

**SOLENT UNIVERSITY, SOUTHAMPTON**  
**FACULTY OF BUSINESS, LAW AND DIGITAL TECHNOLOGY**

**Msc Applied Artificial Intelligence and Data Science**

**Academic Year 2021-2022**

**Research Methods AE1**

**Topic: Improving Candidate Ranking for IT Talent -  
A Python Based Solution Approach**

**Student no. Q15684091**

**Supervisor : Drishty Sobnath**

**Date of submission : September 2022**

**This report is submitted in fulfilment of the requirements of Solent University  
for the degree of MSc Applied Artificial Intelligence and Data Science**

## Acknowledgements

First and foremost, I am grateful to the All-Powerful God for providing me with a sound mind and good health so that I could work on and complete my project.

I would want to express my profound gratitude to my husband Patrick, son Jayden, and all my family and friends for their unwavering support and tolerance.

I want to express my sincere thanks to Drishty Sobnath, my supervisor, for supporting me with my master's research and study, as well as to Femi Isiaq, the module leader for the thesis, for the leadership role he played and for the insightful lectures.

## Abstract

Technology implementation in Hiring Processes has evolved at a fast pace - especially in the last century with the constant innovation in Information Technology. The technology has continued to change but the application is greatly impacted by human interruption. Experts at US Firm Korn Ferry have predicted a global shortage of human capital in the Information and Technology sector up to a deficit of 85 million jobs by 2030. Countries like the United Kingdom and the United States of America are already experiencing the shortage - hence setting up immigration programs that attract such talent.

If a prospective client will hire a plumber based on their skills, or a hire a landscaper based on evidence of previous work without asking for a formal certificate from an ivy league college, why should IT job roles be treated differently? The IT sector is an industry that allows for supervised growth - which means that a freshman spends a certain period before becoming a junior developer (all activities supervised by a senior), just like other skills-based professions - why then should we screen IT talents in the same way as marketing, and operational roles. Social changes also mean that a lot of people are acquiring skills in more non-formal ways than we have always known.

This study believes that a major way to resolve this situation is minimizing exclusion by screening candidates based on their skills and work experience - rather than institutions attended or grades achieved via supervised exams. It is a common belief that humans are responsible for the bias in such screening systems, it is also possible that there is a huge knowledge gap between the Candidates, the algorithms and the software developers. Industry benchmarks that are seen as best practices have largely contributed to this trend.

This project successfully developed a python-based AI resume screening solution, that focus on screening talent specifically for IT roles, based on their acquired programming skills, evidence of work and experience. This does not

mean that transferable skills are not necessary, but the candidate at least gets a shot at being known and recognised for their knowledge - thus expanding possibilities for qualified candidates.

## Contents

Acknowledgements .....	i
Abstract.....	ii
1. Introduction .....	3
1.1 Aims.....	7
1.2 Objective.....	7
1.3 Assumption.....	8
1.4 Limitations.....	8
2. Literature review.....	8
2.1 Evolution of candidate screening.....	9
2.2 Benchmarking.....	10
2.3 Structural analysis of text as data.....	11
2.4 System environment analysis.....	12
2.5 Predictive analysis.....	13
2.5.1. Data Modelling.....	14
2.5.2. DataMining.....	17
2.4 Cross-validation.....	19
2.4 Python programming.....	20
3. Methodology.....	20
4. Exploratory Data Analysis.....	21
4.1 Control Data set.....	21
4.2 Univariate analysis.....	22
4.2.1. Histogram.....	22
4.2.2. Donut plot.....	23
4.3 Qualitative data analysis.....	24
5. The Model.....	25
5.1 Execution.....	28
5.2 Application tools.....	29
6. Result.....	29
7. Discussion.....	35
8. Conclusion.....	36
9. Reference list.....	37
Appendix.....	44

## List of Tables

Table 1 Data Structure ..... 1Error! Bookmark not defined.

## List of Figures

Figure 1 Result of study on which qualifications are most important when evaluation technical talent .....	4
Figure 2 Sample resumes with “coding bootcamp” education experience .....	5
Figure 3. Hierarchical Model.....	15
Figure 4 Relational Data Model.....	16
Figure 5. Multiple levels of logical view.....	17
Figure 6. Data Mining as a stage in the knowledge path discovery.....	18
Figure 7. Word count of categories of skills .....	23
Figure 8. Distribution of categories of skills .....	24
Figure 9. Word cloud .....	25
Figure 10. Model .....	26
Figure 11. Word count of categories of skills .....	23
Figure R1. Home screen to the web App?.....	31
Figure R2. Candidate ready to Upload resume/profile/cv.....	31
Figure R3 Uploading from Local Drive.....	32
Figure R4. Upload Complete.....	32
Figure R5. Resume Analysed and showing skills.....	33
Figure R6. Admin Log In .....	33
Figure R7. Admin Dashboard - list of candidates who have provided their details....	34
Figure R8. Expanded Admin Dashboard .....	34
Figure R9. Possible skill role predictions form a candidate.....	35
Figure R10. Candidate Experience level based on resume analysis .....	35

## 1. Introduction

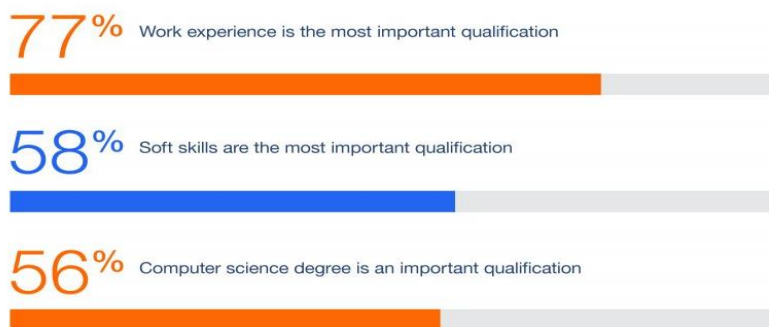
According to United States top consulting firm, Korn Ferry, there will be a shortage of over eighty-five million Information Technology (IT) workers by 2030 (Alan Guarino et al, 2022). This study was presumably focused on the workers not having the skills to execute the job, but there might be other factors or practices that increase the speculation size of this result. The United Kingdom Employment Rights Act 1996, defines a worker as an individual who has entered into a contract to deliver a service for the purpose of earning a reward (Crown, 2022). To satisfy this definition, intending workers are screened and selected by the employer or client - usually from a large pool of applicants. IT workers are not exempted from this process. Unlike other technical skills such as: carpentry, plumbing, photography, painting, acting, to mention a few, IT talents are most likely to be subjected to academic results or institutions attended, rather than focus on skill and quality of work.

With the continued advancement in learning, more people are learning through bootcamps, self-paced courses and Do It Yourself (DIY) platforms like YouTube, LinkedIn, Google Learn digital, IBM Skills, and the list grows. The influence of technology enabled learning have received much attention: “A report from the 7th European Conference on Technology Enabled Learning; 21st Century Learning for 21st Century Skills” (Ravenscroft, et al, 2012) explored the various concepts of technology enabled learning and why they are more desirable and getting popular. Also, Samuel Kai Wah Chu explored innovations that supports project-based learning in their book, “21st Century Skills Development Through Inquiry-Based Learning: From Theory to Practice” (Chu, et al, 20017). Agile “Triad” (Ken Pugh, 2011) have allowed software developers to gain immense knowledge even while working as Juniors in a project. Hence, increased online collaboration between software developers, utilizing available social tools to connect and build solutions. Such projects can easily appear on a work portfolio, but may not make it to the resume since it has no mainstream

contract. This has also had a huge impact on the way various IT talent prepare themselves for the workplace. This however does not say that they have not acquired the skill or participated in reasonable work.

An Indeed.com study that interviewed 1009 employers involved in human resources practice, reported that 86% of the respondents alluded to the fact that finding and hiring tech talent was very challenging (Indeed, Editorial Team, 2016). It also reported that 24% of respondents insist that the type of school attended, while 58% agreed that soft skills are more important than evaluating technical capabilities. Although 77% extoled the evaluation of technical capacity, the one question to ask would be, how do they evaluate this pool of talents? The number of years in employment does not always translate to the quality, quantity and their impact on the projects they have participated in. In the face of a global IT talent shortage, it becomes important to explore all opportunities to ensure that relevant skills are active in the IT talent marketplace. In many cases, the candidates do not get a foot in the door because of their academic background, test scores, previous employer and many more attributes that may necessarily not define their capacity to execute.

Which qualifications are the most important when you evaluate technical talent?



Source: Indeed Research

indeed

This bar graph shows the results of a survey conducted by Indeed October 27, 2016 to November 1, 2016. According to the data, 77% of respondents find that work experience is the most important qualification to evaluate technical talent, 58% of respondents say that soft skills are the most important qualification and 56% of respondents say that a computer science degree is the most important qualification.

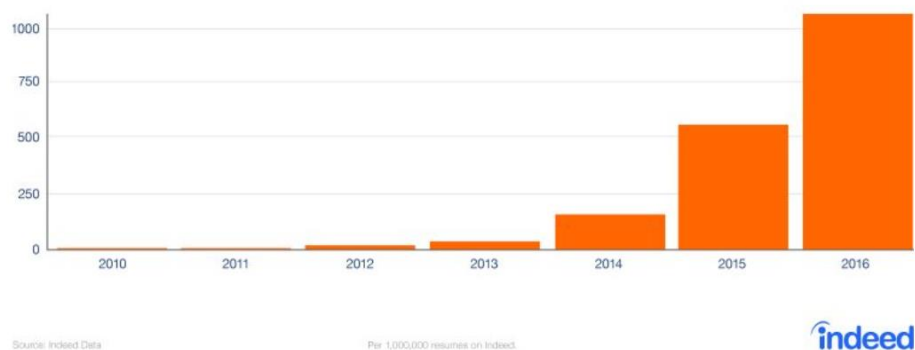
Fig 1: Result of study on which qualifications are most important when evaluation technical talent



Regardless of the number of projects a candidate completed in the past, if they cannot get through the first layer of screening, which is a shortlisting of “high quality” candidates, they cannot be employed. In many situations this definition is based on the resume, which according to job application “best practices”, candidates are advised to be as short and brief as possible. Thus, if the first level of screening an IT talent with a vast portfolio of work is based on such document, then we are intentionally contributing to the gap in available tech skills within our workplaces. Although Indeed.com reported the double growth of considerations for candidates from technical schools with software and engineering expertise, if these people are placed on the usual screening scales, their employment will only be as case of special consideration rather than merited offering.

---

#### Resumes with “coding bootcamp” education experience



---

*This bar chart shows that since 2010, resumes with “coding bootcamp” as an educational experience have grown exponentially, doubling from 2015 to 2016.*

*Fig 2: Sample resumes with “coding bootcamp” education experience*

This approach to hiring IT talent has not only contributed to the reduction in tech talent within the industry, it has also impacted the quality of existing talent. If a fresh talent establishes the fact that all they need to get a high

paying job is academic excellence, they will focus more on scoring high academic grades instead of honing and developing their skills through execution of projects. This will ultimately increase the time a talent can grow from junior to intermediate and eventually senior software developer stage. This practice also continues to hurt small businesses and developing countries because each time their tech staff attains a certain senior level, they move on to a bigger organisation or developed nation for better rewards. There is then a need for practitioners to encourage a new way of first-level screening for IT talents. It is my belief that, with the right dataset attributes, an algorithm can be trained to aggregate and analyse information that will enable recruiters stack candidates based on the quality of their impact on their projects. by screening their portfolios.

Another challenge for first level screening is that, it may seem that the practitioners believe they have things sorted once they shortlist the best grades or academic institutions and may have foreclosed a revisit from design. Slavadore Falletta was one of the early investigators of innovations in human resource management, especially the introduction of predictive analytics; He coined the phrase “HR intelligence” which he described as, “a proactive and systematic process for gathering, analyzing, communicating, and using insightful people research and analytics results to help organizations achieve their strategic objectives” (Falletta, 2008, p. 21). In a 2015 HR magazine article, he argued against Karl-Heinz Oehler’s position supporting the unfettered use of predictive intelligence in HR management (Falletta & Oehler, 2015). Oehler was reflecting from a point of optimization, discussing the operational efficiency and financial implications of using information gathered during recruitment to manage or improve internal talent outcomes. While Falletta was keen on how such practices can create a state of generalization or possibly punishing an individual for outcomes they did not contribute to, but have to live with. That argument, although insightful about the impact and

bias of predictive analysis - when used in entirety, completely focused on operational outcomes for the organisation's staff and super silent on first level screening. Other reports such as, "The Rise of Analytics in HR -The era of talent intelligence is here" (Chen et al, 2017) believe that this is a time organisation must invest in data analytics literacy as well as encourage their people to be data first. This sort of thinking can create a positive pathway to understanding how and why to use data in a justifiable manner.

This study explores a path to improving candidate screening algorithms to enable first level screening of IT talent based on their work quality and Impact. It seeks to make a case for inclusive hiring by placing IT jobs side by side with similar skill-focused jobs while highlighting the new order of learning that has allowed IT talent gain requisite experience without necessarily acquiring top degrees from elite schools. Hence the study set out to achieve the following aims and objectives:

### **1.1 Aims**

- i. Explore a python-based algorithm that can improve the quantity and size of available IT talent in each recruitment exercise.
- ii. Improve inclusivity in hiring processes by implementing a machine learning-supported process, that will guard against the exclusion of IT applicants with impactful and relevant work portfolios - occasioned by screening tool inadequacy.

This will help to discover the possible limiting data attributes that reduce the quality and impact of Information Technology in the first-level screening of IT candidate resumes.

### **1.2 Objectives**

- i. Develop and test a python-based algorithm with efficiency to crunch text bundles.
- ii. Present a comparison between data results for sample resume dataset and portfolio screened dataset.

- iii. Imply possible ways of integrating new program with existing hiring solutions.

### **1.3 Assumptions**

- i. It was assumed that the sample dataset used for this study reflect global qualifications in our present-day reality.
- ii. It was assumed that the keywords assigned for screening tests are standard items and will be understood by HR practitioners and policymakers in the industry.

### **1.4 Limitations**

This study is limited to first-level screening of candidates who are being considered for core Information Technology jobs functions - with special reference to software and hardware maintenance skills. Other job functions such as; sales, customer service, general operations, and human resources, within the IT industry are excluded from this study because they can easily be transferred from one sector to another.

## **2. Literature Review**

This report was preceded by a pilot study conducted and completed in July 2020 and fully integrated into this report for consistency. The pilot study allowed us to investigate industry practices and review existing learnings which enabled us to lay an execution pathway for this study. In this chapter, we review the key aspects of data science that contribute to the final result of data analysis and review existing practices that have shaped the use of analysis in HR management and by extension first level decisions in hiring practices. In essence, these factors contribute to our understanding of how our objectives can be met through the implementation of an improved algorithm for the same purpose.

## 2.1 Evolution of Candidate Screening

As in every kind of competition, the umpire - in this case hiring organisation, will set out their rules of engagement and contestants are required to present their submissions including evidence of claims to be considered. If a contestant does not meet basic stipulated requirements, they have no chance - unless that requirement is modified to accommodate their class. Over the years HR practitioners have continued to evolve and their evolution received a great leap with the advancement in technology - especially data science. Burdened by a large number of applicants and cumbersome paperwork, HR practitioners have continued to find better ways of winning the talent war. A 2008 study on the UK recruiting practices by Mohamed Branine highlighted the changing attitude of employers from focusing on mere qualifications to focusing on the individual by gravitating towards transferable skills (Branine, 2008). Although the study was largely tilted towards fresh graduates, it still pointed out the fact that employers were hooked on the classification of the qualification more than the subject of study. Much of that has not changed when compared to John Spahic's study in 2015 (Spahic, 2015), which explored HR practices in fortune 1000 companies. Organisations have continued to focus on finding new sourcing grounds and improved technology with very little consideration to distinction for subject structure. While discussing the workforce transformation Initiative at Columbia University Business School, Dr. Ann Bartel, shared her excitement about custom programmes, which allow the academia to work with corporations in developing a certain breed of workforce (The European Business Review, 2019). This is a leap from the "Milk round" process that involved organisations making presentations at top universities to attract talent. We have also seen organisations like Microsoft, Google and their likes develop programs such as the Microsoft Learn Student program and Google Student Ambassador program. These programs however have a limitation on individuals who have acquired skills and executed tech focused projects outside the formal academic setting.

Very little interactions have happened between Technical and Vocational Education Training (TVET) and the formal workforce - with exception of countries like Austria, Denmark, Germany and Switzerland (Caves, et al, 2018). Considering that technology might have moved faster than human evolution, there could be a possibility that innovation in hiring algorithms can once again solve the human decision drag.

## 2.2. Benchmarking

It is not unusual for practitioners in any industry to set a standard of operations. A lot of times these standards grow as a culture before they are formally backed by legal affirmations which might stem from government regulations or industry associations. To understand the impact and the various verticals of benchmarking in play, let us first understand what it means to benchmark. Watson, in his revised book on strategic benchmarking described it as the application of objective measurement and scientific methods to analyse existing practices with the aim to determine procedures for improving an organisation's output and processes (Watson, 2007). The first question will be what are the items to be analysed? This definition simply points to the fact that benchmarking cannot be a one-direction affair, it involved multiple stakeholders including external interferences to results. Because it is an intentional effort to improve performance, organisations are always aiming to reach a minimal threshold that will place them in a better position to compete with others. This continuous strive to practice at a certain range of performance eventually leads to a relatively uniform practice that may only defer by execution time. This situation is exactly the state of recruitment across the IT ecosystem: there is a talent war being fought within a defined spectrum which it came about from the need to adapt to "best practices" and the stakeholders are dead focused on grabbing the best of the industry without minding a path of replacement. In order to consider a revision of current HR practices, it was important to understand the contribution of screening

software and algorithms to the competitive landscape of IT talent acquisition - especially in the first stage of candidate screening. This made it possible to understand the possibility and how best an organisation should introduce a new practice and maintain competitiveness in the industry (Hines, 1998). It is also important to note

Since benchmarking stakeholders (HR competitors and software companies) are understood, the first stage of the solution benchmarking process is the collection of industry data (Kodali, 2008). The data collection process for such will be qualitative (Jones, 1995) because we will require to understand details about their adoption as well as operational needs. The information is derived on the basis of general market reviews from acknowledged databases such as; Capterra, G2 Crowd, Software world and benchmarking reports practitioners like Best Practices, LLC. The gathered data does not seek to compare these solutions, it is simply to understand the structure and likeliness of their operations towards establishing a benchmarked bracket (real or imaginary). The analysis will review the first screening stage quality of stakeholders' solutions and compare it with our target functionality proposition. Since benchmarks do have their challenges, it is then important to consider long-term impact when developing an alternate software (Kumar and Harms, 2004).

### 2.3 Statistical Analysis of Text as Data

Various scholars have defined the term “data” as a way of representing and storing information - especially in computing. They are facts and statistics aggregated for referencing or analysis, they are assumed to be true and create a basis for reasoning or calculations (Farouk, 2017). As with every other form of information, it will need to be processed to make it useful to the entity involved - this implies that, the collector, the processor and the end Candidate must use uniform language to reduce unintentional bias. Our study is largely focused on text analysis for economic use; hence this review will seek to

understand how data is processed and presented through machines (computers). Gentzkow et al in a well-researched paper, “Text as Data” had this to say about human data analogy;

*“When humans read text, they do not see a vector of dummy variables, nor a sequence of unrelated tokens. They interpret words in light of other words, and extract meaning from the text as a whole”* (Gentzkow et al, 2019)

This statement may well explain how HR practitioners view the selection of keywords to be used in first-level screening and shed light on the issue we are trying to deal with in this study. The multidimensional nature can also make it more challenging when considered from a literal point of view. Today, Machine Learning and Artificial Intelligence have created a convergence, where text can be analysed alongside every other kind of data by providing encoding solutions like labelEncoder (Pedregosa *et al*, 2011) and its likes to enable analysis of text-based descriptions. But there still exists a gap between the Candidates of these human resource management (HRM) solutions, and this may be as a result of available features - not the inability of scientists to innovate for optimal inclusion. and the understanding of how they present output. It could appear that the challenge is centred around how the Candidates are able to use the many tools available for their analysis data - especially in the human resource environment.

## 2.4 System Environment Analysis

Since we have elected to develop a solution that might require implementation within the industry in future, it becomes important to take a brief walk in understanding the concept of environmental analysis and how it might impact adoption of our solution. The system environment analysis provides the basis for the structural requirements analysis. Subsequently, the system environment analysis has to examine its compatibility to program and implement those



functions in relation to existing software solutions within the adoption environment. This process is highly relevant for the requirements sourcing process due to the necessity to align such application within the existing system environment (MacLean et al., 2004).

An organisation's IT infrastructure consists of interacting subsystems which are part of the operative business (Singer et al., 2009). Embedding changes within such an IT environment which involves other entities creates constraints which are relevant due to the question whether the client is able handle these constraints (MacLean et al., 2004).

In the case that the IT processes are administered by an external provider, it might be necessary to consult the provider as an additional information source. Also, already existing system manuals can be drawn upon to collect information (Vijayan and Raju, 2011). On basis of the information gathering process, a before & after graphic of the IT environment is prepared. One of the interesting approaches to requirement elicitation is the concept of "structural holes" as applied in the development of Open-Source Software Development (Bhowmik, et al, 2015). It basically describes the process of developing requirements in an Open-Source Software (OSS) environment, where individuals within the social space contribute requirements without any formal reliance. Reflecting this model in the Brokerage business, Ronal Burt pointed out that the ideas are contributed and replaced in the social network without necessarily disrupting the flow of activity (Burt, 2001) I see this as a unique way of creating and sustaining innovation without the fear of losing convention.

## 2.5 Predictive Analysis

In this section, we will seek to understand predictive analysis for what it is and then align it to our focus area - HR screening for IT job roles. The challenge of today stems from the extent to which human decisions have affected the adaptation of predictive analysis. Based on the submission from Jac and John, Predictive analysis can be described as a process that utilize multiple

techniques including; data mining, modelling and statistics to review historical and present data for the purpose of predicting future outcomes (Fitz-Enz, et al, 2014). Understanding these techniques aided this work in realizing where modifications can be made in the approach to the practice towards achieving desired results. Since predictive analysis is a case of “probabilities and potential impact” (Fitz-Enz, et al, 2014, p. 29) it is then important to include as much detail as possible about the subject of review. According to Michael Walsh, the result an organisation can achieve depends on how well they apply relevant collected data. Walsh described the application of analytics in human resource management as an art and a science whose outcomes can be different for different Candidates (Walsh, 2021). He concluded that, “within the realm of staffing, analytics can help, us to find better candidates in a shorter period, maybe. We need to make sure that we are using analytical techniques properly” (Walsh, 2021, p. 46).

#### 2.5.1 Data Modelling

A data model can be described as Data modelling means formulating every step and gathering the techniques required to achieve Candidate goals based on a particular data set. Since all the calculations cannot be performed once or in parallel, one needs to list down the flow of calculations - which simply means, pathing the steps to the solution needed. It is a critical step in developing any predictive process, because it helps us understand how parameters are achieved for every result stage. It can also be very useful when implementing modifications to an existing solution - with a good understanding of the model, one can implement modifications without disrupting the entire system, which saves time and cost. Existing data modelling techniques include, the basics; Hierarchical Data Modelling, Relational Data Modelling, and Entity-Relationship Data (ERD) Modelling, Start Schema etc. Industry documentation have always confused the naming of how these models are presented with what they

mean as a form of definition, we will attempt to focus more on describing the key models and representing with diagrams.

Hierarchical Data Modelling is a model that presents data in the form of parent and dependencies. It is also referred to as linear and one of the earliest - it was developed by IBM (International Business Machines) Corporation in 1960. In a one-to-many presentations style, the parent data is in direct association with the child data points. The use of this model has since declined in organisations because it is strenuous to gain in-depth understanding of collected data in a one-to-many representation.

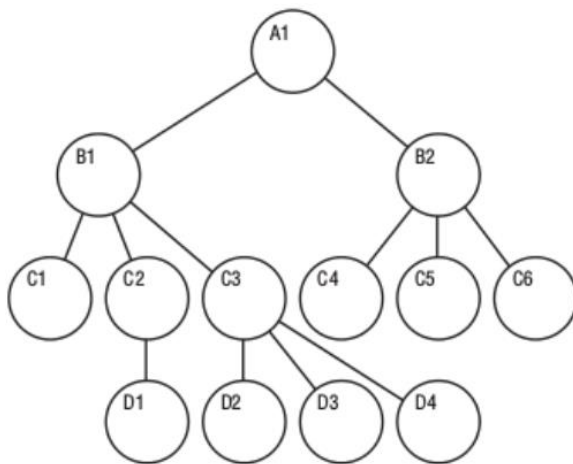


Fig 3: *Hierarchical Model* Image Source: MariaDB

Relational Data modelling is the most widely used data modelling in software engineering. It has its roots in models developed based on the relationship between datasets. It was introduced by Edgar Codd in 1970 and still remains the go to technique for most Candidates because of its

efficiency in managing complex data analysis.

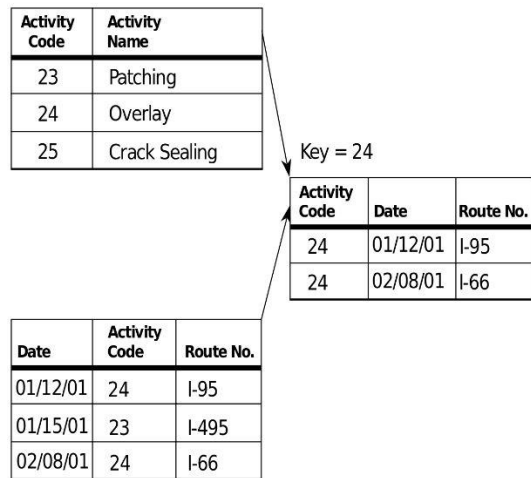


Fig 4: Relational Data Model Image Source: Polarwinco

The technique is executed by using structured query language (SQL) to obtain data in the form of tables while maintaining its relationships - ensuring consistency and data integrity.

iii. Entity-Relational Data (ERD) Modelling technique is a logical structure that allows for creating relationships between data points based on specific solution development requirements. Introduced by Peter Chen in 1976 (Chen, 1977), and presents a unification between relational model, the network and the entity set models. Chen sees it as a revolutionary milestone in data modelling because it embodies all the strengths in all three models - which at the time, were seen to have their own unique strengths and individual arguments created preferences based on their subjective needs. ERD implementation takes a view of understanding the various logical levels at which data can be represented minding their relationships without losing their unique identities. According to Piotr, ERD diagrams make it easier to see the big picture, understand table relationships while creating a possibility to use visual cues in communicating key information such as; location, colour, shape and

proximity (Kknow, 2017). Piotr also pointed out limitations such as the inability to handle large data models due to space constraint which leads to support of fewer details among other issues.

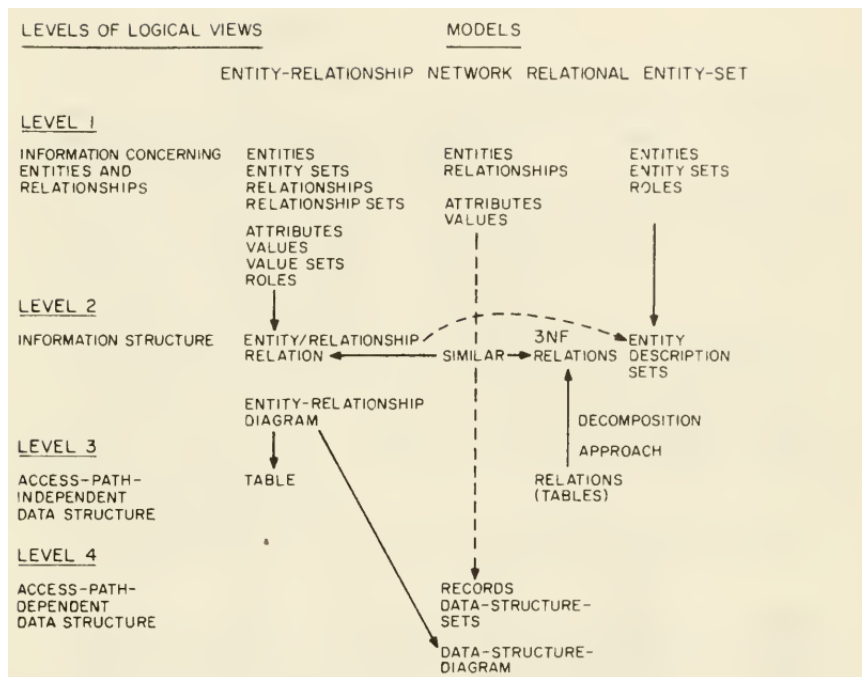


Fig 5: Multiple Levels of Logical View

Image Source: (Chen, 1977, p. 11)

## 2.5.2 Data Mining

Every decision stem from the information available to the decision maker. It is easier to process a very little information with the human brain at a given time, but if one has to run an enterprise or plan properly a large amount of information will be involved. Thus, the need to store collected information - in computer science, such information is stored as identifiable data with their unique properties. In order to make sense of the data, it must be arranged in a particular manner for easy access and processing. The larger the information the more queries are required for data usage, hence the necessity of analysing and understanding the stored information. This challenge led to the development of data mining (Han

and Kamber, 2006). Han and Kamber defined data mining as “the process or method that extracts or “mines” interesting knowledge or patterns from large amounts of data” (Han and Kamber, 2006, p. 3). This data can be accessed from various information repositories including; a database, data warehouse, and the web. It can also be streamed into the system in a dynamic way. The essence of data mining is to transform randomly collected or raw data into knowledge that can be processed for decision making.

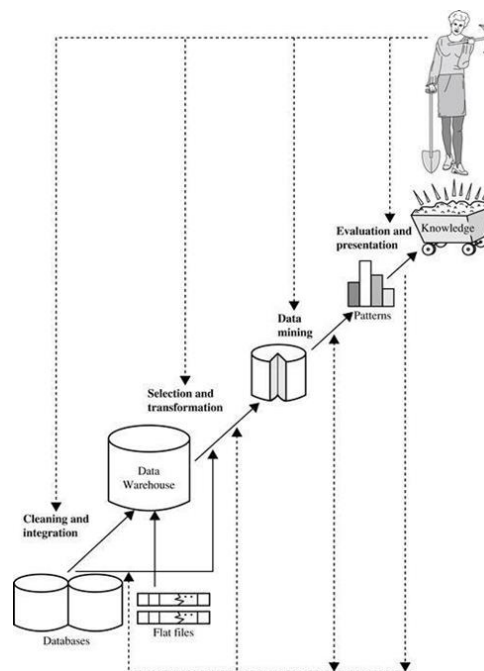


Fig 6: Data Mining as a stage in the knowledge path discovery

Image Source: (Han and Kamber, 2006, p. 11)

A standard data mining architecture for a HR tool will include; A data source - this could be a database of received resumes and portfolios, a data warehouse of historic sorted internal data, the web, or any kind of information repository that the Candidate has set up, a server that fetches the information from these sources, a knowledge base - which guides search in order to determine how interesting the data is and, in this case,

will include key text words, and a pattern evaluation module that helps to focus the search towards patterns. A data mining engine consisting of functional modules is also required to perform functions such as association, classification, cluster evolution, and deviation analysis. Finally, a graphical interface that will enable the Candidate interact visually with the data mining system.

Similarly, data mining execution follows a set of processes; first, the data is cleaned to remove noise or inconsistent data, this gives way for any form of data integration (combining data from multiple sources) that might happen. Relevant data is then selected and consolidated into forms (data transformation) appropriate for data mining. Data mining is executed by employing relevant methods to extract interesting patterns. The extracted patterns are evaluated based on interestingness measures that will identify interesting patterns representing the sourced knowledge. Finally, the mined knowledge is presented using a visualization and representation technique.

## 2.6. Cross Validation

To determine the genuine prediction error of models and to fine-tune model parameters, Cross-validation is one of the most used data resampling techniques. It also prevents overfitting - which is a situation that a model fits so well to the existing dataset but becomes alien to an unseen data (). While the procedure of cross-validation is similar to the random subsampling approach, no two test sets may overlap due to the manner the sampling is carried out. The learning set ( $D_{learn}$ ) is a dataset that may be used to assess a prediction model, and random subsampling techniques are used to produce the training set ( $D_{train}$ ) and test set ( $D_{test}$ ) from the learning set. There are so many subsampling methods, but distinguished by their process of execution. In

the 10-fold Cross validation, where  $k=10$ , Cross validation accuracy is achieved by taking the average of all ten accuracies achieved on the validation sets.

The cross-validated estimate of the prediction error is calculated thus;

$\hat{cv}$ , is calculated thus;  $\hat{cv} = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}_{-k}(x_i))$  (Hastie, et al, 2008)

## 2.7. Python Programming

Python is a high-level programming language, complex enough to perform all kinds of machine learning computations. Its extensive use has continued to ensure accommodation of all kinds of integration and easy set up for all kinds of application development. Since it is an interpreted language, it makes reduces the time spent on development because there is no need for linkages or compilations. Programmers may write logical, simple programmes for both small and big tasks using Python's strong organisational capabilities, such as functions, nested blocks, modules, classes, and packages. Python consistently uses objects and object-oriented programming and can provide the speed for computing intensive tasks because of its extension in C and C++ (Drake, et al, 2003). Being that our project is focused on text analysis Python becomes a go to tool for a successful implementation. According to Bengfort et al, considers text analysis techniques as primarily applied machine learning, they advised that a language rich scientific and numeric computing libraries is required (Bengfort, et al, 2018). Tools like Scikit-Learn, NLTK, Gensim, spaCy, NetworkX, and Yellowbrick are all among the libraries available Python's powerful suit of libraries.

## 3. Methodology

Methodologies for qualitative research were used in this study. Utilizing qualitative research methods can help you better comprehend the attitudes,



capacity and skills of candidates applying for jobs. We conducted our qualitative study using the method of observation.

As we commenced the final execution process of this project, we took into cognisance, our experience and findings during the pilot study. We had set out to modify an existing Python Algorithm originally designed to screen traditional resumes with the standard expectation and current practices. Our pilot study however posed the challenges of achieving this; which mainly will require a longer period to collate samples and try to show disparity - within a limited resource and time frame. But the basis for this project was to explore ways that individuals with requisite skills and experience can get through resume/profile screening without having to attend “top leagues” schools or gain exam based high scores.

#### 4. Exploratory Data Analysis

A careful review of data was performed utilising exploratory data analysis. the data adopted was cleaned and ensured it is in a suitable state with no missing values before beginning a data analysis or subjecting it to a machine learning algorithm. we took into account any persisting trends and noteworthy relationships that may be present in your data. In executing this project, we decided to follow a single proof path; that is to show that if we apply modifications to human decisions on algorithm key words, we can expand the array of qualified candidates who pass through the first screening cycle. Hence our algorithm goal was tilted towards key words that depict experience and knowledge of core IT functions.

##### 4.1 Control Data Set

	A	B
1	Category	Resume
2	Data Science	Skills * Programming Languages: Python (pandas, numpy, scipy, scikit-learn, matplotlib), S
3	Data Science	Education Details
4	Data Science	Areas of Interest Deep Learning, Control System Design, Programming in-Python, Electric Machinery, Web
5	Data Science	Skills R Python SAP HANA Tableau SAP HANA SQL SAP
6	Data Science	Education Details
7	Data Science	SKILLS C Basics, IOT, Python, MATLAB, Data Science, Machine Learning, HTML, Microsoft Word, Microsoft
8	Data Science	Skills Python Tableau Data Visualization R Studio Machine
9	Data Science	Education Details

Table 1: Dataset structure

Image Source: (Dutta, 2022)

Our test data contains twenty-five job roles creating twenty-five categories with 963 entries. The data sample is a public domain data made available for practice on Google Datasearch (Dutta, 2022). The dataset contains a sizeable IT job role attribute which will allow our test study to determine the kind of information collected at first-level resume screening. It is also a role agnostic information scrape showing no distinction in patterns. The results formed the basis for assigning ground truth attributes our solution.

## 4.2. Univariate analysis

Each variable in a data set is individually analysed in a univariate analysis. Univariate analysis checks at the values' central tendency as well as their dispersion. It gives an account of how the variable has been handled. Each distinct variable is explained. Each variable in a data set is individually analysed in a univariate analysis. It looks at the values' central tendency as well as their dispersion. It gives an account of how the variable has responded.

### 4.2.1. Histogram

As a histogram spanning categorical variables, we utilised a count plot. This displays the frequency with which each skill category appeared in the data. Figure 7 illuminates top two most prevalent talents, testing and Java

Developer, are ranked at 84 and 70, respectively. The SAP developer category has the lowest number of entries.

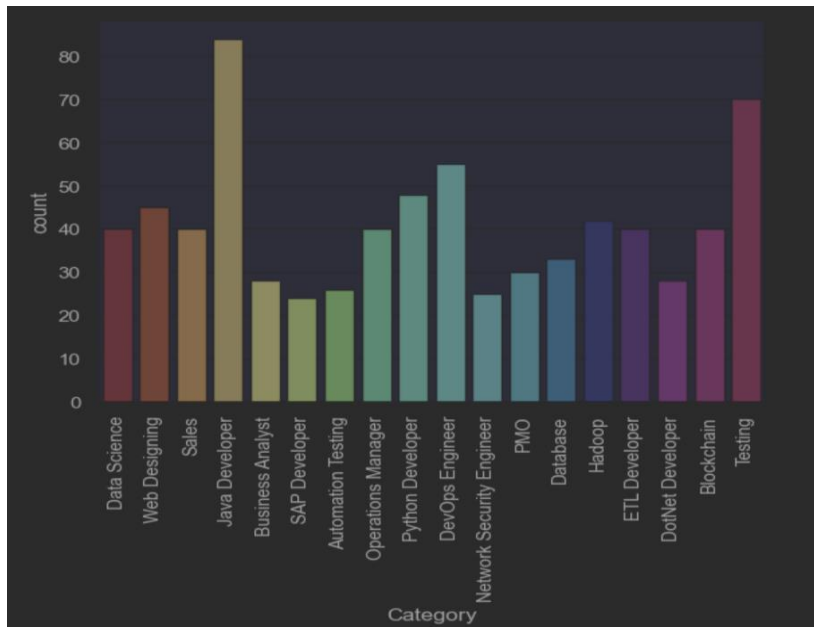


Fig 7: Word count of categories of skills

Image Source: refer to codes

#### 4.2.2 Donut plot

A donut chart is simply a pie chart with the central portion cut off. It was used to examine the distribution of the skill categories. Figure 8 below shows that web design and data science account for the biggest percentage of dispersion. Among 16 other categories, both held 20.9% of the total.

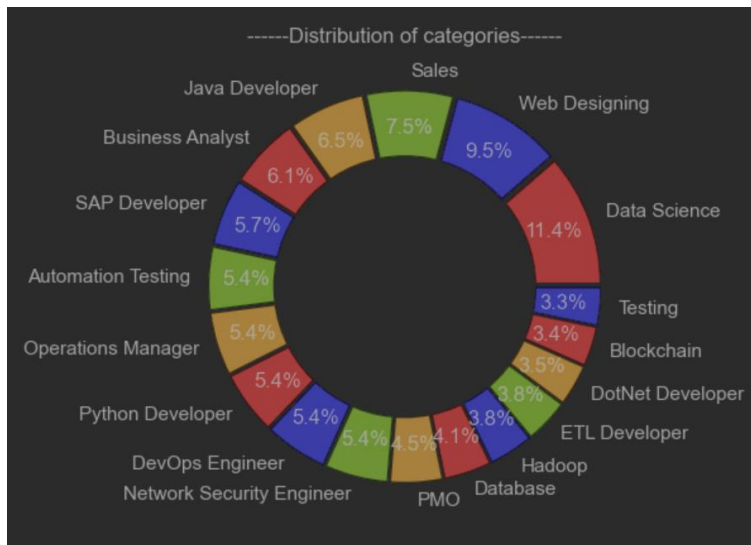


Fig 8: Distribution of categories

Image Source: refer to codes

### 4.3 Qualitative data analysis

In order to show qualitative data, word clouds are employed. Word clouds are a spatial information visualisation technique that organises text-based content into alluring visual clusters for qualitative data analysis. Word clouds, which are just collections of globes in different colours, show how frequently certain words are used. The phrases' sizes differ based on how frequently they appear in the data source; the more frequently a phrase was used, the bigger it is.

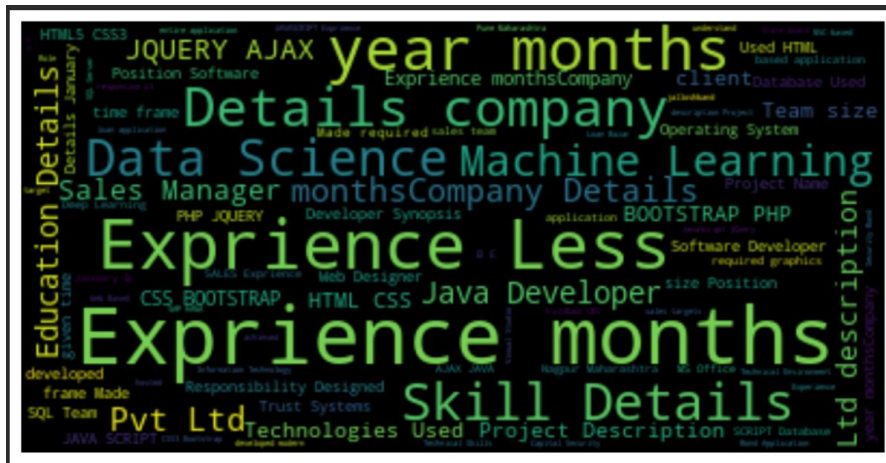


Fig 9: Word cloud

Image Source: refer to codes

As can be seen in the figure 9 above, the key words that are frequently used are "Experience," "less," and "months." We decide to develop a software that can recognise the actual skills required for a job from a resume in order to prevent distortion brought on by these phrases, which will have no influence on our study.

## 5. The model

Based on the iterations mentioned in above, we are now proceeding with building a python-based AI Portfolio Analyser. It is developed in such a way that a Candidate can upload their resume without recourse to the backend activity.

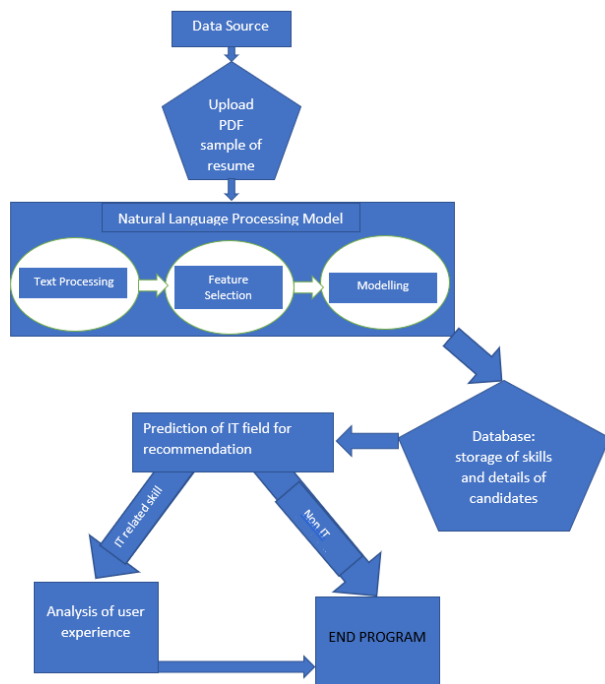


Fig 10: Model

Our AI analyser will receive and analyse the uploaded resume using Natural Language Processing (NLP) In this case we will specifically use Name Entity Recognition.

The following information will be captured;

- Candidate’s Name
- Candidate’s Email Address,
- Number of Pages on Resume
- Candidate’s Phone Number
- The Candidate’s skill set.

It also predicts the experience level of the Candidate;

- Beginner level
- Intermediate level

- Experienced level)

Based on the skill set in the Candidate's resume, it could also predict the kind of job the Candidate is looking for.

The system is programmed to suggest resume booting activities such as new courses the candidate might attempt to increase their resume score in future applications. If the candidate is having any recommended tips, it's also going to appear in green. With this tip, the candidate could further improve their resume by adding the recommending tips to their resume.

The resume is then scored based on the skill sets available in the candidate's resume, which will be communicated to a knowledge base of key words set preset in the solution. These keywords are based on our learnings from our control data, which enabled us to understand the different information recorded as skills for a particular job role.

The Software programme consist of two parts;

- There's a Candidate path
- Admin path

The Admin dashboard allows authorized persons to view all resumes uploaded by candidates through mobile Web/App interface. All analytical data are also available to the Admin. The report on a batch or an individual can be read and downloaded. We applied relational database management for storing and accessing our data.

More so, all the information extracted from the Candidate's resume will be connected to a SQL database which allows for storing of important data should we decide to carry out more for in-depth analysis or use ML on the data. So, based on the Candidates that are coming in to the page to input the data, we'll be able to collect their email, their name, their resume score, the timestamp,

predicted field, Candidate's level of experience, actual skills, recommended skill, etc.

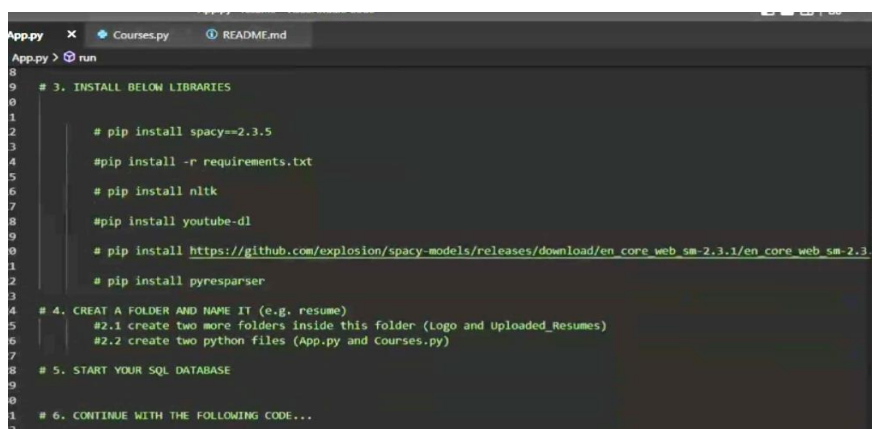
With this, before the candidate even appeared at the interview, we have everything that we need. With this strategy, we'd be able to analyze and know the best candidate

## 5.1 Execution

The first thing was to source for sample resumes which we were web generated based on real human sample resumes found on resource data from a public repository. We sourced a total of eight samples for testing our solution.

Once this is completed, it was time to set up the Python environment. For this project, our major tool was the python programming language which provides great libraries to deal with data analysis, and the building of machine learning models. Once the model was ready, then we were set to set up and run the program.

The required Libraries were then installed into the programme as seen in figure below.



```
App.py x Courses.py README.md
App.py > run
8
9 # 3. INSTALL BELOW LIBRARIES
10
11 # pip install spacy==2.3.5
12
13 # pip install -r requirements.txt
14
15 # pip install nltk
16
17 # pip install youtube-dl
18
19 # pip install https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-2.3.1/en_core_web_sm-2.3
20
21 # pip install pyresparser
22
23
24 # 4. CREAT A FOLDER AND NAME IT (e.g. resume)
25 #2.1 create two more folders inside this folder (Logo and Uploaded_Resumes)
26 #2.2 create two python files (App.py and Courses.py)
27
28 # 5. START YOUR SQL DATABASE
29
30
31 # 6. CONTINUE WITH THE FOLLOWING CODE...
32
```

Fig. 11: Python Development Environment



## 5.2. Application Tools

- i. Pandas - was installed for text manipulation
- ii. pdfminer3 - Supported the parsing of resumes to extract data. pdfminer3 is primarily used to extracting and analysing data from PDF documents. pdfminer3 obtains the exact location of texts in a page, as well as other information such as fonts or lines. It embodies an extensive PDF parser that can be used for purposes other than text analysis
- iii. Pyresparser - Is a simple resume parser used to extract information from resumes.
- iv. streamlit - used for deployment of our web application. It is an App framework commonly used in Machine Learning and Data Science.
- v. Pandas - It aided with our data manipulation. Pandas is a python-based library vast in data analysis and data manipulation. It is built on the Numerical Python (NumPy) which means that most of the features of NumPy are present.
- vi. pafy - Supports the recommendation of Youtube video courses. They are small single modules used for video streaming in Python. Pafy holds stream-specific data such as resolution, url and bitrates
- vii. plotly - Is a visualization software that enables the plotting of analysed data in the solution backend
- viii. MySQL - Acts like a connector between Python and MySQL.

- ix. streamlit-tags - Supported the creation of keywords in the solution. It also supports Ad tags.
- x. Pillow - supports the installation of various packages

## 6. Results

After an exciting process of developing an alternative first level screening algorithm for IT job roles, we will now present the results as follows;

Fig. R1. Home screen to the web App

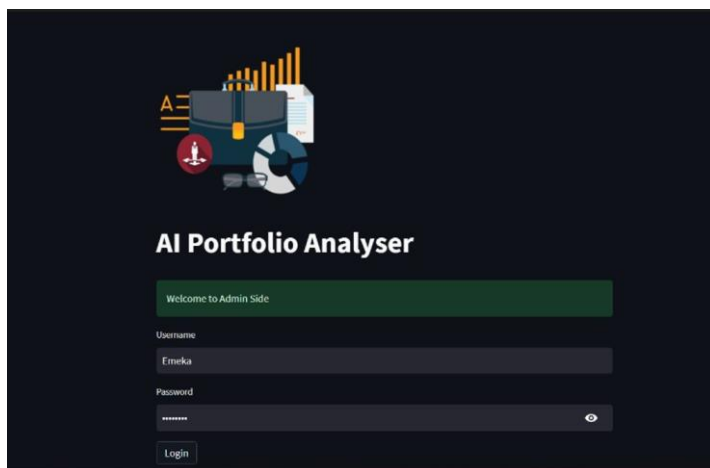


Fig R2. Candidate ready to Upload resume/profile/cv

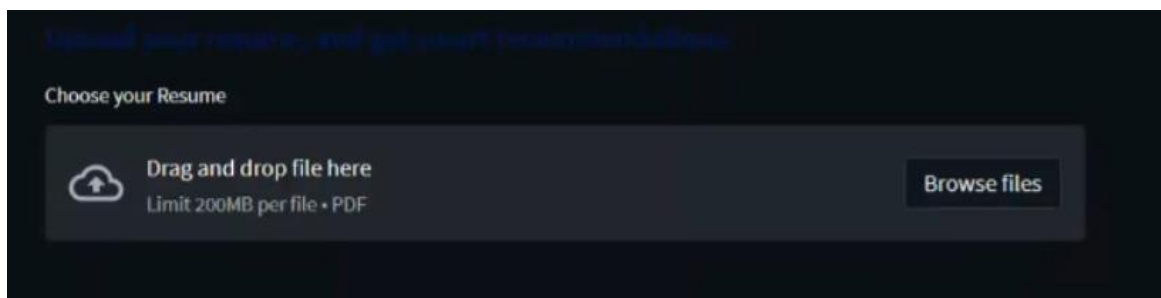


Fig. R3. Uploading from Local Drive

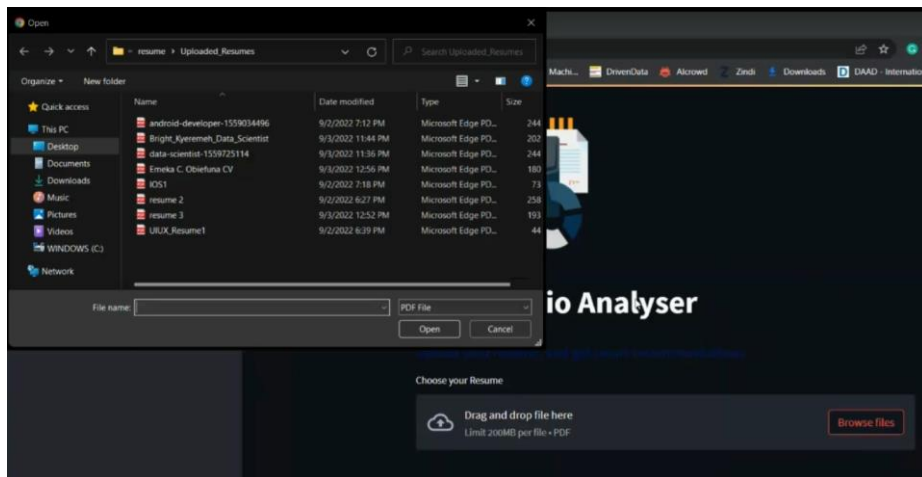


Fig. R4. Upload Complete

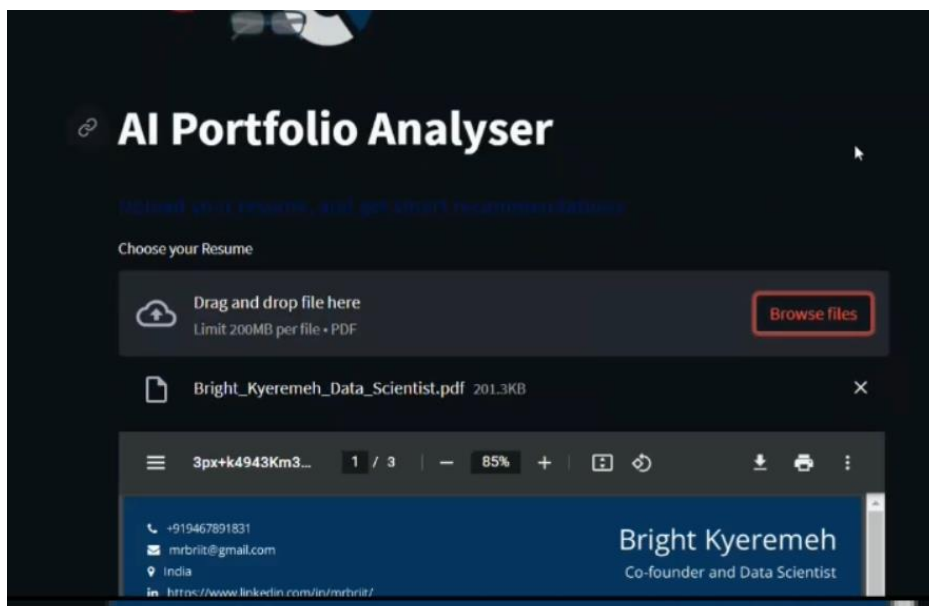


Fig. R5. Resume Analysed and showing skills

Contact: 9194678918

Resume pages: 3

**You are at experience level!**

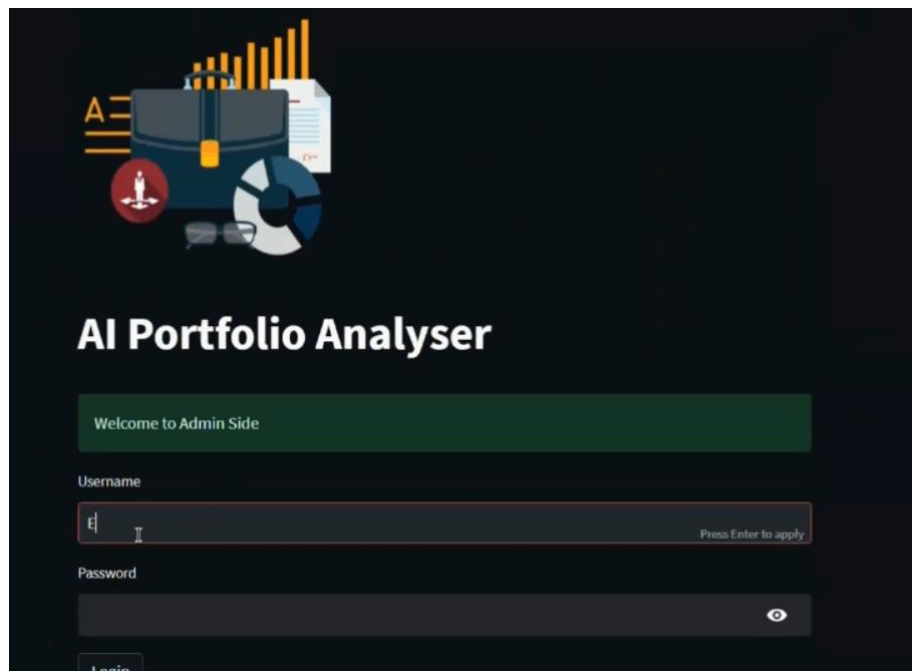
**Your Current Skills**

Pandas × Selenium × Windows × Research × Startup × Testing ×  
Ibm × Design × Digital marketing × Analyze × Analysis × Plan ×  
Video × Cloud × System × Transactions × Data analysis × Banking ×  
Os × Statistics × Algorithms × Strategy × Engineering × Teaching ×  
Numpy × Datasets × Technical × Writing × Machine learning × Sql ×  
Scrapy × Proposal × Social media × Python × Hypothesis × Tensorflow ×  
Tableau × Excel × Aws × Flask × English × Economics ×  
Mathematics × Health × Database × Email × Linux × Mysql ×  
International × Seaborn × Marketing × Analytics × Mining ×

See our skills recommendation below

**\*\* Our analysis says you are looking for Data Science Jobs.\*\***

Fig. R6. Admin Log In



The image shows the admin login page for the 'AI Portfolio Analyser'. At the top left, there is a graphic with a briefcase, a bar chart, a document, and a person icon. Below the graphic, the title 'AI Portfolio Analyser' is displayed in a large, bold, white font. Underneath the title, a green banner contains the text 'Welcome to Admin Side'. The login form consists of two input fields: 'Username' and 'Password'. The 'Username' field has a cursor and a 'Press Enter to apply' hint. The 'Password' field is obscured by a dark grey bar with a toggle icon. A 'Login' button is located at the bottom left of the form.

**AI Portfolio Analyser**

Welcome to Admin Side

Username

Press Enter to apply

Password

Login

Fig. R7. Admin Dashboard - list of candidates who have provided their details

Welcome Emekal

### User's Data

ID	Name	Email	Resu. Timestamp	Total
0	1 art director	hello@allisonbeer.com	20 2022-09-02_18:39:53	1
1	2 Android Developer	info@qwikresume.com	40 2022-09-02_19:13:02	2
2	3 OBIEFUNA NNAEMEKA	itsdonmonc@gmail.com	20 2022-09-02_19:13:43	2
3	4 OBIEFUNA NNAEMEKA	itsdonmonc@gmail.com	20 2022-09-02_19:17:11	2
4	5 0135New York	info@resumekraft.com+1-202-555-0135New	60 2022-09-02_19:18:19	2
5	6 Bright Kyeremeh	mrbrit@gmail.com	20 2022-09-02_19:19:22	3
6	7 HOWARD ONG	hello@reallygreatsite.com	40 2022-09-03_12:52:08	1
7	8 OBIEFUNA NNAEMEKA	itsdonmonc@gmail.com	20 2022-09-03_12:56:49	2
8	9 Data Scientist	info@qwikresume.com	40 2022-09-03_23:36:33	2
9	10 Bright Kyeremeh	mrbrit@gmail.com	20 2022-09-03_23:44:13	3

[Download Report](#)

Fig. R8. Expanded Admin Dashboard

ID	Name	Email	Resume Timestamp	Total	Predicted Field	User Level	Actual Skills
0	1 art director	hello@allisonbeer.com	20 2022-09-02_18:39:53	1	UI-UX Development	Fresher	['Design', 'Brand', 'Photoshop', 'Indesign', 'Prototyping', 'Strategy', 'Css', 'Illustrator', 'Wordpres
1	2 Android Developer	info@qwikresume.com	40 2022-09-02_19:13:02	2	Web Development	Intermediate	['Analysis', 'Python', 'Construction', 'Javascript', 'Android', 'Cloud', 'Email', 'Publishing', 'Specif
2	3 OBIEFUNA NNAEMEKA	itsdonmonc@gmail.com	20 2022-09-02_19:13:43	2	Data Science	Intermediate	['Analysis', 'Python', 'Technical', 'Research', 'English', 'Statistics', 'Writing', 'Machine learning', 'Y
3	4 OBIEFUNA NNAEMEKA	itsdonmonc@gmail.com	20 2022-09-02_19:17:11	2	Data Science	Intermediate	['Statistics', 'Technical', 'Reports', 'Training', 'Programming', 'Mysql', 'Communication', 'Sql', 'W
4	5 0135New York	info@resumekraft.com+1-202-555-0135New	60 2022-09-02_19:18:19	2	Android Development	Intermediate	['Xml', 'Communication', 'Ios', 'Architecture', 'International', 'Rest', 'Swift', 'Soap', 'Programmin
5	6 Bright Kyeremeh	mrbrit@gmail.com	20 2022-09-02_19:19:22	3	Data Science	Experienced	['Banking', 'Health', 'Teaching', 'Seaborn', 'Strategy', 'Technical', 'Analyze', 'Tensorflow', 'Mysql',
6	7 HOWARD ONG	hello@reallygreatsite.com	40 2022-09-03_12:52:08	1		Fresher	['Reports', 'Communication', 'Sas', 'Budget', 'Data analytics', 'Analysis', 'International', 'Accoun
7	8 OBIEFUNA NNAEMEKA	itsdonmonc@gmail.com	20 2022-09-03_12:56:49	2	Data Science	Intermediate	['Marketing', 'Reports', 'Database', 'Technical', 'Python', 'Programming', 'Presentation', 'Comm
8	9 Data Scientist	info@qwikresume.com	40 2022-09-03_23:36:33	2	Data Science	Intermediate	['Litigation', 'Design', 'Python', 'Programming', 'R', 'Etl', 'Aws', 'Data analysis', 'Sphinx', 'Analysis
9	10 Bright Kyeremeh	mrbrit@gmail.com	20 2022-09-03_23:44:13	3	Data Science	Experienced	['Tableau', 'Teaching', 'Os', 'Linux', 'Video', 'Design', 'Mysql', 'Testing', 'Cloud', 'System', 'Market
10	11 Bright Kyeremeh	mrbrit@gmail.com	20 2022-09-04_22:14:16	3	Data Science	Experienced	['Pandas', 'Selenium', 'Windows', 'Research', 'Startup', 'Testing', 'Ibm', 'Design', 'Digital market

This dashboard allows for human decision application based on ranked resumes.

Fig. R.9 Possible skill role predictions form a candidate

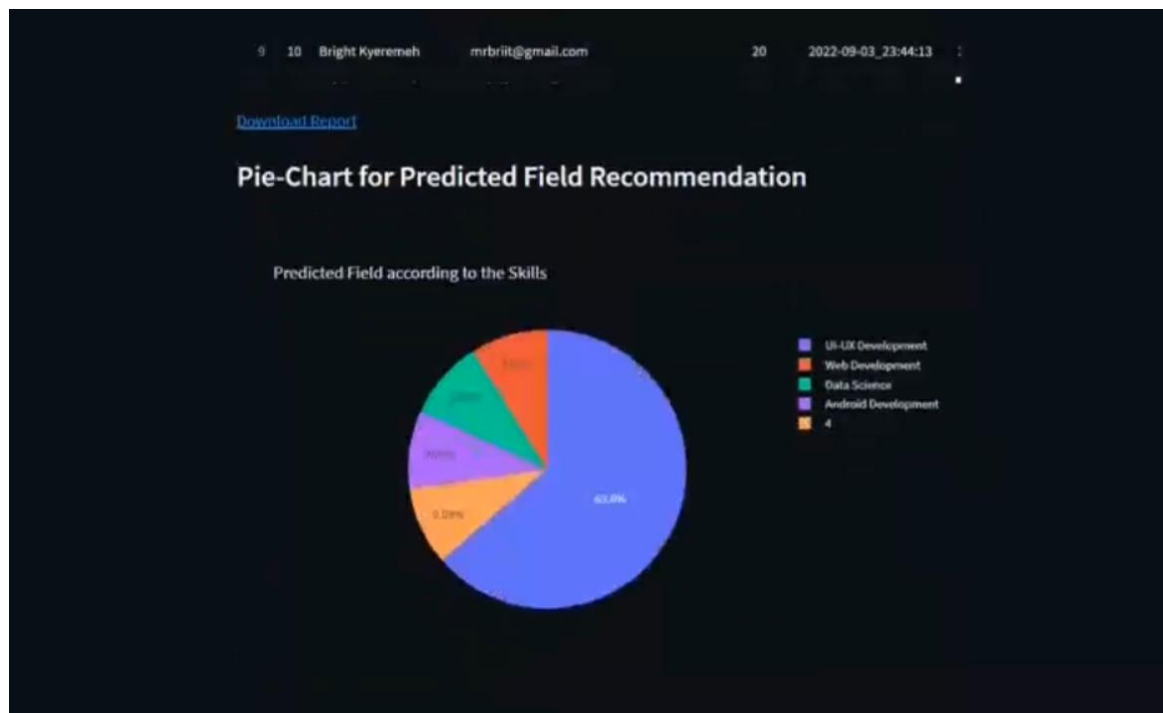


Fig. R10. Candidate Experience level based on resume analysis



## 7. Discussion

The first thing that comes to mind here is access, and equity. The solution we have successfully developed strictly focus on screening candidates based on key words that focus on their ability to execute. Solutions such as this will encourage hard work and deep learning without the constraint of formalised inspection. This simply means that if I can prove that I possess a particular skill by showing where I have consistently applied them, then I stand a chance of been reviewed for my personality and other yardsticks that the organisation might layout. This could be a very strong confidence booster for the applying individual because all he needs to do to get a foot in the doors is know his work.

The simplicity of the solution from a front-end perspective also allows for both recruiter and applicant have a pleasing ride with the application of technology. It also starts to build trust across certain aspects of the society, because people will no longer be seen for their colour, grade, or institution attended before they can at least have an opportunity to speak. The grading process and recommendation of courses to learn to improve classification percentile is a great tool for responding to candidates who have not been selected. Emails can be automated to relay such message, which will reduce the incidence of Ghosting - see (teamtaylor (2019). Derous et al in a paper presented at the 18th Conference of the European Association for Work and Organizational Psychology, May 18, 2017, Dublin, Ireland, agreed that there are strong questions about fairness at the resume funnelling stage (Derous and Ryan, 2018). Although the study is based on ethnic bias, it leans towards the same conversation - the ability of candidates to be turned back not because they are unqualified but because they were expected to express themselves in a certain way.

It is also a thing to note about text analysis, have a first-hand experience of how “stopwords” process is applied clearly align with our earlier explanation on how possibly the knowledge in how humans see words and how machines adapt them for use. There must then be international efforts to ensure that developers of such solutions understand the impact and reasoning of the users, Cowgill states, “These algorithms will can remove human biases exhibited in historical training data, but only if the human training decisions are sufficiently noisy; otherwise, the algorithms will codify or exacerbate existing biases” (Cowgill, 2020). Which aligns to the fact that machines susceptible to follow the level of eagerness by humans and will take them in exactly as received.

## 8. Conclusions

Conventions are generally hard to overturn - especially if the impacted are; naïve, inexperienced or content. It is also hard to execute if the principal actor (HR leaders) may have accepted the situation and focus on winning against each other in the talent war with very few of their available specifications dwindling due to economic and social trends. A summary of Rowan Gibson’s book on Innovation and creative thinking points to the fact that innovation is not a myth, that it is methodical, systemic and can be achieved. (Gibson, 2015).

The world is today faced with so many problems and technology is paving the way to resolving a lot of them, but humans must lead the way. Individuals must be judged on their area of expertise before they are rejected. The shortage of tech talent is real and big corporations have been looking to new horizons to grab talent - Amazon in Africa (Rest of the World, 2022), Microsoft’s new locations in Africa (Tech News Africa, 2022), Google, and many more. Countries like the United Kingdom and South Africa have set up programs to attract talent - these initiatives still carter for fewer people, a larger number could have had an opportunity if there could be one little shift in recruiting mindset.



As we have discovered in our study, most HR management solutions are built into a wider CRM, a few stand-alone solutions are focused on talent retention, and internal operations rather than supporting equitable hiring practices. Our review has also shown that benchmarking might have contributed largely to the lack of will to try new ways of first level screening. Understanding how text is analysed in machine learning have also shown that there is a knowledge gap in how these solutions work, hence the inability of practitioners to put it to good use or explore innovations to existing practices. Our sample data showed that academic qualification and soft skills were priority in the scrapped data - even for IT related jobs.

Executing the project has also provided the opportunity to learn about the loads of available tools and the ability of Python to aid software developers in any kind of thought they want to put to practice. Which means there is no shortage of tools to execute innovative solutions - especially if they will solve a pain point of lacking number of IT skills. At the end of the day, everybody wins. There are enough people to support expansion and innovation for big corporations and there are also enough skills to help start-ups grow. conclusion, having successfully

Although we have proven that technology can play a role, human interest and will to adapt is critical for the changes to begin to happen. As we have seen from Python to its adjoining libraries and data analytic methods reviewed, AI and ML are willing servants but can turn dangerous masters if not well instructed.

## 9. Reference list / Bibliography

Alan Guarino et al (2022) 2030: The Very Future of Work,  
<https://www.kornferry.com/insights/briefings-magazine/issue-30/2030-the-very-human-future-of-work> [Accessed on July 3, 2022]

Bengfort B., Bibro, R. and Ojeda, T. (2018) Applied Text Analysis with Python. Boston, O'Reilly

Berrar D., Dubitzky W., (2013) Overfitting, in: W. Dubitzky, O. Wolkenhauer, K.-H. Cho, H. Yokota (Eds.), Encyclopedia of Systems Biology, Springer, 2013, pp. 1617-1619.

Bhowmik, T., Niu, N., Singhanian, Prachi., and Wang, W. (2015) On the Role of Structural Holes in Requirements Identification: An Exploratory Study on Open-Source Software Development New York, Association for Computing Machinery, 10.1145/2795235

Caves, K. & Renold, U. (2018). Goal-Setting for TVET Reform: A Framework for Identifying the Ideal System in Nepal. Journal of Education and Research. 8. 10.3126/jer.v8i1.25477.

Chen, J., Cheng, C., Collins L., Chhabria P., and Cheong H. (2018) The Rise of Analytics in HR: The era of talent intelligence is here, LinkedIn Blog, <https://tinyurl.com/3wsdd2mz> [Accessed on August 24, 2022]

Chen, P. P-S. (1977), The Entity Relationship Model - Towards a Unified View of Data, Cambridge Massachusetts, Centre for Information Systems Research, Massachusetts Institute of Technology.

Chu S.K., Reynolds, R. B., Tavares, N. J., Notari, M., Lee, C. W. Y. (2017) 21st Century Skills Development Through Inquiry-Based Learning: From Theory to Practice, Singapore, Springer. DOI 10.1007/978-981-10-2481-8

Cowgill, B., (2020) Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening, Columbia University

Crown (2022) Official UK Government Website, <https://www.gov.uk/employment->



<https://web.stanford.edu/~gentzkow/research/text-as-data.pdf> [assessed: September 2, 2022]

Gibson, R. (2015) *The Four Lenses of Innovation: A power Tool for Creative Thinking*, New Jersey, Wiley

Gutierrez, D. (2020) *Why you should be Using Jupyter Notebooks*, <https://opendatascience.com/why-you-should-be-using-jupyter-notebooks/>, Open Data Science Society [Accesses on August 27th, 2022]

Han, J., and Kamber, M. (2006), *Data Mining: Concepts and Techniques Solution Manual*, Second Edition, University of Illinois at Urban-Champaign, Morgan Kaufmann.

Hastie T, Tibshirani R., Friedman J., (2008) *The Elements of Statistical Learning*, 2nd edition, New York /Berlin/Heidelberg, Springer,

Hines, P. (1998), "Benchmarking Toyota's supply chain: Japan vs UK", *Long Range Planning*, Vol. 31(6), pp. 911-918.

<https://www.legislation.gov.uk/ukpga/1996/18> [Accessed on July 5, 2022]

Indeed, Editorial Team (2016) *Is the War for Tech Talent Hurting Innovation? Hiring Managers, Recruiters Respond*, <https://www.indeed.com/lead/impact-of-tech-talent-shortage>, [Accessed on July 4, 2022]

Kerridge, C. (2019) *Employer ghosting: Why it's bad for candidates and how to prevent it*. Available: <https://blog.teamtailor.com/en/employer-ghosting-why-its-bad-for-candidates-and-how-to-prevent-it> [assessed: September 7, 2022]

Köchling, A., Wehner, M.C. *Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development*. *Bus Res* 13, 795-848 (2020). <https://doi.org/10.1007/s40685-020-00134-w>

- Kodali, G.A.R. (2008), "Benchmarking the benchmarking models",  
Benchmarking: An International Journal, Vol.15(3), pp. 257-291
- Kononow, P. (2017), ER Diagram vs Data Dictionary - Which is Better for Documenting Data Models, <https://tinyurl.com/KononowP>, Dataedo
- Kumar P. P. (2005) Effective Use of Gantt Chart for Managing Large Scale Projects, Cost Engineering; Vol. 47, Issue. 7: 14-21
- Lindebaum D., Vesa M., and Den Hond F. (2019). Insights from the machine stops to better understand rational assumptions in algorithmic decision-making and its implications for organizations. Academy of Management Review. <https://doi.org/10.5465/amr.2018.0181>. [Accessed on August 4, 2022]
- M. Storey, A. Zagalsky, F. F. Filho, L. Singer and D. M. German (2017) "How Social and Communication Channels Shape and Challenge a Participatory Culture in Software Development," in IEEE Transactions on Software Engineering, vol. 43, no. 2, pp. 185-204, doi: 10.1109/TSE.2016.2584053.
- MacLean, R., Stepney, S., Smith, S., Tordoff, N., Gradwell, D., Hoverd, T. and Katz, S. (2004), Analysing Systems: determining requirements for object-oriented development. Hemel Hempstead: Prentice Hall International
- Mann, G., and O'Neil, C. (2016). Hiring algorithms are not neutral. Harvard Business Review 9. <https://hbr.org/2016/12/hiring-algorithms-are-not-neutral>. [Accessed on August 4, 2022]
- Mohamed Branine, (2008),"Graduate recruitment and selection in the UK: A study of the recent changes in methods and expectations",  
<https://tinyurl.com/mahduk>, Career Development International, Vol. 13 Iss: 6 pp. 497 - 513
- Mohammed, F. (2017) What is data? Scridb.com Available:  
<https://www.scribd.com/document/336246896/What-is-Data#> [accessed:

Pedregosa et al (2011), Scikit-learn: Machine Learning in Python, MLR 12, pp. 2825-2830 Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html> [accessed: September 2, 2022]

Pugh, K. Lean-Agile Acceptance Test-Driven Development: Better Software Through Collaboration,

Ravenscroft, A., Lindstaedt, S., Kloos, C.D., Hernandez-Leo, D. (2012) 21st Century Learning for 21st Century Skills, Heidelberg, Dordrecht, London, New York, Springer. LNCS 7563

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5, 532-538.

Rest of The World (2022), Amazon's Africa expansion means a hiring drive in Nigeria Available: <https://restofworld.org/2022/amazon-is-set-to-expand-its-africa-footprint-with-a-hiring-drive-in-nigeria/> [assessed: September 5, 2022]

Singer, L., Brill, O., Meyer, S. and Schneider, K. (2009), "Utilizing Rule Deviations in IT Ecosystems for Implicit Requirements Elicitation", Second International Workshop on Managing Requirements Knowledge (MaRK'09).

Spahic, J., (2015) Exploring HR Intelligence Practices in Fortune 1000 and Select Global Firms, Drexel University.

Tech News Africa (2022) Microsoft Opens 2 New Offices Africa Available: <https://techafricanews.com/2022/03/31/microsoft-opens-2-new-offices-for-adc-in-africa/>

The European Business Review, (2019), Custom-Built Programmes for The Future Workforce, Issue: September/October  
<https://www.scribd.com/article/450374751/Custom-Built-Programmes-For-The-Future-Workforce> [Accessed on August 28th, 2022]

Vijayan, J. and Raju, G. (2011), "A New approach to Requirements Elicitation Using Paper Prototype", International Journal of Advanced Science and Technology, Vol. 28, pp. 9-16.

Walsh, M. J. (2021), HR Analytics Essentials You Always Wanted To Know, <https://www.scribd.com/>, Vibrant Publishers, ISBN: 9781636510354

Watson, G.H. (2007), Strategic Benchmarking Reloaded with Six Sigma: Improving Your Company's Performance Using Global Best Practice. Hoboken, New Jersey: Wiley

Appendix  
Refer to attached Ethics approval