**		

By comparing the previous tables for each job, the researcher determined eight important features by ranker algorithms and the attibutes most frequently noted in the results, as shown in the Table 41:

Table 41: The most important features by ranker algorithms fo each job specification

	Programmer's	Managers's	Technical statistician's
1	Qualification	Qualification	Address code
2	University type	Specification	Job title
3	Experience	Job type	Qualification
4	Specification	Experience	Specification
5	Job title	Age	Age
6	Average	Average	University type
7	Age	University type	Average
8	Department	Address code	Experience

In the functional requirements for the position of a programmer and manager, the resulst revealed that the academic qualification is the most important among others.

For the programmer position the type of university is more important than the years of experience which is more important than Aaverage and age.

But the functional requirements for managers position, Specification and Experience are more important than Average and University type.

The results also show that address code of technical statisticians is the most important attributes among others; job title and qualification are more imbortant than Age and University type. The lowest important criteria for technical statisticians are average and experience.

First experiment programmer's job requirments:

The researcher extracted the characteristics that should be available in any new programmer; she she found these requirement and Criteria:

1. Depending on all of the preceding analysis, the job specificationto to be advertised for the programmer position is:Qualification: Bachelor degree or higher.
2. University: Government University
3. Experience: 1 - 14 years
4. Specialization: Information Technology majors
5. Job title: expert programmer
6. Average: a good university estimate or higher
7. Age: 25 - 44 years old
8. Department: The department is of the greatest important, as it must be within the employee's specialization and scientific qualification.

Second experiment manager's job requirments:

The researcher extracted the characteristics that should be available in any new IT manager, she found these requirement and Criteria, So the job Specification to be announced for IT manager position is:

1. Qualification: Bachelor degree or higher.
2. Specialization: Information Technology majors
3. Job title: expert programmer manager.
4. Experience: 11 - 19 years
5. Age: 45 - 54 years old
6. Average: a good university estimate or higher
7. University: Government or Out of Jordan
8. Address code: Amman

Third experiment technical statistician's job requirments:

The researcher extracted the characteristics that should be available in any new programmer; she she found these requirement and Criteria.

Based on the foregoing, the most important criteria for technical statistician position to be advertised for that the job include the following:

1. Address code: Amman
2. Job title: technical Statistician
3. Qualification: Bachelor degree or higher
4. Specialization: Specialties according to the relevant department, if the vacancy is in the agricultural department, for example, the specialization must be agricultural.
5. Age: 35 - 54 years old
6. University: Government
7. Average: acceptable university estimate or higher
8. Experience: 6 - 14 years

After three job specification documents have been prepared, the job is usually advertised through public or company's websites; a date is set for receiving resumes. To complete the experiment, several job-related resumes and others unrelated to the job were selected, for testing the model .10 PDF resumes for each job were tested. The model was built for matching the best resumes with job specification by Python code, with 80% accuracy, 0.89 Recall, and 0.87 precision. |The the proportion of resumes matching each job specifications is shown in Table 42:

Table 42: Results of accuracy of matching the pdf résumes submitted with job specification's

Programmer's		Manager's		Technical Statistician's	
CVs	**Matching %**	**CVs**	**Matching %**	**CVS**	**Matching %**
Resume1	90	Resume1	80	Resume1	88
Resume2	65	Resume2	78	Resume2	70
Resume3	36	Resume3	70	Resume3	54
Resume 4	35	Resume 4	55	Resume 4	40
Resume5	35	Resume5	40	Resume5	33
Resume6	35	Resume6	35	Resume6	25
Resume7	27	Resume7	30	Resume7	25
Resume8	26	Resume8	26	Resume8	17
Resume9	24	Resume9	24	Resume9	15
Resume10	23	Resume10	23	Resume10	15

CHAPTER 11: DISCUSSION

This chapter discusses whether the results obtained were verified proved the hypothesis answered the research question and provided some additional scientific contribution to previous works related to the topic of the research, followed by a comparison with results of previous works.

This research presents an intelligent automated system that creates a job specification and then automatically filters the CVs matching the required job specifications, according to the matching percentage and best accuracy.

The results of creating job specifications and determining their requirements showed a high performance, with an accuracy of 91%, when using the k-Nearest Neighbors algorithm. And the model also showed the most important features for each job titles.

The results of creating job description were limited, due to the use of insufficient archived data, and to the lack of some information such as skills and experience certificates. The hope is that future studies will provide more comprehensive data.

As for the most important features that should be available to the applicant for the job, they were limited because the data used were for a government department which is governed by some restrictions in hiring. These are competitive arrangement that depends on the role of civil service not the applicant's skills, but in this research, work was hard to obtain the most important criteria from the available data.

The performance of this approach for matching CVS with job specification in this methodology using résumés with PDF Format and government dataset for designing job descriptions shows that better accuracy is (80%), when using the Random Forest and KNN, compared to the work of (Roy, Chowdhary and Bhatia 2020), who got an accuracy of 38.99 % when using Random Forest and an accuracy of 78.53% with Linear SVM by using résumés in Excel Format.

As we know the majority of CVs are submitted in PDF format so as not to be modified and to protect the candidate's data. As for the Excel data, it is only used for the submission about a customized form through a website or the like. This gives an additional advantage to this research because it has the ability to analyze and extract data from pdf.

Additionally, this work enhances the work of (El Mohadab, Bouikhalene and Safi 2020) that used dataset of scientific research from the Postgresql database that contains all records of doctoral students at the university and NLP with a decision tree classifier with 94.3% Precision and 97.1% Recall while this approach used PDF of different sources and designs, and used NLP with a Random Forest of 89% Recall, of 87% precision.

This research also differs from (Buil, Catalán and Martínez 2020) work, in that it uses candidate data that was entered in a standard way through a website by a questionnaire. It also differs from the work of (Yahiaoui, et al. 2019) who used SQL DB Server to select the right skilled people from User profile database by the Gephi graphical analysis tool, a competence repository,and a web scraping tool.

The study of (Ajoudanian & Abadeh, 2019) was also a different work as they used novel fuzzy C-means clustering with precision of 85.3%.

CHAPTER 12: Conclusions and Recommendations

This work aimed at enhancing the way of creating job specifications, and to that the requirements are not inappropriately randomized, and to reduce false job advertisements that do not satisfy needs of the same job.To enhance attracting the right employee for the right job, the methodology applied was by DM and ML methods to extract a good job specification, using Jordanian department of statistics data set.

This approach has been created to handle and predict the most important criteria and features that must exist in job specification. The issue of imbalance was treated using the re-sampling method to increase the performance of prediction model. In the second step, the approach used the feature selection methods of data sets, such as ranker search method with information to gain attribute and other methods; it was previously mentioned and summarized in previous tables 12, 13, 14.

The methodology of this research adds an innovative way to write job specification, evaluation results were excellent, using the KNN and Random Forest models. The approach also shows results improvement in matching phase of resumes in PDF format and gives better results than previous works; accuracy was 80% for the random jungle model and the matching percentages were CVs are different and the highest matching rate was 90%. The following is a summary of contributions of this research:

- The researcher used reliable government department data for actual employees, from which she predicted the requirements for some new jobs.
- Enhance the way of determining Job Requirements & Selection criteria.
- Predict of the most important criteria and features that must exist in Job Specifications.

- Add an innovative way to write the job Specification to ensure that the job advertisement is not random and biased to the circumstances of a particular person.
- Enhances the results of the matching phase of resumes in PDF format than previous work.
- The research results were promising in terms of accuracy compared to previous works, and also in the way of determining the appropriate CV.

There were limitations and difficulty in obtaining an archived employee database,. Therefore, we recommend more experiments at the stage of preparing the job specification and using a larger database. We also recommend using another classification model to match resumes with the job specifications.

References

Acikgoz, Yalcin. “Employee recruitment and job search: Towards a multi-level integration.” *Human resource management review* (Elsevier) 29 (2019): 1–13.

Ajoudanian, Shohreh, and Maryam Nooraei Abadeh. “Recommending human resources to project leaders using a collaborative filtering-based recommender system: Case study of gitHub.” *IET Software* (IET) 13 (2019): 379–385.

BaniMustafa, Ahmed. “Enhancing learning from imbalanced classes via data preprocessing: A data-driven application in metabolomics data mining.” *The ISC International Journal of Information Security* (Iranian Society of Cryptology) 11 (2019): 79–89.

BaniMustafa, Ahmed Hmaidan, and Nigel W. Hardy. “A strategy for selecting data mining techniques in metabolomics.” In *Plant Metabolomics*, 317–333. Springer, 2011.

Banimustafa, Ahmed, and Nigel Hardy. “A Scientific Knowledge Discovery and Data Mining Process Model for Metabolomics.” *IEEE Access* (IEEE) 8 (2020): 209964–210005.

Bayes, Thomas. “LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S.” *Philosophical transactions of the Royal Society of London* (The Royal Society London), 1763: 370–418.

Belgiu, Mariana, and Lucian Drăguţ. “Random forest in remote sensing: A review of applications and future directions.” *ISPRS journal of photogrammetry and remote sensing* (Elsevier) 114 (2016): 24–31.

Bellhouse, David R. “The Reverend Thomas Bayes, FRS: a biography to celebrate the tercentenary of his birth.” *Statistical Science* (Institute of Mathematical Statistics) 19 (2004): 3–43.

Bondielli, Alessandro, and Francesco Marcelloni. “A data-driven approach to automatic extraction of professional figure profiles from Résumés.” *International Conference on Intelligent Data Engineering and Automated Learning.* 2019. 155–165.

Brandão, Catarina, Rita Silva, and Joana Vieira dos Santos. “Online recruitment in Portugal: Theories and candidate profiles.” *Journal of Business Research* (Elsevier) 94 (2019): 273–279.

Brat, G, D Drusinsky, and D Giannakopoulou. “Experimental Evaluation of Verification and Validation Tools on Martian Rover Software.” *Kluwer Academic Publishers* (Formal Methods in System Design) 25 (2004).

Brownlee, Jason. “Machine learning mastery with Weka.” *Ebook. Edition: v. 1.4*, 2019.

Buil, Isabel, Sara Catalán, and Eva Martínez. “Understanding applicants’ reactions to gamified recruitment.” *Journal of Business Research* (Elsevier) 110 (2020): 41–50.

Cervantes, Jair, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. “A comprehensive survey on support vector machine classification: Applications, challenges and trends.” *Neurocomputing* (Elsevier) 408 (2020): 189–215.

Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. “Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews].” *IEEE Transactions on Neural Networks* (IEEE) 20 (2009): 542–542.

Chapman, Pete, et al. “CRISP-DM 1.0: Step-by-step data mining guide.” *SPSS inc* 9 (2000): 13.

Charte, Francisco, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. “Addressing imbalance in multilabel classification: Measures and random resampling algorithms.” *Neurocomputing* (Elsevier) 163 (2015): 3–16.

Chen, Jie, Chunxia Zhang, and Zhendong Niu. “A two-step resume information extraction algorithm.” *Mathematical Problems in Engineering* (Hindawi) 2018 (2018).

Chien, Chen-Fu, and Li-Fei Chen. “Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry.” *Expert Systems with applications* (Elsevier) 34 (2008): 280–290.

Chopra, Abhimanyu, Abhinav Prashar, and Chandresh Sain. “Natural language processing.” *International journal of technology enhancements and emerging engineering research* (Citeseer) 1 (2013): 131–134.

Chung, H. Michael, and Paul Gray. “Data mining.” *Journal of management information systems* (Taylor & Francis) 16 (1999): 11–16.

Courtin, Christophe, and Miguel Tomasena. “A benchmarking platform for analyzing corpora of traces: the recognition of the users' involvement in fields of competencies.” *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies.* 2016. 178–186.

Coverdill, James E., and William Finlay. *High tech and high touch: Headhunting, technology, and economic transformation.* Cornell University Press, 2017.

“CRISP-DM: Towards a standard process model for data mining.” 1 (n.d.).

Deng, Zhenyun, Xiaoshu Zhu, Debo Cheng, Ming Zong, and Shichao Zhang. “Efficient kNN classification algorithm for big data.” *Neurocomputing* (Elsevier) 195 (2016): 143–148.

“Does machine learning really work?” *AI magazine* 18 (1997): 11–11.

El Mohadab, Mohamed, Belaid Bouikhalene, and Said Safi. “Automatic CV processing for scientific research using data mining algorithm.” *Journal of King Saud University-Computer and Information Sciences* (Elsevier) 32 (2020): 561–567.

Farhat, Nabil H., and Vladimir Vapnik. “Support-Vector Networks.” *IEEE Expert Intell. Syst. Their Appl* 7 (1992): 63–72.

Fix, Evelyn, and J. L. Hodges. “Discriminatory analysis, nonparametric discrimination.” (consistency properties. Technical Report 4, United States Air Force, School …) 1951.

Flach, Peter A. “The geometry of ROC space: understanding machine learning metrics through ROC isometrics.” *Proceedings of the 20th international conference on machine learning (ICML-03).* 2003. 194-201.

Freitas, Alex A. “A survey of evolutionary algorithms for data mining and knowledge discovery.” In *Advances in evolutionary computing*, 819--845. Springer, 2003.

Galdi, Paola, and Roberto Tagliaferri. “Data mining: accuracy and error measures for classification and prediction.” *Encyclopedia of Bioinformatics and Computational Biology* (Elsevier Amsterdam, The Netherlands), 2018: 431–6.

García, Salvador, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining.* Edited by Springer. Springer, 2015.

Gentleman, Robert, and Vincent J. Carey. “Unsupervised machine learning.” In *Bioconductor case studies*, 137–157. Springer, 2008.

Georgiou, Konstantina, Athanasios Gouras, and Ioannis Nikolaou. “Gamification in employee selection: The development of a gamified assessment.” *International journal of selection and assessment* (Wiley Online Library) 27 (2019): 91–103.

Gholami, Jafar, Farhad Pourpanah, and Xizhao Wang. “Feature selection based on improved binary global harmony search for data classification.” *Applied Soft Computing* (Elsevier) 93 (2020): 106402.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. “The WEKA data mining software: an update.” *ACM SIGKDD explorations newsletter* (ACM New York, NY, USA) 11 (2009): 10–18.

Ho, Tin Kam. “Random decision forests.” *Proceedings of 3rd international conference on document analysis and recognition.* 1995. 278–282.

Jiang, Tammy, Jaimie L. Gradus, and Anthony J. Rosellini. “Supervised machine learning: a brief primer.” *Behavior Therapy* (Elsevier) 51 (2020): 675–687.

Jordan, Michael I, and Tom M Mitchell. “Machine learning: Trends, perspectives, and prospects.” *Science,* 2015: 255--260.

Kaelbling, Leslie Pack, Michael L. Littman, and Andrew W. Moore. “Reinforcement learning: A survey.” *Journal of artificial intelligence research* 4 (1996): 237–285.

Kesavaraj, Gopalan, and Sreekumar Sukumaran. “A study on classification techniques in data mining.” *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT).* 2013. 1-7.

Khoshgoftaar, Taghi M., Kehan Gao, and Naeem Seliya. “Attribute selection and imbalanced data: Problems in software defect prediction.” *2010 22nd IEEE International conference on tools with artificial intelligence.* 2010. 137-144.

Kohavi, and Provost. “The Case Against Accuracy Estimation for Comparing Introduction Algorithm.” *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning.* 1998. 445--453.

LaBerge, David. “A recruitment theory of simple behavior.” *Psychometrika* (Springer) 27 (1962): 375–396.

Larose, Daniel T., and Chantal D. Larose. *Discovering knowledge in data: an introduction to data mining.* Vol. 4. John Wiley & Sons, 2014.

Laurim, Vanessa and Arpaci, Selin and Prommegger, Barbara and Krcmar, and Helmut. “Computer, Whom Should I Hire?--Acceptance Criteria for Artificial Intelligence in the Recruitment Process.” In *Proceedings of the 54th Hawaii International Conference on System Sciences,* 5495. 2021.

Lee, Choong Kwon, and Hyo-Joo Han. “Analysis of skills requirement for entry-level programmer/analysts in Fortune 500 corporations.” *Journal of Information Systems Education* 19 (2008): 17.

Liu, Huan, and Lei Yu. “Toward integrating feature selection algorithms for classification and clustering.” *IEEE Transactions on knowledge and data engineering* (IEEE) 17 (2005): 491-502.

Luque, Amalia, Alejandro Carrasco, Alejandro Martín, and Ana Heras. “The impact of class imbalance in classification performance metrics based on the binary confusion matrix.” *Pattern Recognition* (Elsevier) 91 (2019): 216-231.

Meshram, Sarita Gajbhiye, Vijay P. Singh, Ozgur Kisi, Vahid Karimi, and Chandrashekhar Meshram. “Application of artificial neural networks, support

vector machine and multiple model-ANN to sediment yield prediction." *Water Resources Management* (Springer) 34 (2020): 4561–4575.

Mo, Yunjeong, Dong Zhao, Jing Du, Matt Syal, Azizan Aziz, and Heng Li. "Automated staff assignment for building maintenance using natural language processing." *Automation in Construction* (Elsevier) 113 (2020): 103150.

Nagarhalli, Tatwadarshi P., Vinod Vaze, and N. K. Rana. "Impact of Machine Learning in Natural Language Processing: A Review." *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV).* 2021. 1529–1534.

Necula, Sabina-Cristiana, and Cătălin Strîmbei. "People analytics of semantic web human resource résumés for sustainable talent acquisition." *Sustainability* (Multidisciplinary Digital Publishing Institute) 11 (2019): 3520.

Paredes-Valverde, Mario Andrés, María del Pilar Salas-Zárate, Ricardo Colomo-Palacios, Juan Miguel Gómez-Berbís, and Rafael Valencia-García. "An ontology-based approach with which to assign human resources to software projects." *Science of Computer Programming* (Elsevier) 156 (2018): 90–103.

Punjani, Dipti N., and Kishor Atkotiya. "A Comprehensive Study of Various Classification Techniques in Medical Application using Data Mining." 2018.

Punyakanok, V., D. Roth, and W. Yih. "The Importance of Syntactic Parsing and Inference in Semantic Role Labeling." *Computational Linguistics* (Computational Linguistics) 34 (2008).

Quinlan, J. Ross. *C4. 5: programs for machine learning.* Elsevier, 2014.

Rodriguez, Leah G., and Enrico P. Chavez. "Feature selection for job matching application using profile matching model." *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS).* 2019. 263–266.

Roy, Pradeep Kumar, Sarabjeet Singh Chowdhary, and Rocky Bhatia. "A Machine Learning approach for automation of Resume Recommendation system." *Procedia Computer Science* (Elsevier) 167 (2020): 2318–2327.

Ruuska, Salla, Wilhelmiina Hämäläinen, Sari Kajava, Mikaela Mughal, Pekka Matilainen, and Jaakko Mononen. "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle." *Behavioural processes* (Elsevier) 148 (2018): 56–62.

Salappa, A., Michael Doumpos, and Constantin Zopounidis. "Feature selection algorithms in classification problems: an experimental evaluation." *Optimisation Methods and Software* (Taylor & Francis) 22 (2007): 199–212.

Schröer, Christoph, Felix Kruse, and Jorge Marx Gómez. “A Systematic Literature Review on Applying CRISP-DM Process Model.” *Procedia Computer Science* (Elsevier) 181 (2021): 526–534.

Solutions, Alternate Computing. *JSchnizzle.* Alternate Computing Solutions Inc. Bear, Delaware, 2017.

Srivastava, Shweta. “Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining.” *International Journal of Computer Applications* (Citeseer) 88 (2014).

Strohmeier, Stefan, and Franca Piazza. “Artificial intelligence techniques in human resource management—a conceptual exploration.” In *Intelligent techniques in engineering management*, 149–172. Springer, 2015.

Swanson, Don R. “Searching natural language text by computer.” *Science* (JSTOR) 132 (1960): 1099–1104.

Taymouri, Farbod, Marcello La Rosa, Marlon Dumas, and Fabrizio Maria Maggi. “Business process variant analysis: Survey and classification.” *Knowledge-Based Systems* (Elsevier) 211 (2021): 106557.

Tsai, Hsien-Tang, Herbert Moskowitz, and Lai-Hsi Lee. “Human resource selection for software development projects using Taguchi_s parameter design.” *European Journal of operational research* (Elsevier) 151 (2003): 167-180.

Turing, A. “Computing Machinery and Intelligence, in ed. Epstein, R., Roberts, G., and Beber, G.” *Parsing the Turing Test–Philosophical and Methodological Issues in the Quest for the Thinking Computer*, 1956.

Turing, Alan M., and J. Haugeland. *Computing machinery and intelligence.* MIT Press Cambridge, MA, 1950.

Vanam, Murali Krishna, Barkat Amirali Jiwani, Arelli Swathi, and V. Madhavi. “High performance machine learning and data science based implementation using Weka.” *Materials Today: Proceedings* (Elsevier), 2021.

Verma, Archit. “Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA.” *IJACSA) International Journal of Advanced Computer Science and Applications*, 2019.

Vijayarani, S., R. Janani, and others. “Text mining: open source tokenization tools-an analysis.” *Advanced Computational Intelligence: An International Journal (ACII)* 3 (2016): 37–47.

Visa, Ari, Jarno Einolander, and Hannu Vanharanta. “New tools to help in the recruitment process.” *Procedia Manufacturing* (Elsevier) 3 (2015): 653–659.

Vuk, Miha, and Tomaz Curk. “ROC curve, lift chart and calibration plot.” *Metodoloski zvezki* (Anuska Ferligoj) 3 (2006): 89.

Witten, Ian H., and Eibe Frank. “Data mining: practical machine learning tools and techniques with Java implementations.” *Acm Sigmod Record* (ACM New York, NY, USA) 31 (2002): 76-77.

Wong, Tzu-Tsung. “Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation.” *Pattern Recognition* (Elsevier) 48 (2015): 2839-2846.

Wyse, Adam E. “Analyzing job analysis data using mixture Rasch models.” *International Journal of Testing* (Taylor & Francis) 19 (2019): 52–73.

Yahiaoui, Soumaya, Christophe Courtin, Pierre Maret, and Laurent Tabourot. “Decision-making system for recommending and evaluating competences networks based on interaction data.” *Applied Network Science* (SpringerOpen) 4 (2019): 1-15.

التوظيف الذكي بإستخدام الذكاء الاصطناعي وتنقيب البيانات

أعدت من قبل

دارين علي محمد ابو ربيع

اشرف عليها

أ.د محمد الفيومي

المشرف المشارك

د احمد بني مصطفى

الملخص

عمليه التوظيف هي من أهم القرارات التي يمكن أن تأخذها اي مؤسسة وأكثرها صعوبه, فعمليه التوظيف يجب أن تهدف الى الحصول على الاشخاص الأكثر ملائمة للوظيفة والأكثر قدره على تحقيق أهداف المؤسسة وأهدافها الاستراتيجيه.

فتوظيف الشخص الغير مناسب قد يؤثر على سمعه المؤسسة، وقد يؤدي إلى اشكالات ونزاعات قانونيه، والي خسائر لايمكن تحملها من قبل المؤسسه.

تقدم هذه الدراسه حلا لهذه المشكله عن طريق تو؛ يف تقنيات الذكاء الاصطناعي في عمليه التوظيف والتي تشمل التعلم الآلي، تنقيب البيانات ،معالجه اللغات الطبيعيه، وذلك لتجنب التحيز واختيار افضل المرشحين المناسبين للوظيفة مما يساعد على زياده فعاليه عمل المؤسسة وضمان نموها وازدهارها.

يتضمن الحل المقترح استنباط اهم العوامل والمؤشرات المرتبطة بنجاح الموظفين في المؤسسة، وذلك من خلال تحليل بيانات الموظفين الحاليين، باستخدام نماذج تنقيب البيانات، ومن ثم استخدام هذه العوامل والمؤشرات في صياغه الوصف للوظيفة المطلوبه وتحديد متطلباتها وشروط التقدم اليها, ثم مطابقه السير الذاتيه للمتقدمين لهذه الوظيفه آليا.

تم في هذه الدراسه تصميم ثلاثه تجارب تطبيقيه لاختبار هذه الطريقة باستخدام مجموعة من البيانات التي تم الحصول عليها من دائره الاحصاءات العامة الاردنيه، باستخدام السجلات الوظيفية ل 529 موظف, موزعه على 19 خاصية، حيث تم بناء 9 نماذج تعلم الي في كل تجربه.

استخلصت هذه الدراسه ان افضل هذه النماذج هو الذي تم بناءه باستخدام K-Nearest Neighbours (KNN), حيث تم الحصول على دقه تصنيف تبلغ 91٪، يليه النموذج الذي تم بناءه بواسطه Random Forest (%89) , Random Committee (%86).

اما في عمليه المطابقة للسير الذاتيه مع المؤشرات المستخلصه من عمليه تنقيب البيانات، فقد تم الحصول على دقه تعادل 80٪، علما بأن هذه النتائج هي أفضل من تلك التي تم الحصول عليها في الدرسات المشابهه.