

ARBA MINCH INISTITUTE OF TECHNOLOGY (AMIT)

ARBA MINCH UNIVERSITY

“**Web Based CV analysis and Illegibility Prediction System**”

**Group Members:**

**Name**  **ID No**

Dagim Yosef Ramit/532/11

Genet Gezahegn Ramit/dwdd/11

Lamrot Assefa Ramit/ddwd/11

Yeabsira Abera Ramit/dwd/11

Zekarias Kobota Ramit/2035/11

Submitted to Department of Software Engineering, Faculty of Computing and Software Engineering, AMIT, Arba Minch University, in partial fulfillment for the requirement of the Degree of Bachelor Science in (Software Engineering)

Arba Minch, Ethiopia

**Feb 30, 2023**

# Declaration

This is to declare that the proposal is done with 5th year Software Engineering Student under the title of “Web based CV analysis and Illegibility prediction system”

Project advisor: Signature Date

Project Examiners Signature Date

1.

2.

3.

# Acknowledgment

First of All, we would like to thank almighty GOD for giving us the strength to complete this proposal. Secondly, we would like to express our deepest sincere gratitude to our Advisor Ms. Amin Tuni(Ass. Professor) for his guidance, support and continues encouragement for making this proposal successful. Thirdly we would like to thank all software engineering staff members. Finally we would like to thank all our classmates and software engineering students for their strong supports.

Table of Contents

[Declaration i](#_Toc128406605)

[Acknowledgment ii](#_Toc128406606)

[List of Tables vi](#_Toc128406607)

[List of Figures vii](#_Toc128406608)

[Acronyms/Abbreviation viii](#_Toc128406609)

[Abstract ix](#_Toc128406610)

[Chapter One 10](#_Toc128406611)

[Introduction 10](#_Toc128406612)

**[1.1 Background of the Study](#_Toc128406613)** [10](#_Toc128406613)

**[1.2 Team Composition](#_Toc128406614)** [11](#_Toc128406614)

**[1.3 Statement of the Problem](#_Toc128406615)** [12](#_Toc128406615)

**[1.4 The Proposed System](#_Toc128406616)** [12](#_Toc128406616)

**[1.5 Research Questions](#_Toc128406617)** [13](#_Toc128406617)

**[1.6 Objectives of the Project](#_Toc128406618)** [14](#_Toc128406618)

***[1.6.1 General Objective](#_Toc128406619)*** [14](#_Toc128406619)

***[1.6.2 Specific Objective](#_Toc128406620)*** [14](#_Toc128406620)

**[1.7 Significance and Beneficiary of the System](#_Toc128406621)** [15](#_Toc128406621)

***[1.7.1 Significance of the System](#_Toc128406622)*** [15](#_Toc128406622)

***[1.7.2 Beneficiary](#_Toc128406623)*** [15](#_Toc128406623)

**[1.8 Scope and Limitation](#_Toc128406624)** [16](#_Toc128406624)

***[1.8.1 Scope](#_Toc128406625)*** [16](#_Toc128406625)

***[1.8.2 Limitation](#_Toc128406626)*** [17](#_Toc128406626)

**[1.9 Organization of the Document](#_Toc128406627)** [17](#_Toc128406627)

**[1.10 Time Schedule](#_Toc128406628)** [18](#_Toc128406628)

[Chapter Two 19](#_Toc128406629)

[Literature Review 19](#_Toc128406630)

**[2.1 CV analysis](#_Toc128406631)** [19](#_Toc128406631)

***[2.1.1 CV and CV analysis](#_Toc128406632)*** [19](#_Toc128406632)

[2.2 Shortlisting 19](#_Toc128406633)

[2.2.1 Technical and legal considerations that go into a proper shortlisting process: 20](#_Toc128406634)

**[Determine Shortlist criteria](#_Toc128406635)** [20](#_Toc128406635)

***[2.2.2 CV Screening Cost](#_Toc128406636)*** [21](#_Toc128406636)

***[2.2.3 Technical Standards and Quality of CV analysis](#_Toc128406637)*** [21](#_Toc128406637)

***[2.2.4 Summarizing](#_Toc128406638)*** [21](#_Toc128406638)

**[2.3 Related Projects That Has Been Developed](#_Toc128406639)** [22](#_Toc128406639)

***[2.2.1 Comparison](#_Toc128406640)*** [23](#_Toc128406640)

**[2.2.2 Strong Side of the reviewed project](#_Toc128406641)** [23](#_Toc128406641)

**[2.2.3 Limitations](#_Toc128406642)** [23](#_Toc128406642)

**[2.2.4. Solution](#_Toc128406643)** [24](#_Toc128406643)

**[2.4 Conclusion](#_Toc128406644)** [24](#_Toc128406644)

[Chapter Three 25](#_Toc128406645)

[Methodology 25](#_Toc128406646)

**[3.1 Introduction](#_Toc128406647)** [25](#_Toc128406647)

**[3.2 System Development Methodology](#_Toc128406648)** [25](#_Toc128406648)

***[3.2.1 Data Collection Methodology](#_Toc128406649)*** [26](#_Toc128406649)

***[3.2.2 System Development Approach](#_Toc128406650)*** [27](#_Toc128406650)

***[3.2.3 System Development Tool](#_Toc128406651)*** [28](#_Toc128406651)

**[3.3 Methods for Illegibility prediction Feature of Our System](#_Toc128406652)** [30](#_Toc128406652)

***[3.3.1 Overview](#_Toc128406653)*** [30](#_Toc128406653)

***[3.3.2 Machine Learning](#_Toc128406654)*** [30](#_Toc128406654)

***[3.3.2.1 Supervised Learning](#_Toc128406655)*** [31](#_Toc128406655)

***[3.3.3 Data Collection](#_Toc128406656)*** [32](#_Toc128406656)

***[3.3.4 Preprocessing](#_Toc128406657)*** [33](#_Toc128406657)

*[3..3.5 Data Analysis Method](#_Toc128406658)* [36](#_Toc128406658)

***[3.3.6 Modeling Techniques](#_Toc128406659)*** [37](#_Toc128406659)

***[3.3.7 Programming Technique and Tools](#_Toc128406660)*** [37](#_Toc128406660)

[Conclusion 37](#_Toc128406661)

*[References](#_Toc128406662)* [37](#_Toc128406662)

# List of Tables

[Table1.1 Team Composition 11](#_Toc128406663)

[Table2.1 Comparison of Existing Projects 23](#_Toc128406664)

[Table3.1 List of Features we used 33](#_Toc128406665)

# List of Figures

Figure 1.1 Time Schedule 18

# Acronyms/Abbreviation

# Abstract

This project aims to study and understand the current system problems and provide an automated system to enable the CV analysis and ranking to cope with technological advancements using machine learning algorithms. It will enable a more effective way to shortlist submitted candidate CV from a large number of applicants providing a consistent and fair CV ranking policy, which can be legally justified and will predict illegible person for the position. The system will rank the CV based on the experience and other key skills which are required for particular job profile, helping the HR (human resource) department to easily shortlist the candidate based on the CV ranking policy.

**Key Words**: Machine Learning, Automated, Shortlist, rank

# Chapter One

# Introduction

Since they have so many advantages and a vast range of uses, computers are now a part of every aspect of life. The human race has become completely reliant on computers in several areas as well. One of the remarkable and much known products of technology advancement is the conversion of manually-operated system into automated system. Automation produces a great impact in the lives of mankind, particularly in the field of industry, business, medicine, technology and education.

Analyzing the Curriculum vitae of the candidates applying for particular job position is becoming difficult task in hiring process of an organization. To avoid these and other related problems we are proposing to develop Web based CV analysis and Illegibility prediction system which enables the applicants to register and apply online for the particular job posted by HR by uploading their CV and after that the system analyzes their CV and shortlist the candidate(s) best suited for the particular job position. We are initiated to develop this project because the Arba Minch University Human Resource department is having difficult condition in analyzing CV of each candidates applying for the particular job that takes long time and more human power.

The main purpose of this project is to develop and implement a computerized CV analysis system to help HR to provide reliable services. These systems tend to us to find possible solution about the problem, it helps applicants to access the system easily and also provide illegibility prediction model developed by using machine learning.

## **Background of the Study**

Arba Minch University is based in South-West Ethiopia. The University was initially founded as the Arba Minch Water Technology Institute (AWTI) and it was officially inaugurated as a full -fledged university in June 2004 and started offering both undergraduate and graduate programs in the following institutes, colleges and schools:

* [Arba Minch Institute of water Technology](https://www.amu.edu.et/academics/main-menu-academics-institutes/main-menu-academics-institutes-awti" \t "/home/Zeka/Documentsx/_blank)
* [Arba Minch Institute of Technology](https://www.amu.edu.et/amit-home" \t "/home/Zeka/Documentsx/_blank)
* [College of Agriculture](https://www.amu.edu.et/college-of-agriculture-home" \t "/home/Zeka/Documentsx/_blank)
* [College of Business and Economics](https://www.amu.edu.et/college-of-business-and-economics-home" \t "/home/Zeka/Documentsx/_blank)
* [College of Natural Sciences](https://www.amu.edu.et/cncs-home" \t "/home/Zeka/Documentsx/_blank)
* [College of Medicine and Health Sciences](https://www.amu.edu.et/cmhs" \t "/home/Zeka/Documentsx/_blank)
* [College of Social Sciences and Humanities](https://www.amu.edu.et/cssh-home" \t "/home/Zeka/Documentsx/_blank)
* [School of behavioral and pedagogical sciences](https://www.amu.edu.et/academics/main-menu-schools/main-menu-schools-school-of-pedagogical-and-behavioural-sciences" \t "/home/Zeka/Documentsx/_blank)
* [School of law](https://www.amu.edu.et/academics/main-menu-schools/main-menu-schools-school-of-law" \t "/home/Zeka/Documentsx/_blank)
* [Sawla Campus (a multi-disciplinary Academic areas campus)](https://www.amu.edu.et/academics/main-menu-colleges/sawla-campus" \t "/home/Zeka/Documentsx/_blank)

Arba Minch University is currently working as one of National Research Universities and has more than 36,000 registered students in regular, evening, weekend and distance programs in first, second and third degree.

There are a lots of management area in Arba Minch University from the main president office to the lowest areas and Human Resource(HR) management is one of the main management area in Arba Minch University that is focused on the recruiting and retention of employees within a company. HR typically finds hires and trains employees. It oversees employee relations, manages benefit programs. It's the place an employee goes with questions about their position at the company, to address concerns, and to air grievances. So our project will focus on the work of Human Resource department of Arba Minch University main campus, specifically Faculty of Computing and software engineering.

## **1.2 Team Composition**

Table1.1 Team Composition

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Project Title | **Web based CV analysis and Illegibility prediction system for AMU FCSE** | | | | |
| Prepared by | S.No | Name | ID.No | Email/Mobile | Responsibility |
| 1 | Dagim Yosef | Ramit/532/11 | 0938967148 | Participate in all works |
| 2 | Genet Gezahegn | Ramit//11 |  | Participate in all works |
| 3 | Lamrot Assefa | Ramit//11 |  | Participate in all works |
| 4 | Yeabsira Abera | Ramit//11 |  | Participate in all works |
| 5 | Zekarias Kobota | Ramit/2035/11 | kobezekibelay@gmail.com | Participate in all works |
| **Date** | Feb17, 2023 | | | | |
| Advisor | **Mr. Amin Tuni (Ass. Professor)** | | | | |

## **1.3 Statement of the Problem**

In existing manual system when a human recruiter is performing CV analysis, they are actually doing two things at the same time. First, they are processing the CV to understand what it saying (this is known as parsing). Second, they will have specific vacancy or set of vacancies in mind and they will as be assessing whether the CV they are looking at could be a good match (whether this is a candidate that should be offered to client for consideration)

The lack of a reliable and trustworthy personality prediction system through CV analysis is the other problem as it limits the potential for companies to properly identify potential employees to match their organizational values and goals. This results in increased recruitment costs, high employee turnover rates and decreased job satisfaction.

**Generally, the main and observed problems are: -**

* **Complexity: -** analyzing CV of all candidates is difficult and time taking task.
* **Man power wastage**: - HR department use more human power to analyze all the CV of the candidates who applied for the particular jo position.
* **Wastage of time**: -HR lose their time by analyzing and probing each and every CV one by one.
* **Wastage of money**: -Candidates waste their money by printing their CV to apply manually for the particular job.
* **Redundancy**: - There is redundancy in ranking list of applied CV’s by corruption from someone who is one of the applicants.

## **1.4 The Proposed System**

CV analysis involves a multi-step process of resume reviews. The intent is to extract key information from resumes about job applicants in the most efficient manner possible. So, the proposed system is aimed to perform basic and crucial tasks of the CV analysis process. It contains a well-organized CV analysis system designed to provide fast and easy way of analyzing the CVs of the candidates applying for the particular job position. Such as an easy way for candidates to check available jobs for applying. And also, the system allows managing all personal informations of the candidates. This includes posting, deleting and updating job posts. The system also provides an easy way for HR and to see the list of applied candidates list and allow them to compute for the particular job position. Generally, Our System concerned on managing candidates, providing easy and understandable way of analyzing the candidates CV, and making easy and fast communication held between candidates and HR. And also provides an illegibility prediction model for predicting the illegibility of the candidates for the particular job position.

## **1.5 Research Questions**

The research questions are based on the purpose of the study and theoretical rationale. The purpose of our project is to provide a working environment for applicants CV analysis and illegibility prediction system in Arba Minch University FCSE by studying the existing working environment.

**Questions**

To conduct this Project, the following question was raised:

* To what extent will the analysis of the existing CV analysis system determine the need of creating the computerized one?
* To what extent the computerized CV analysis system reaches the HR and candidates in order to provide a proper platform that will help them to manage and access system, respectively?
* How will developing a CV analysis system helps a applicants and HR in managing and selecting candidates required for the job position?
* How to study and analyze the functional requirement specification of the system?
* How to validate the newly developed system?
* Who are the applicants, target groups?
* Where the required data set can be obtained?
* How to make the obtained data sets suitable for the algorithm?
* How obtained data set can be analyzed?
* How to choose best result of algorithm?

## **1.6 Objectives of the Project**

### ***1.6.1 General Objectiv******e***

The general objective of this project is to develop a web-based Personality CV analysis and Prediction for AMU FCSE by using machine learning algorithms.

### ***1.6.2 Specific Objective***

* Collect information about the existing/current CV analysis system in the University.
* Identify the problems in the existing system.
* Find out solution for the problems in the existing system.
* Analyze the requirements to propose solution for the existing system.
* Designing the analyzed requirement in order to implement the newly proposed solution.
* Implement a fully automated CV analysis system for the AMU FCSE.
* Provide a complete organized and reliable system that with least possibilities of any error.
* Collect data sets from online sources.
* Extract the data set in CSV format and clean the dataset to make dataset suitable for our algorithm.
* Develop a model illegibility using Regression machine learning algorithm technique.
* Integrate the model with the website
* To minimize the manpower workload for recruitment process.
* To reduce the time for analyzing and selecting individual Cv.
* To maximize effectiveness and efficiency recruitment process.
* To develop secured and reliable system.
* To minimize cost by reducing employee turnover rates and to increase job satisfaction.

## **1.7 Significance and Beneficiary of the System**

### ***1.7.1 Significance of the System***

* Candidates apply for particular job by using their phone or computer
* CV will be analyzed on time and shortlisted.
* Misinformation provided by brokers will be avoided because there is no need of intermediates in our system.
* More secured than the existing system.
* Save time of applicants and HR, applicants will not lose their time in applying for the job manually by involving in the queue of many peoples and HR will not lose their time in registering and analyzing each and every CV of the candidates that applied for the particular job position.
* Save money of applicants as well as HR departments.
* Allow users to apply for the job online even they are not in Arba Minch or the location where the job position placed; they can apply before coming and they can directly go to the home after they come.
* Reduce need of man power.
* Predict the illegibility of the candidates so it helps to reduce excess time required to shortlist the candidates and unfair actions usually done by the recruiters.

### ***1.7.2 Beneficiary***

The target beneficiary groups are:

* **Applicants: -** our system prevents applicants from extra expense for printing CV and wasting time for applying their CV manually and they can get the service anywhere anytime.
* **HR department: -** the HR are beneficiary regarding to the analysis of CV on time and also, they can save their time, money, and energy at the same time.
* **Developers: -** we are also beneficiary from the system because it helps us to gain knowledge on how to develop a system and get some experience of system development.
* **Other peoples: -** other peoples may also benefited by referring the document of our project to retrieve information related to the CV analysis and shortlisting method.

## **1.8 Scope and Limitation**

### ***1.8.1 Scope***

The scope of the project can be described as the overall features of what the new system is capable of doing. The scope this project is to develop web-based personality analysis and prediction system for AMU FSCE which is able to accurately predict a candidates illegibility and capabilities from their resume by extracting data from the person’s resume.

This system accepts CV from user as pdf format and then analyses the uploaded CV by using data mining classification algorithm to predict the personality of the user.

**Generally, our system has the following scopes:**

* The system allows applicants(candidates) to register and login to the system
* It also allows the applicants to view the job posted by the Human resource department after logging in to the system.
* The system allows Human Resource (HR) department to add new jobs with details to the system that will be displayed in logged in candidates page.
* The system allows the candidates to apply and upload their CV for the posted jobs after logging in to the system.
* It allows HR to view the list of shortlisted candidates.
* It also HR to view the profiles of the registered candidates.

Our project is also developed to solve the problem of manual CV analysis by making the analysis in the automated way, shortlisting candidates and predicting illegibility based on their CV. The project examines the prediction model by using machine learning algorithms techniques to improve the prediction of illegibility of candidates only for Arba Minch University FCSE.

### ***1.8.2 Limitation***

The main limitation of our project is that the data used or collected and the sample used for the illegibility prediction is taken from online sources hence we did not find suitable data about CV analysis to identify required skills, experience and other related information about the candidate in the University, this leads our model’s illegibility prediction may not accurately shortlist candidates required for the particular job position. The other limitation of our system is, it does not support different languages like Amharic, Gamogna. Hence occasional users can’t use the system easily.

## **1.9 Organization of the Document**

This document is organized into six chapters as follows:

**Chapter One:** This chapter includes background of the study, introduction, motivation, statement of the problem, objectives, research methodology, scope and limitation of the study, and application of the study.

**Chapter Two**: Basically, contains literature review and related works of the study. Gives review of literature on CV analysis and illegibility prediction system.

**Chapter three:** Presents the methodology of the work in detail. What design is better for Illegibility prediction system, why the specific design is selected and basically the methodologies and which machine learning technique are better for our study.

**Chapter four:** under this we see experiment and Implementation, source of dataset, preparation and data-set description including sample code.

**Chapter five:** Include result and modeling. In this chapter we analyze all regression method and compare them with their mean absolute error (MAE) and select the best one from listed regression method.

**Chapter Six:** contains conclusion and recommendation (future work) of the study

## **1.10 Time Schedule**

Figure 1.1 Time Schedule

# Chapter Two

# Literature Review

This chapter contains a study of the relevant literature for the system that will be developed. An analysis of the major components of current knowledge, such as important discoveries and theoretical and methodological contributions to a particular field, is the goal of a literature review. Positioning the current study within the body of literature will provide context for the reader in particular (Cooper, 1998).

It can act as a guide for developing a new system that is better or more beneficial than the existing one. Analyzing pertinent initiatives that are currently under development is crucial for improving the system.

## **2.1 CV analysis**

### ***2.1.1 CV and CV analysis***

To understand what the CV analysis is, it is necessary first to understand what CV(Curriculum Vitae) is. CV is a document that job applicants use to showcase their academic and professional accomplishments. It is use to apply for position within areas where a person’s specific knowledge or expertise is required. Curriculum vitae are usually long from resume and it must include the information that the recruiter needs to verify the skills, experience, and educational qualifications of an applicant. In many countries a CV is usually the first document that a perspective employer looks at when screening candidate for job interviews, scholarship programs, grant applications, or bursaries. When we come to CV analysis, it is concerned with viewing the contents(Skills, experience, languages and etc.) in the CV of particular candidates to shortlist candidates that have the skills and experience or ability the companies requires for the particular job position.

# 2.2 Shortlisting

The holy grail of recruitment is finding a quick, easy, and accurate way to automate candidate shortlisting. As organizations start to embrace the idea that recruiting and retaining talented employees represents a competitive advantage, Deloitte’s Human Capital Trends Report finds that HR departments are being tasked to redesign everything they do to change the way organizations hire, manage, and support their people. But the more things change, the more they stay the same. Despite these demands for change, the classic challenges of recruitment remain including how to find, attract, screen, and shortlist candidates. Shortlisting is often the most challenging and time-consuming step in the recruitment process.

**Shortlisting** is the process of identifying the candidates from your applicant pool who best meet the required and desired criteria for the open req and who you want to move forward onto the next step of your recruitment process, which is usually some form of interview.

In the recruitment process, shortlisting comes after sourcing (finding and attracting) and before interviewing and further assessing. Screening and shortlisting often happens simultaneously: as we screen resumes from candidates, we shortlist the best ones to move forward.

# 2.2.1 Technical and legal considerations that go into a proper shortlisting process:

### **Determine Shortlist criteria**

These are the essential and desirable criteria needed to do the job and the minimum level that the shortlisted candidate should have. These criteria should be related to on-the-job performance and ideally should be captured in the job description.

Developing the right criteria is a balance between standards that are high enough to ensure good quality candidates move forward, but not too strict that you’d be unnecessarily screening out a lot of qualified candidates.

**Create a shortlist scorecard**

Take the essential and desirable criteria you’ve identified above and create a shortlist scorecard for your candidates. The purpose of this scorecard is to list out each criteria so that you can assign a rating for each screened in candidate.

**Determine the length of shortlist**

When it comes to deciding the length of your shortlist, most people have a predetermined number in mind. For example, interviewing four to six candidates to get one successful hire.

For high volume recruitment roles like customer service representatives where you need to hire hundreds of people, you might shortlist any candidate that meets your criteria.

**Screening resumes to shortlist candidates**

According to industry stats, 75% of applicants are unqualified and 88% are not strong enough to move forward to an interview. When 75 to 88 percent of the resumes you receive for an open requirement are ones you have to screen out, it’s obvious why shortlisting is the most time-consuming part of recruitment.

### ***2.2.2 CV Screening Cost***

CV screening cost or analysis cost is a key in the improvement of recruitment processes of a company or organization. In past this reflects variations in the quality of hiring, but there are also significant cost variations that do not reflect quality changes or differences. Expenditure on Screening is an important item in recruitment budgets. Screening costs in recruitment vary greatly from one company or organization to another and from one place to other.

### ***2.2.3 Technical Standards and Quality of CV analysis***

The quality of selecting qualified candidates depends mainly on the following factors: the technical conditions, the required experience for the position, skills, and required education level. In many countries especially in the developing countries large numbers of educated persons live without job more specially persons with high level of skills because of poor and unfair CV analysis.

### ***2.2.4 Summarizing***

As we realize from the above points CV analysis and screening is a very complex process that is impossible to handle all the things about renting a house in a single website because CV screening has different forms that varies considerably. So, in our proposed system we try to minimize the concept by selecting essential parts of the CV screening and analysis activities in order to avoid complexity in the system. i.e., our project considers the applicants by which the CV is analyzed by the system and Human resource department is takes action on shortlisting. The HR post their job and enters the details of their job requirements. These details help to standardize the job position that can be posted by the HR. And our system provides services for all the applicants that want to apply for the jobs that are posted by the admin (HR). And our system also provides a shortlisting of the CV based on the requirement of particular job position that is posted by the HR.

## **2.3 Related Projects That Has Been Developed**

In this context we refer two projects developed in India and our team tries to identify the problems in the two projects and tries to provide solution for those problems.

1. ***Online Resume Mart***

**Authors:** Prithviraj Banerjee

Computer Science Engineering, University of Bridgeport, India

As we observed the objectives of this project is to design and develop a web service which is a place for Job Seekers and Job Providers to meet. The database is used to collect all the details of the job seeker and provider. Online resume collects multiple resumes from the Job seeker and the job provider is provided with the data required as per need of the company.

***2. A Resume Analyser Application for Matching Candidates with Job Requirements, Using Parsing Algorithm***

**Authors:** Abdiaziz Abdi Ali, Farhan Hersi Olow, Abdinasir Abdullahi Dahir**,**Abdifitah Mohamed Hassan, Mohamed Ahmed Mohamud.

**Jamhuriya University of Science and Technology Mogadishu-Somalia,** Faculty of Computer and Information Technology

The main purpose of this project is to define the main attributes required for the matching resume data and job offer in the job matching system. This attributes are extracted from job seekers resumes by using resume parser. When the attributes are defined, the proposed framework is guided to follow word similarity algorithms for the PostgreSQL trigram module to match the job seekers' skills against the requirements of the job posted by the prospect employers.

### ***2.2.1 Comparison***

**The table below compares the above two projects**

Table2.1 Comparison of Existing Projects

|  |  |  |  |
| --- | --- | --- | --- |
| **Author** | **Title** | **Method** | **Description** |
| Prithviraj Banerjee | Online Resume Mart | Web based | The system that provide web service which is a place for Job Seekers and Job Providers to meet. |
| Abdiaziz Abdi Ali, Farhan Hersi Olow, Abdinasir Abdullahi Dahir,Abdifitah Mohamed Hassan, Mohamed Ahmed Mohamud | A Resume Analyser Application for Matching Candidates with Job Requirements, Using A Parsing Algorithm | Web Based with AI | A web-based application to help user to register and apply for particular job profile , after the user upload their CV this system analysis the uploaded CV to shortlist the candidate that matches for particular job position. |

## **2.2.2 Strong Side of the reviewed project**

As we reviewed the above two projects, the second project(A resume analyzer application for matching candidates with job requirements, using a parsing algorithm) has the same idea with our proposed system that is related with analyzing candidates CV by using machine learning algorithms and finally predicts illegible candidate(s) that matches particular job profile.

### **2.2.3 Limitations**

**As we observe there is some limitations in above two projects. those limitations are**:

We reviewed the above described projects and the limitation of the second project is that it only focus on how to make the meeting between job seekers and job providers that means it does not provide a way to shortlist the job seekers candidates based on their CV and also does not provide illegibility prediction.

### **2.2.4. Solution**

Our system tries to solve the limitation of the above projects by predicting illegibility of candidates for the particular job profile by using machine learning algorithms and by providing the way that HR post particular job detail as additional functionality in the second project above.

## **2.4 Conclusion**

This chapter discusses the collected literature review. This helps us to gain understanding about the CV analysis in recruitment process of an organization, helps us to know what things included in the system. And also helps us to discover the problems of the previous projects which need to be improved and overcome in our project.

# Chapter Three

# Methodology

## **3.1 Introduction**

Methodology is the process, steps or stages used to collect information and data for the purpose of making decisions. The methodology is chosen from the software development life cycle (SDLC) model. Methodology also specifies the methods and technologies used to develop the system such as, the method used to gather the data, approaches used to design the system, and software and hardware requirements used to implement the system.

## **3.2 System Development Methodology**

A system development methodology provides guidelines for every facet of the system development life cycle. Some methodologies contain massive written documentation that defines everything that the developer may need to develop at any point in the project. For developing our project we will follow agile model methodologies.

**Agile Model**

Agile SDLC model is a combination of iterative and incremental process models with focus on process adaptability and customer satisfaction by rapid delivery of working software product. Agile Methods break the product into small incremental builds. These builds are provided in iterations. Each iteration typically lasts from about one to three weeks.

**We choose agile model because:**

* Agile uses an **adaptive approach** where there is no detailed planning and there is clarity on future tasks only in respect of what features need to be developed. There is feature driven development and the team adapts to the changing product requirements dynamically. The product is tested very frequently, through the release iterations, minimizing the risk of any major failures in future.
* **Customer Interaction** is the backbone of this Agile methodology, and open communication with minimum documentation are the typical features of Agile development environment. The agile teams work in close collaboration with each other and are most often located in the same geographical location.

### ***3.2.1 Data Collection Methodology***

Our team uses primary and secondary data collection methods to collect information about the current CV analysis process in Arba Minch University HR department. We use two data collection methods to collect the data need for the team project those are:

* Primary data collection method and
* Secondary data collection methods

#### 3.2.1.1 Primary Data Collection

When the data collected by the researcher for the first time it is called Primary Data.it is original in nature and specific to the research problem under study.

Our team uses the following data collection techniques to collect primary data:

* **Observation:** it is a fact-finding technique where the system analyst either participate in or watches a person perform activities to learn about the system, our team observes the current CV analysis process in HR department of AMU and we gather information such as, how the analysis performed, what is the problem in the current system and many other information. Our team also referred other senior projects. And we observe the entire brokers working procedures.
* **Interview:** It is another fact-finding technique where by the system analyst collects information from individual face to face interaction. We collect data by interviewing Staffs in Arba Minch University HR department and also some customers about the candidates CV analysis processes.

#### 3.2.1.2 Secondary Data Collection

When the data collected by someone else for purpose other than the researchers current project and already has undergone the statistical analysis is called as Secondary Data.

* **Document Analysis:** our team tried to discover written document related to our project.

### ***3.2.2 System Development Approach***

In our project we used object-oriented system development methodology (OOSD) for designing. The team used the object-oriented design methodologies for the development of the system among the different methodologies. Because it is better way to construct, manage and assemble objects that are implemented in our system.

Object oriented design methodology has two phases: -

* **Object Oriented Analysis (OOA):** During this phase the team will look at the problem domain and with the aim of producing a conceptual model of the information that exists in the area which will be analyzed. And this model the functions of the system (use case modeling), identifying the business objects, organize the objects and also the relationship between them and finally model the behavior of the objects.
* **Object Oriented Design (OOD):** During this phase the model interactions and behaviors that support the use case scenario, and finally update object model to reflect the implementation environment. And also transforms the conceptual model produced in object-oriented analysis to take account of the constraints needed to our system, so that we will use this phase to improve the use case model to reflect the implementation environment.

We decide to use object-oriented system analysis and design methodology because of the following reason:

* Object-oriented techniques work well in situations where complicated systems are undergoing contentious maintenance, adaptation and design.
* **Simplicity:** software objects model real world objects, so the complexity is reduced and the program structure is very clear.
* **Re-usability**: the object oriented provides opportunities for reuse through the concepts of inheritance.
* **Modifiability: -**when we need to add new feature to the system, we only need to make changes in one part of the applicable class.

### ***3.2.3 System Development Tool***

#### 3.2.3.1 Hardware Tools Required

* Personal computer (PC): almost all tasks of our project are performed on computer.
* Flash disk (16GB): required for data movement to store & transfer data from one PC to another PC.
* Modem and LAN of 512 mbps broad band connection.
* Network cable: help us to extract relevant information about our project from internet.
* Stationeries (pen, paper): for writing all necessary documentations associated with the project.
* Note book: to take notes during data collection and for other.

#### 3.2.3.2 Software Tools Required

**Python**: **-** Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. We list down some of the key reasons we select python as a programming tool:

* **Python is Interpreted** − Python is processed at run-time by the interpreter.
* **Python is Interactive** − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
* **Python is Object-Oriented** − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
* **Python is a Beginner's Language** − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

**Django Framework**: Django is a free and open-source web application framework written in Python. It is used for rapid web development and clean, pragmatic design. It is built by experienced developers to make repetitive tasks easier, so we can focus on writing apps instead of reinventing the wheel. Some features that make Django an ideal framework for web application development are as follows:

* **Super-fast**: Django development is extremely fast. Our ideas can take the shape of a product very quickly.
* **Fully loaded**: Django has dozens of projects that can be integrated to carry out common tasks such as user authentication, authorization, and content administration.
* **Versatile**: Django can be used for almost any kind of project, from CMSs to e-commerce apps to on-demand delivery platforms.
* **Secure**: Django also has support to prevent common security issues, including cross-site request forgery, cross-site scripting, SQL injection, and clickjacking.
* **Scalable**: Django websites can scale fast to meet high traffic demands.
* **Maintainable:** Django code is written using design principles and patterns that encourage the creation of maintainable and reusable code. In particular, it makes use of the Don't Repeat Yourself (DRY) principle so there is no unnecessary duplication, reducing the amount of code.
* **Portable:** Django is written in Python, which runs on many platforms. That means that you are not tied to any particular server platform, and can run your applications on many flavors of Linux, Windows, and mac OS. Furthermore, Django is well-supported by many web hosting providers, who often provide specific infrastructure and documentation for hosting Django sites.

**SQLite**: - SQLite is a very easy to use database engine. We select to use SQLite because It is self-contained, server-less, zero-configuration and transnational. It is very fast and lightweight, and the entire database is stored in a single disk file

SQLite doesn't run as a server process, which means that it never needs to be stopped, started, or restarted and doesn't come with any configuration files that need to be managed.

**Microsoft office Word**: -for documenting the corresponding deliverable associated with the project

Notepad++, visual studio: -For writing a code or program of the system.

Browsers: -Chrome, Firefox, Linux browser

## **3.3 Methods for Illegibility prediction Feature of Our System**

### ***3.3.1 Overview***

Research methodology is the technique to scientifically solve the research problem. Detailed description has made about the procedures followed to conduct the study. It includes we review Machine Learning method, discuss why regression methods are what we need to predict illegibility of candidates, discussions about research design, sources of data, data collection techniques, further as tools utilized to design and develop the prototype system. The aim is to clarify the procedures and steps adopted that address the research problem.

### ***3.3.2 Machine Learning***

Machine learning is a discipline that evolved from artificial intelligence, but it focuses more on cognitive learning capabilities. AI has many other aspects that attempt to model human function and intelligence (such as problem-solving). However, ML is a subset technology specific to the use of data to simulate human learning Sharp. Machine learning algorithm will be classified into four, namely supervised, unsupervised, semi-supervised and Reinforcement learning, however the foremost known algorithms are supervised and unsupervised learning. From the categories, supervised learning requires a training data set with determined labeled class while unsupervised don’t require to possess labeled class. An issue solved through classification and regressions methods are classified under supervised learning, due to data sets have known output values. Clustering problems may be solved with an unsupervised approach. Machine learning draws on concepts and results from many fields, including statistics, AI, philosophy, scientific theory, biology, scientific discipline, computational complexity, and control theory.

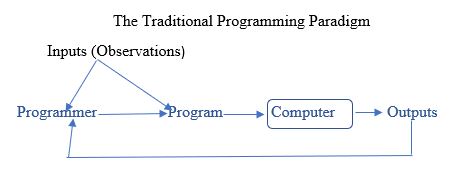


Figure 3. 1: Traditional programming paradigm

Machine learning is field of study that gives computers the ability to learn without being explicitly programmed.

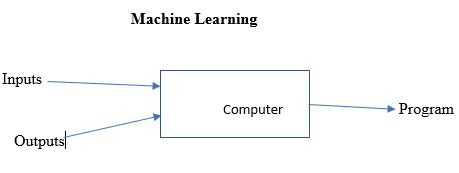


Figure 3. 2: Machine learning programming paradigm

### ***3.3.2.1 Supervised Learning***

The main idea of supervised learning is to learn a mapping between the input and the output whose correct values provided by a supervisor (Antonelo, E. A., & Schrauwen, B. (2010, May)).

Supervised learning is that the machine learning task of inferring a function from labeled training data (Antonelo, E. A., & Schrauwen, B. (2010, May)). There are two main types of supervised learning, classification and regression, where there is an input and output, and the main role is to find a mapping between the input and the output. In classification, the task is to assign the training input to one of the predefined classes. The training data accommodates a group of coaching examples. The outputs may be real numbers in regression or class labels in classification. As classifying an input into two or more classes, supervised learning is usually referred to as classification. Generally, Classification concerns on building predictive model for function with discrete range, while regression concern on continuous range model building.

**1. Regression Method**

Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It’s used as a method for predictive modeling in [machine learning](https://www.seldon.io/what-is-machine-learning/), in which an algorithm is used to predict continuous outcomes.

Solving regression problems is one of the most common applications for machine learning models, especially in supervised machine learning. Algorithms are trained to understand the relationship between independent variables and an outcome or dependent variable. The model can then be leveraged to predict the outcome of new and unseen input data, or to fill a gap in missing data.

**2. Classification Method**

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Classes can be called as targets/labels or categories. Unlike regression, the output variable of Classification is a category, not a value. In classification algorithm, a discrete output function(y) is mapped to input variable(x).

### ***3.3.3 Data Collection***

Our team collects the data by using A technique called collecting pre-existing datasets from official sources because this option is faster and requires little or no technical knowledge. The datasets on this types of source are usually available in CSV, JSON, HTML, or Excel format.

The secondary sources of knowledge are collected by using document analysis. Moreover, secondary sources of knowledge are gathered from the Internet, research, journal, conference, and reading textbook related to our title to develop our knowledge on the prediction of illegibility for the particular job position. After we understand factor that affect the price of rental house, we collect more than 15049 record of dataset from the internet. After we clean all unreliable data, missed value, duplicate records, and other unnecessary values only remain 14922 records of datasets.

Table3.1 List of Features we used

|  |  |  |
| --- | --- | --- |
| Name of the attribute | Description | Data Type |
| Area | Living area in square feet | Int |
| No of bathrooms | Number of bathrooms | Int |
| No of bedrooms | No of bedrooms | Int |
| Renter\_type | Either landlord or broker | String |
| Furnishing |  | String |
| Locality | Specific location | String |
|  |  |  |
| Price | Rental price of the house | Int |

### ***3.3.4 Preprocessing***

Data Preprocessing is done in order to transform the data set into a clean data set for better Machine Learning Model. Data preprocessing are applied to data in raw format, which is not feasible for analysis. In our case we collect the data from online sources, so there will be missing value, data in various format, and incorrect data.

In Ethiopia, there are several housing websites (ethiobethoch, qefira, realethio, zegebeya, elfegn, ethiopianhome, etc.) that used to advert items including houses for their customer. But for our study, there is no website that have a data for Arba Minch city all the websites contain data only for Addiss Ababa city so we collect the data from the internet and we try to customize the dataset in a way suitable with our specific location. All data processing steps are done in Excel’s CSV format. In order to solve those problem our team use the following data preprocessing techniques:

#### **3.3.4.1 Data Cleaning**

Data in raw form are not good for analysis. To be useful for predictive modeling the data must first be cleaned. The purpose of data cleaning is to improve the quality of data by identifying and removing errors and inconsistencies from data. Data quality problems are often present due to mistakes during data entry, missing information, duplicates, redundant data, or other general invalid data. The data cleaning process includes removing missing or duplicate information, filtering unnecessary data, and combining different data representations to provide access to consistent and accurate data. We removed all data not meeting the following criteria:

* The house has a non-missing house rent price.
* The house has a non-missing house total area.
* The house must be in Ethiopia.
* The house is a residential house, city home, or apartment.

The data in the CSV need to be checked whether it has any null values, as the data source has missing values every attribute has checked using the filters and missed/blank values are removed which helps to increase the accuracy level of the dataset. To remove invalid input from dataset we use feature engineering methods explained briefly on chapter five.

#### **3.3.4.2 Data description**

After data cleaning, our data set consists of 14922 data records and nine (8) attributes (Renter\_type, Specific location, total area, number of bedrooms, furnishing, Property\_Type, Number of Bath room, and price of the house) with complete information. The characteristic features or attributes of the house (Locality, total area, number of bedrooms, presented month for sale, presented year for sale) that are used to predict the house price are called independent or predictor variables, and house price is called target variable or dependent variable. The price of house is depending on other attributes listed above. Here, we will provide a brief description of dataset features or attribute in the following table.

|  |  |
| --- | --- |
| Feature/Attribute | Description |
| Renter\_type | Describe who is the renter (owner of the house or agents) |
| Bedroom | The number of bedrooms |
| Bathroom | The number of bathrooms |
| Property\_Type | The type of property to be rented either apartment, house or room |
| Area | Total number of area |
| Furnish\_Type | Is it furnished, semi furnished or unfurnished |
| Price | The rental price of the house |

**Data type**

The housing data we used in this study have 15049 rows and nine (9) attributes. The description of the data set data type is given below: -

Description of attribute values:

AGENT 13664

OWNER 1271

BUILDER 114

Name: Renter\_type, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2.0 7392

1.0 4393

3.0 2758

4.0 462

5.0 39

6.0 3

8.0 2

Name: bedroom, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

BHK 14612

RK 437

Name: layout\_type, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Apartment 13878

Studio Apartment 437

Independent House 411

Villa 207

Independent Floor 116

Name: property\_type, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Secha 6677

Limat 5264

Sikela 3108

Name: locality, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

18,000 830

15,000 814

20,000 720

16,000 685

12,000 667

...

16,001 1

19,190 1

35,001 1

15,660 1

6,750 1

Name: price, Length: 335, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1000.0 656

1200.0 591

650.0 583

600.0 554

1100.0 471

...

1716.0 1

1394.0 1

2259.0 1

1274.0 1

1772.0 1

Name: area, Length: 1011, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Unfurnished 6383

Semi-Furnished 5708

Furnished 2958

Name: furnish\_type, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2 7766

1 4233

3 2413

4 425

5 78

East facing 14

6 7

West facing 5

Grfloor 4

North facing 2

NorthEast facing 1

11 of 11floor 1

Name: bathroom, dtype: int64

### *3..3.5 Data Analysis Method*

We analyze the data to understand our data using different graphical representations such as heat map, bar graph

### ***3.3.6 Modeling Techniques***

We build our model by using three different algorithms (Linear regression, KNN, Random Forest) and by evaluating the performance of the models on the test data we select one algorithm that performs well with our dataset.

### ***3.3.7 Programming Technique and Tools***

we use a machine learning algorithm, a regression approach for predicting the price of the house. Regression a data mining task of predicting the value of a target (numerical variable) by building a model based on one or more predictors the predictors can either be numerical or categorical variables. To test/run collected sample data we also use python Jupiter notebook and other anaconda applications. And to make my reference standard reference style, I use Mendeley reference management software.

# Conclusion

Computer technology has revolutionized many aspects of life, particularly in the field of industry, business, medicine, technology and education. This project aims to develop a Web based CV analysis and Illegibility prediction system to help HR provide reliable services. It will help applicants to access the system easily, register online and apply for particular job position and provide illegibility prediction model developed by using machine learning. Generally the system provides suitable way to shortlist candidates that apply for particular job position

# *References*

Analyzing and shortlisting the CV of candidates who apply for some job is becoming difficult and time taking process in a lot of companies. The main objective of this project is to study and understand the current system problems and provide an automated system as a solution to enable the CV analysis and ranking to cope with technological advancements. The project we will develop is Web based CV analysis and Illegibility prediction system. We inspired to develop this project because there is a workload in the human resource department analyzing all the applicants CV manually and also there is unfair processes in CV ranking to choose illegible person for the position required. To avoid these and other related problems we are proposing to develop Web based CV analysis and Illegibility prediction system. In order to develop this system we used observation and interview data collection methodologies and the source of data for the functionality of the system is primary as well as secondary data. Our project will enable a more effective way to short list submitted candidate CV from a large number of applicants providing a consistent and fair CV ranking policy, which can be legally justified and will predict illegible person for the position. Candidate here will register him/herself with all its details, apply for the particular job posted by the admin and will upload their own CV into the system, which will be further used by the system to shortlist their CV. The system will rank the CV based on the experience and other key skills which are required for particular job profile. This system will help the HR (human resource) department to easily shortlist the candidate based on the CV ranking policy. It will focus not only in qualification and in experience but also focuses on other important aspects, which are required for particular job position and it will help the human resource department to select right candidate for particular job profile, which in turn provide expert workforce for the organization.