Arab American University

Faculty of Graduate Studies

Machine Learning Based Analysis of the Impact of University Specialization on Unemployment Rate

By

Aisar Misara Atieh Hussein

Supervisor

Prof.Ana Fernández Vilas

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in the data science and business analytics

March / 2024

I

Machine Learning Based Analysis of the Impact of University Specialization on Unemployment Rate

By

Aisar Misara Atieh Hussein

This Thesis was defended Successfully on **22/02/2024** and approved by:

Committee members	Signature
1. Prof. Ana Fernández Vilas: Supervisor	……
2. Dr. Amani Yousef Owda: Internal examiner	
3. Prof. Manuel Fernández Veiga: External examiner	

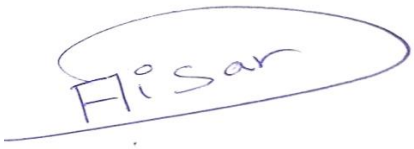
Declaration

I certify that this thesis titled "**Machine Learning Based Analysis of the Impact of University Specialization on Unemployment Rate**" is my work, has been completely done by myself, and does not contain work from other related works or research done in the same domain, and only submitted especially to complete a master's degree in data science and business analytics.

Student Name: Aisar Misara Atieh Hussein

Student Number: 202113261

Signature



Date:13-07-2024

Dedication

My sincere gratitude goes out to my family (my father, brothers, sisters, my wife Bisan, (Nadeem and Joury my children)) in their unwavering support and encouragement throughout this academic journey. The love and understanding they have shown me have been a source of strength and inspiration to me.

Acknowledgment

I am deeply grateful to Professor Ana for her support and guidance during the completion of this thesis. Her commitment, encouragement, and expertise played an important role in shaping and refining it. Her mentorship, patience, and unwavering belief in my abilities has been invaluable to my academic success.

Abstract

Palestine, like most developing and Arab countries, suffers from the problem of unemployment. This problem in the Palestinian economy is linked to many economic and social variables, especially growth. As in 2022, the unemployment rate among those participating in the labor force in Palestine stood at 24.4%, compared with 23.0% in 2015. Moreover, a significant percentage of Palestinian youths and graduated face limited job prospects, leading to frustration and disillusionment. High levels of youth unemployment not only jeopardize individual opportunities but also pose a broader societal challenge, potentially contributing to social unrest. According to figures from the Palestinian Central Bureau of Statistics (PCBS), the unemployment rate for individuals aged 20 to 29 with an intermediate diploma, bachelor's degree, or higher has reached more than 48%. This translates to half of the graduates being unemployed.

Given the importance of the issue of unemployment among graduates in Palestine, in this research we learned about the reality of unemployment among graduates, and worked to apply machine learning techniques to classify labor force data to determine the reality of young graduates in Palestine in terms of work and unemployment, and determine the impact of some variables on employment status such as (gender, age, specializations university, region, governorate,...), as unemployment deserves special attention among university graduates and their field of study.

Two sets of classification models were built: one with five algorithms (Random forest (RF), Decision tree (DT), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors Algorithm (KNN), Adaptive Boosting (AdaBoost)) for all graduate data, and another with three algorithms (RF, DT, and KNN) for data grouped by university major.

RF achieved an impressive 95.8% accuracy on the test data, DT was close behind with 93.4%. Other models like XGBoost, KNN, and AdaBoost performed well too, with accuracies 91.6, 88% and 91.8% respectively. Also the RF excelled in specialization classification, achieving impressive accuracy across all four categories: business and administration and law (94.9%), engineering, manufacturing and construction (94.7%), education (94.1%), and health & welfare being (94%).

Keywords: machine learning, labor force survey, random forest, graduates unemployment.

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
LF	Labor Force
KNN	K-Nearest Neighbors Algorithm
DT	Decision Tree
RF	Random Forest
EDA	Extra Tree Classifier
FN	False Negative
ETC	Extra Tree Classifier
EDA	Exploratory Data Analysis
FP	False Positive
FPR	False Positive Rate
ANN	Artificial Neural Networks
AUC	Are Under the curve
MRL	Multinomial regression logistic
LR	Logistic Regression
PCBS	Palestinian Central Bureau of Statistics

Chapter One: Introduction

1.1 Background

The youthful demographic stands as the dynamic and industrious faction within any society, tasked with a crucial role in navigating through crises and overcoming challenges. Serving as the backbone of the community, Palestine has a wide segment of youth within its societal fabric, with the percentage of youth aged between 20 and 29 over a fifth of the Palestinian demographic in 2023, this age group represents a significant 22% of Palestine's population [1], presenting a considerable challenge for policymakers aiming to fulfill the ambitions of these young individuals, notably their primary concern of securing employment opportunities.

Investing in education stands as the genuine asset for Palestinians. Given its significance at both individual and societal levels, statistics reveal that among every 100 young adults aged 20 to 29, 20 individuals have obtained a bachelor's degree or higher, underscoring the pivotal role education plays in their lives, Figure 1.1 illustrates the substantial growth in Palestinian enrollment in education. Over the years, data shows a sustained increase in the number of students attending universities. For instance, in 1994, approximately 29,000 students were enrolled in universities or college, while by 2022, this figure increased significantly to around 214,000 students. Consequently, the number of graduates also surged from about 3,000 in 1994 to an estimated 46,000 in 2022 [2].

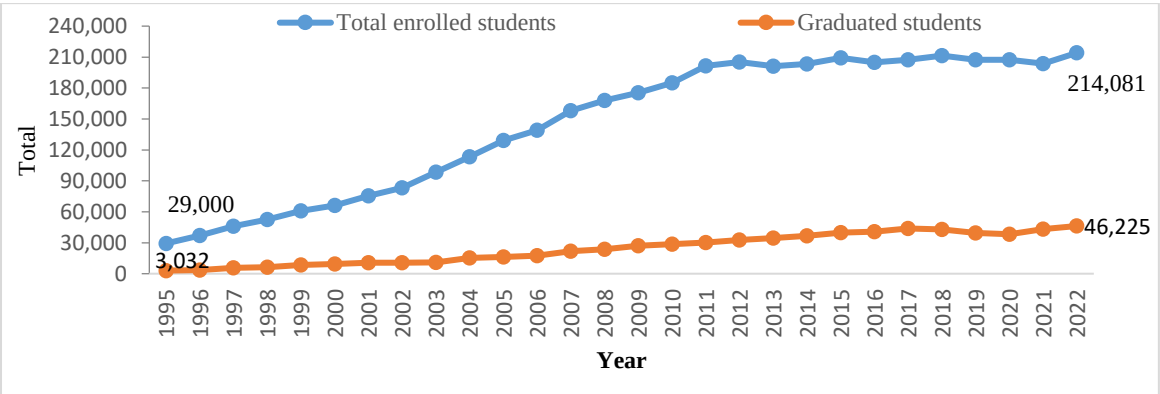


Figure 1.1: Total number of enrolled students, and number of graduated students in the Palestinians university during 1995- 2022

One of the problems that the world suffers from among young people and educated people is unemployment, as it constitutes a major and global challenge, regardless of a country's developmental stage. It stands as a pressing socio-economic issue with global repercussions, impacting individual quality of life. The root cause often lies in an inadequate demand within

the economy, resulting in decreased labor requirements, in addition to the weakness of the labor market in the country, this potentially leading to reduced work hours or layoffs [3]. As this situation not only squanders human potential but also generates adverse economic effects, acting as a substantial barrier to holistic growth and social stability. Policymakers and stakeholders should pinpoint specific interventions aimed at bolstering employment opportunities for all segments of the population. Of particular concern is the fact that a substantial portion of the unemployed comprises young individuals, exacerbating the problem further.

Palestine, like most developing Arab countries, suffers from the problem of unemployment. This problem in the Palestinian economy is linked not only to economic but also social variables, especially population growth, government actions, direct credit facilities, consumption, size of labor market, etc. In 2022, the unemployment rate among those participating in the labor force (LF) in Palestine stood at 24.4%, compared with 20.9% in 2011 [4]. Moreover, a significant percentage of Palestinian youths and graduates face limited job prospects, leading to frustration and disillusionment. High levels of youth unemployment not only jeopardize individual opportunities but also pose a broader societal challenge, potentially contributing to social unrest. According to figures provided by PCBS, the unemployment rate for individuals aged 20 to 29 with an intermediate diploma[1], bachelor's degree, or higher has reached more than 48%, almost half of the graduates being unemployed [5]. From Figure 1.2, it becomes evident that unemployment rates in Palestine are on the rise, with the most substantial proportion observed among university graduates. This highlights the profound level of frustration experienced by young graduates due to the prevailing conditions within the Palestinian labor market, in addition the shortage of the local government action to solve this problem.

Palestinian youth and graduates face a bleak employment landscape, fueling frustration and disillusionment. This widespread youth unemployment isn't just a personal obstacle; it's a societal ticking time bomb with the potential to trigger unrest. Statistics paint a grim picture: the PCBS reports that nearly half (48%) of young adults (20-29) with intermediate diploma,

[1] It refers to a two-year post-secondary degree awarded by community colleges and technical schools

bachelor's degrees, or higher are unemployed[2] [1]. Figure 1.2 further underscores this alarming trend, revealing that unemployment rates in Palestine are climbing, with university graduates bearing the brunt. This stark reality breeds deep frustration among young graduates, a direct consequence of both the harsh realities of the Palestinian labor market and the apparent lack of government action to address this critical issue.

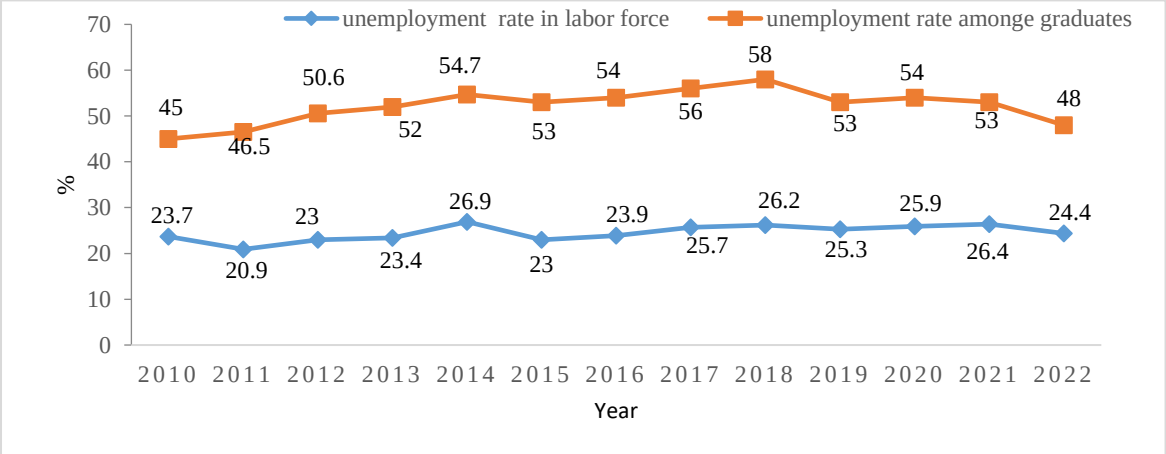


Figure 1.2 Unemployment rate in Palestine , and unemployment rate among graduates who possess an intermediate diploma certificate or higher between the years 2010 and 2022

The surge in graduation unemployment in Palestine stems from multiple factors. Firstly, it relates to the constrained and undersized job market, unable to match the pace of population growth and the predominant youthful demographic structure. Moreover, insufficient information exists regarding the dynamics of supply and demand within the labor market. The unemployment prevalent among Palestinian graduates mirrors the reality of higher education institutions and their curriculum challenges [6]. These encompass issues concerning scope, the struggle to keep pace with scientific advancements, and a misalignment with job market demands. Additionally, problems persist within vocational education and auxiliary programs, stemming from insufficient resources, inadequate equipment, and underdeveloped infrastructure. The conventional educational methodologies further compound these issues. Collectively, these factors contribute to adverse impacts on educational outcomes, failing to align with the requisites of the labor market. Consequently, numerous young graduates

[2] A person who is unemployed is a person aged fifteen years and above who has not worked during the reference period, has not been absent from their job, has been available for work, and has actively sought a job during the last four weeks

encounter fresh hurdles, finding themselves unable to secure jobs commensurate with their educational background and the skills acquired during their university education. So, the Palestinian universities and colleges predominantly offer conventional specializations, lacking modern disciplines that hold high demand within the job market.

1.2 Problem Statement

The issue of unemployment has persisted as a challenge for societies throughout the ages, and one of the most serious forms of unemployment is the unemployment of graduates, as unemployment of graduates represents the largest share of the unemployed in Palestine, in 2022 about 46,000 students graduated from Palestinian universities [1] in various specializations, while only 8,000 job openings for graduates have been absorbed by the local market; this disparity poses a developmental challenge.

Given the importance of the issue of unemployment among graduates in Palestine, in this research we will learn about the reality of unemployment among graduates, specifically, machine learning techniques will be applied to analyze LF data and extract conclusions about the current situation of young graduates in terms of work and unemployment. Also, the impact of some socio-economic variables will be studied (gender, age, specializations, region, governorate, etc.) and special attention will be put among university graduates and their field of study. Throughout this thesis, we will analyze data sourced from the LF survey conducted periodically by the Palestinian Central Bureau of Statistics, the survey aims to understand the Palestinian labor market by estimating the number of people in the workforce (employed and unemployed) and their characteristics (age, sex, education, etc.),and tracking changes over time: By conducting the survey regularly, and the data provides insights for policymakers to evaluate and develop better employment policies to improve overall economic productivity.

1.3 Research Objectives

1. Learn about the labor market, and the reality, development, and burdens of the unemployment problem in Palestine, with reference to current unemployment indicators.
2. Analyzing the LF and classifying it according to gender, age, and educational level.
3. Studying the extent of compatibility between the graduates' qualifications and the requirements of the Palestinian labor market.
4. Propose suitable models to classify the status of unemployment among graduates in Palestine.

5. Explore the potential of the models to assist policymakers and other stakeholders in taking action to reduce graduate unemployment.

1.4 Research Contribution

Reducing unemployment stands as a paramount economic goal for policymakers in any country. The objective of this study is to create a learning model using real data from the LF Survey, which periodically collects information from PCBS. The model will classify graduates based on various characteristics, primarily focusing on their scientific specialization and other features.

Some studies have categorized employment in Palestine, yet they have not specifically concentrated on the segment of graduates. Notably, to my understanding, there is a scarcity of studies employing longitudinal analysis concerning graduates aged between 20 and 29, despite the assumed significance of such research on this demographic segment. To address this gap, my contribution will include conducting a study using machine learning techniques to classify university and college graduate students between the ages of 20 to 29 years as employed or unemployed, through 15 characteristics obtained from a LF data survey.

1.5 Outline of the Thesis

As for the rest of the thesis, it will be arranged as follows: chapter two will encompass a comprehensive review of unemployment literature, specifically focusing on graduates. It will delve into studies concerning the algorithms used for data classification, whereas chapter three will involve an in-depth analysis of the data, exploring the correlations between the variables. Graphical representations will be utilized to extract deeper insights from the dataset, in addition to introduce proposed methodologies for data classification using machine learning algorithms, along with techniques for acquiring and optimizing optimal parameters. While Chapter four will showcase and discuss the outcomes of various classification methods employed across different scales. Lastly, Chapter five will encapsulate the conclusion, recommendations, and future directions derived from the research findings.

Chapter Two: Literature Review

2.1 Introduction

In this chapter, we examined past research and literature concerning unemployment in both the general LF and specifically among recent graduates. Additionally, we explored various methodologies involving machine learning algorithms utilized for classification purposes. In this chapter, we will cover the following topics: Unemployment Crisis in the Countries, Classification techniques for analyzing the LF in Palestine and other Arab countries, then summary of literature reviews alongside a table summarizing relevant studies.

2.2 Related Work

2.2.1 Unemployment Crisis in the Countries

The primary aim of study [7] is to identify the factors contributing to the unemployment crisis among graduates in Malaysia. The finding demonstrated significant correlation between employer preferences, candidate attributes, and economic instability with the unemployment crisis. These findings emphasize the need for better preparation of fresh graduates who are actively seeking employment, and it underscores the importance of the education ministry considering a revision of the higher education curriculum. This revision should align more closely with the requirements of the job market and the expectations of stakeholders, ultimately improving graduate employability.

The article [8] examined the determinants of unemployment in Nigeria using time series data. A significant relationship exists between gross domestic product (GDP), inflation, degree of openness, and private domestic investment. The study also revealed that the government should encourage private investment at all levels and also reduce the restrictions on trade and support the openness of trade.

The study [9] examines employability among management graduates in Sri Lanka despite significant government investment in higher education. Many graduates face challenges in finding suitable employment. A survey of 1,333 out of 10,033 graduates shows most are employed but seeking new jobs. Key factors in securing employment include education, personality, work experience, English proficiency, and professional qualifications. However, age and lack of experience are barriers. The study suggests policy measures to improve employment rates and employability among management graduates.

The research [10] findings highlight the significant issue of graduate unemployment and emphasize the need for further investigation into its impact on mental and psychological health in India. The research proposes urgent steps to mitigate the rising youth unemployment rates, including the introduction of specialized internship initiatives, the adoption of a merit-based and impartial hiring process, and the provision of career counseling services to students by university faculty members.

The article [11] highlights the importance of higher education in training future community members and the expectations of graduates to find higher-value jobs in Vietnam . The fear of unemployment affects how students plan their careers, and they tend to choose majors in the natural sciences rather than humanities or social sciences. The study surveyed 329 respondents on the reasons why fresh graduates are easily unemployed. The results show that there are 7 reasons, in order of importance: quality of graduates, professional qualifications, unsuitable work, lack of career orientation, job skills, market conditions, and untransparent recruitment.

A study on unemployment among Klang Valley graduates [12] reveals that strong grades alone aren't enough. Fresh graduates need proficiency in English and soft skills like leadership, communication, and critical thinking. The research suggests universities should prioritize skills training to create well-rounded graduates with strong work ethic. It also recommends further studies with a larger sample size, qualitative analysis, and a broader exploration of factors affecting graduate unemployment in Malaysia.

While the other article [13] identifies other skills that cause unemployment, its delves into the pressing concern of unemployment among Malaysian graduates and investigates the underlying factors contributing to this predicament, the study identifies that employability skills, job mismatch, and unrealistic salary expectations emerge as the primary culprits driving graduate unemployment, while English proficiency and graduate attributes are noteworthy, their correlation with the unemployment issue appears to be less robust. In light of these findings, the article advocates for policy interventions that target both the individual attitudes of graduates at the micro-level and institutional aspects, particularly the courses offered by universities, at the macro-level.

Whereas article [14] highlights the significance of aligning the enrollment of individuals in higher education institutions with the real workforce requirements to maintain a stable and sustainable job market. It stresses that an unsustainable surge in the count of highly educated graduates might lead to a rise in unemployment and continued emigration of highly skilled

professionals. The outcomes of this study offer valuable insights for policymakers and can guide the formulation of effective strategies tailored to the higher education sector.

The article [15] explores the significance of emerging fields that present potential opportunities for students seeking to secure employment. It examines the impact of the Fourth Industrial Revolution on the MENA (Middle East and North Africa) region's workforce, presenting it as a dual challenge and opportunity. On one hand, it poses challenges due to the potential for automation in various work activities. On the other hand, this revolution provides a chance to generate valuable formal sector jobs, especially in fields like data analysis, computer science, and engineering. Well-established industries in the region, such as oil and gas, aviation, transportation, and healthcare, can play a crucial role in spreading technology and nurturing the development of smaller enterprises.

In paper [16] discusses the youth unemployment crisis in the (MENA) region, highlighting high unemployment rates among youth, particularly those with tertiary education, low entrepreneurship, and vocational education participation, and the need for policy measures to improve labor market.

A study explores the gender unemployment gap in developing countries, focusing on Nigeria's high female unemployment rate [17]. Researchers employed a machine learning model to pinpoint factors affecting women's unemployment in Nigeria. The model identified age, ethnicity, marital status, and religion as important contributors to these high rates.

While article [18] examines the factors that determine the time it takes college graduates in Spain to obtain their first job after graduation. Using parametric and nonparametric duration models, the authors find that men are more likely to find employment than women, graduates of private universities gain their first job sooner than those from public universities, previous work experience and starting the job search before graduation lead to quicker employment, and Arts and Humanities graduates face the most difficulty in finding work. The study also suggests that graduates with international experience and expert knowledge of communication are more likely to find employment.

The Author [19] investigates the individual factors impacting the length of unemployment among individuals in the Turkish labor market. Survival analysis is employed as the econometric method to identify the sociodemographic and economic variables affecting unemployment duration. Furthermore, the research sheds light on the profound psychological,

social, and economic consequences associated with long-term unemployment among university graduates.

By using logistic regression (LR) the authors in the study [20] analysis data from the employer and graduate survey in Rwanda, the research identifies significant associations between the duration of first unemployment and factors such as the graduate's age, job search method, acquired skills, further graduate studies, monthly salary, The findings emphasize the need for the higher education sector to align curricula and delivery methods with labor market requirements to enhance graduate employability.

2.2.2 Classification Techniques for the Analysis of Labor Force

Several research endeavors have delved into employing data mining techniques to forecast and classification of students' employability. Among the various methodologies, classification and prediction algorithms have emerged as the primary methods extensively utilized in this field.

2.2.2.1 Studies in Palestine

In a study [21] centered on graduates' data in Palestine, the research aimed to compare Discriminant Analysis (DA) and Multinomial Logistic Regression (MLR) techniques, crucial for handling categorical dependent variables by Analyzing data from 2005-2006 with nine independent variables and a dependent variable (Employment Status of graduates) across three categories (Employment, Unemployment, Outside of LF), MLR achieved 67.2% correct classification compared to DA's 65.2%. Additionally, the ROC curves highlighted substantial differences in the area under the curve (AUC), with MLR at 91.42% and DA at 53.52%. Consequently, the MLR model outperformed the DA model for this dataset.

While the study [22] discusses the importance of tracking graduates to understand the labor market, it aimed to enhance the ability of Palestinian higher education institutions to understand the needs of the labor market by reviewing the literature and initiatives taken by these institutions and the private sector in the West Bank and Gaza Strip, as the recommendations highlighted the absence official tracking system for graduates within higher education institutions in Palestine. We also revealed that activities related to tracking graduates are limited and are in the early development stages, and are mainly implemented by a few universities and colleges, by distributing questionnaires for both graduates and employers.

Study [23] employs (MLR) and (DA) as classification methods to model LF data, the output was 83.5% correct classification by MLR compared to 81.1% in DA. The AUC curve is

significantly higher in MLR at 91.89% versus 52.8% in DA. The models classify observations into Employment, Unemployment, and Outside LF.

Researchers in [24] compared three machine learning methods for classifying employment status. The researchers used data from the LF dataset (2019), focusing on individuals aged 15-65. Eight questionnaire responses (independent variables) served as the basis for classification into employed or unemployed categories. The Artificial Neural Network (ANN) achieved the highest accuracy (82.7%), followed by LR at 81.6% and DA at 79.5%. Similarly, the AUC supported this trend, with ANN (85.8%) outperforming LR (81.7%) and DA (72.8%)

The result in the study [25] reveals a notable correlation between employment status in Palestine and residency, age, educational level, and educational specialization. Graduates residing in the West Bank exhibit higher employment prospects compared to those residing in the Gaza Strip. Additionally, older individuals tend to have increased employment opportunities, with each additional year contributing to enhanced odds of employment. Moreover, individuals holding a master's degree have a greater likelihood of employment compared to those with a bachelor's degree. Furthermore, graduates specializing in humanities and social sciences tend to have higher employment possibilities than those specializing in applied and natural sciences.

2.2.2.2 Studies in Arab Countries and Others

In a study [26] involving 658 women in Saudi Arabia, the research identified key factors influencing unemployment, including marital status, sources of income from family and husband, as well as the alignment of academic majors with the labor market. These factors significantly impacted whether women were employed or not, as observed in both methods used. Employing ANN and LR models exhibited a strong performance, achieving an 86% accuracy rate and using AUC as the evaluation metric, the study found no significant difference in the results when applying the AUC metric to these models.

This paper [27] aims to propose an effective classification model for predicting and evaluating the attributes of student datasets, aligning them with the workforce demands in the academic field. By utilizing both supervised and unsupervised Machine Learning Algorithms such as KNN, Naïve Bayes (NB), DT, ANN, LR, and Support Vector Machine (SVM), this study seeks to create a model that assists university management in strategizing long-term plans.

The study [28] utilized machine learning to predict students' employability by analyzing 27,000 data points encompassing Mock Job Interview Evaluation Results, On-the Job Training

(OJT) student performance rating, and general point average (GPA) of students from School Year 2015 to School Year 2018. Three algorithms (DT, RF, and SVM) were employed. SVM demonstrated the highest accuracy at 91.22%, outperforming DT (85%) and RF (84%).

This research [29] endeavors to predict the employment status of Malaysian graduates by analyzing factors contributing to employability and gauging graduate satisfaction with their academic activities and ICT skills, the study explored various classifiers and variable selection methods to enhance employment status predictions, with LR emerging as the best classifier for different consolidation approaches. Additionally, K-Means Clustering uncovered seven clusters, highlighting the satisfaction disparity among graduates, particularly a cluster expressing high satisfaction in ICT skills but lower satisfaction with curricular activities.

In this study [30], the researcher introduced a data mining model aimed at predicting employability in Morocco three classification algorithms (DT, LR, and NB) were employed. The study identified significant variables contributing to the prediction of graduates' employability. The DT algorithm exhibited an accuracy of 81.70% in classifying employment status.

A study in Morocco [30] used data mining to predict graduate employability. Three classification algorithms (DT, LR, and NB) were employed to identify factors affecting employment. The DT algorithm achieved the highest accuracy (81.70%) in classifying graduates as employed or unemployed.

This study [31] suggests a predictive artificial intelligence model, aims to bridge the gap between computing graduates' skills and industry requirements. It identifies crucial skills needed for employment in the computing field, experimental results indicate that ADA, SVM, and LR classifiers outperform others, achieving an accuracy of 89%

Building on data mining techniques, research [32] proposes a model to predict graduate employment status (employed, unemployed, or undetermined). The study compares various approaches, including Bayesian and tree methods. Notably, the Bayesian algorithm achieved the highest accuracy, reaching an impressive 99.7%.

The research [33] focused on comparing various classification algorithms to construct an ensemble predictive model for anticipating graduates' graduates from an electronics engineering program employability through robust data mining techniques. An ensemble model comprising RF,SVM, and NB demonstrated the highest cross-validated accuracy score,

reaching 93.3%. The study revealed through association rule mining and permutation feature importance analysis that grit significantly influences employability, encompassing the ability to acquire technical skills and professional certifications.

In the research [34] aimed to develop a model for predicting employability and assess its performance using statistical and machine learning techniques. Key variables influencing employability were identified, including graduation year, education level, field of study, university type, ranking, and region. Four models, namely LR, DT, RF, and KNN, were compared. Results indicated that RF and KNN models were suitable for employability prediction, achieving 70.5% and 70.1% accuracy respectively, with AUC values of 74% and 76% for the RF model. Notably, important variables for prediction included university ranking, master's degree, statistics major, national university status.

In study [35] a novel approach, combining DT and RF models, has been introduced for employment prediction. This method exhibits feasibility and adaptability, offering potential applicability in guiding schools towards enhancing employment prospects, and the elements of academic achievement and graduation qualification emerge as pivotal aspects influencing students' employability, the accuracy reached 81%.

The objective of the study [36] was to categorize graduates as employed, unemployed, by using five data mining algorithms: NB,LR, Multilayer Perceptron, KNN, and DT. LR emerged with the highest classification accuracy, reaching 92.5%. The resulting classification model holds potential benefits for college management, offering insights into the quality of graduates and avenues to refine the curriculum to align with industry requirements.

The research in [37] aims to forecast the employment paths of IT graduates, distinguishing between their entry into IT-related careers or other non-IT fields. Utilizing data mining algorithms in supervised learning, the research identifies key attributes and compares their accuracy. It employs various classification techniques such as NB, J48, Simple Cart, LR, and chaid algorithms. Historical data from the university job placement office tracer study spanning from 2011 to 2015, along with academic records of BSIT graduates, constitute the dataset. The outcomes underscore three crucial factors—IT_core, IT_professional, and gender—that significantly influence IT employability.

In study in [38] introduced a novel model aimed at forecasting the employment prospects of Khon Kaen University graduates. Employing Bayesian methodologies encompassing various techniques such as Naïve Bayesian Simple, Averaged One-Dependence Estimators (AODE),

Averaged One-Dependence Estimators with subsumption resolution (AODEsr), Bayesian networks, and Naïve Bayesian updateable, the outcomes highlighted the Averaged One-Dependence Estimators with subsumption resolution (AODEsr) algorithm as the most accurate, achieving a remarkable 98.3% precision rate. Following closely behind, the AODE method demonstrated a commendable accuracy of 96.1%

The research in [39] focuses on assessing graduates' employability, a crucial metric for higher educational institutions (HEIs). It analyzes a dataset of 27,000 records from 2015 to 2018, including 3,000 observations and 12 features like mock job interview evaluations, on-the-job training performance, and general point averages. Using six learning algorithms, including DT, RF,SVM,KNN, and LR, the study found that SVM achieved the highest accuracy at 91.2%, surpassing DT (85%) and RF (84%).

Author in [40] aims to create a predictive model for graduates' employment status with data through a tracer survey of Oman Tourism College alumni, predicting whether graduates are in full-time employment, part-time work, self-employment, or unemployed based on their profiles. The research found that the RF algorithm within DT methods achieved an accuracy of 96.3%, outperforming the REPTree algorithm at 88.1%. BayesNet, another algorithm variant, ranked third with an 84.5% accuracy rate. Attribute ranking using information gain and ranker methods identified occupation as the most influential factor for employability, followed closely by job sector. Other factors influencing employment status included specialization, degree, age, and various skill sets like personality development, cultural competency, leadership, interpersonal abilities, creativity, and problem-solving skills.

Whereas the study in [41] identified several key DM techniques for employability studies, including SVM, NB, PCA, LR, k-means algorithm, KNN, DT, ANN, text mining, and item response theory. The review indicates that, in terms of accuracy, LR, DT, ANN, RF, and Text-IRT stand out as the most effective DM techniques for studying employability.

This paper [42] conducts an exploratory review of 99 articles from the "Education and Training" journal focusing on "Graduate Employability" from 2005 to 2021. Through rigorous cleaning, classification, and analysis using NVIVO, the study reveals a diverse range of methodologies employed by authors. These include quantitative, qualitative, mixed, and experimental approaches, utilized to explore graduate attributes, employability skills, and related constructs. The studies aim to enhance employability, understand self-employability

perceptions, and gauge employer perspectives. Additionally, the review identifies three prominent conceptual frameworks to use for measuring graduate employability.

While the study [43] proposes early employability prediction and determining key factors influencing students' employability to mitigate this problem. Supervised Machine Learning methods were employed on data from 500 Engineering students in hyderabad to predict their employability. DT and SVM algorithms surpassed NB and KNN, achieving 98% accuracy in predicting employability. Critical determinants like communication skills, aptitude, mentorship, family income, and college teaching quality significantly impact students' employability, suggesting a need for early prediction and focusing on these factors for enhancing graduates' job readiness.

Table 2.1 Related studies and existing models for classification of labor force

Ref	Summary	Independent Variable /	ML Model	Output features	Accuracy	Country
[17]	Analyzing Factors Influencing Women Unemployment	Age groups, ethnicity, marital status, and religion, martial status, pregnancy term, region, educational level	RF,NB,AdaBosst, SVM	Status labor of women: (Employment, unemployment)	RF:86.7%.	Nigeria
[21]	Compare DA and MLR techniques, to classify the (Employment Status of graduates)	Age, Labor forces status, region, gender	DA, MLR	Employment Status of graduates: (Employment, unemployment)	MLR 67.2% DA :65.2%.	Palestine
[23]	Classify observations into Employment, Unemployment,and Outside LF (data 2012)	Gender, Age, Educational attainment, refuges status, region, industry group, relationship to the head of household, labor force status.	MLR,DA	Employment Status of LF: (Employment, unemployment, outside of labor force)	MLR:83.5% DA:81.1%	Palestine
[24]	Discover the optimal model capable of elucidating the correlation between various types of labor force and additional influencing factors. (data 2019)	Age, labor force status, region	DR, ANN, LR	Employment Status of LF: (Employment, unemployment, outside of labor force)	LR:81.6%, DA:79.5%, ANN:82.7%	Palestine
[26]	Identify key factors	Marital status, sources of income	ANN, LR	key factors influencing the	LR: 86%	Suadi arabia

	influencing unemployment,	from family and husband, academic majors		overall unemployment status among women		
[27]	Propose an effective classification model for predicting and evaluating the attributes of student datasets, aligning them with the workforce demands in the academic field.	Gender, program, academic	KNN,NB, DT, ANN, LR, and SVM	Model that assists university management in strategizing long-term plans.	NB have the highest	Malaysia
[28]	Predict students' employability by analyzing 27,000 data points	Student Performance Rating, and General Point Average (GPA)	DT,RF,SVM	Students' employability	SVM:91.2% DT:85% RF:845%	Philippines
[29]	Predict the employment status of Malaysian graduates by analyzing factors contributing to employability and gauging graduate satisfaction.	Academic activities and ICT skills,	LR,K-means	Employment status of graduates	LR: 61%	Malaysia
[30]	Predicting employability in Morocco	Socioeconomic variables, Demographics and activities	DT,LR,NB	Classifying employment status.	DT:81.70%	Morocco
[31]	Predictive artificial intelligence model , aims to bridge the gap between	Graduates' skills and industry requirements	SVM, and LR	Preparing Computing Graduates for High-Demand Industry Roles	LR: 89%	Suadi Arabia

	computing graduates' skills and industry requirements.					
[32]	Predict the employment status of graduates, distinguishing between employed, unemployed, or undetermined situations.	21 independent variable (gender, GPA, martial status, faculty,..)	NB,DT, C4.5	Predicting the employment outcomes of recent graduates (employed, unemployed, or otherwise)	NB:99.7%	Thailand
[33]	Construct an ensemble predictive model for anticipating graduates' graduates from an electronics engineering program	Technical skills and professional certifications	RF,SVM, and NB	Predictive model for anticipating graduates' graduates	NB:93.3%.	Philippines
[34]	Develop a model for predicting employability and assess its performance.	Graduation year, education level, field of study, university type, ranking, and region	LR,RF,KNN	Employability prediction	KNN: 70.5% RF: 70.1%	Thailand
[35]	Using data mining in technology to predict employment based on five key factors	Academic achievement and graduation qualification emerge as pivotal aspects influencing students' employability	DT,RF	Predict employment	DT:81%	China
[36]	Categorize graduates as employed, unemployed.	Program, graduation year, education level, field of study, university type,	NB,LR,KNN,J48	Employment status of graduates	LR:92.5%	Malaysia
[37]	Forecast the employment paths of IT graduates, istinguishing between their entry into IT-related careers or other non-IT fields	11 variables such as (gender, location, It-related, CGPA, It-professions)	NB,LR,J48	Forecast the employment of IT	LR:78.4%	Indonesia

[38]	Forecasting the employment prospects of Khon Kaen University graduates.	Education gender, board of, degree, years of work experience.	AODEsr	Employment (employment, unemployment)	AODEsr:98.3 %	Taiwan
[39]	Assessing graduates' employability, a crucial metric for higher educational institutions (HEIs).	12 features like job interview evaluations, on-the-job training performance, and general point averages	DT,RF,SVM		SVM:91.2% DT:85% RF 84%.	Philippines
[40]	Create a predictive model for graduates' employment status with data through a tracer survey of Oman Tourism College alumni	Age,major,etc..	DT,REPTree,NB,	Graduates' employment	DT: 96.3%	Oman
[43]	Proposes early employability prediction and determining key factors influencing students' employability	communication skills, aptitude, mentorship, family income, and college teaching quality significantly impact students' employability	DT,KNN, SVM,NB	Determining key factors influencing students' employability	KNN:98%	India

2.3 Summary of the Literature Reviews

The literature review encompasses an extensive array of research focusing on unemployment factors, particularly emphasizing recent graduates' employability. It delves into various elements influencing employment status, including academic qualifications, skills, age, language proficiency (especially english), region, gender, ethics, and market demand. The studies explore these factors using a diverse range of algorithms such as LR, DT, ANN, KNN, SVM, RF. Additionally, classification methods like DA and MLR are employed. These collective efforts aim to improve comprehension, prediction, and classification related to graduate employment.

In Palestine, employment classification has been the subject of two studies focusing on various factors. These studies primarily encompassed the entire labor force aged 15 to 64 during 2014 and 2019. The emphasis was on utilizing logistics and classification methods while selecting

specific variables from surveys. Another study targeted graduate classification in Palestine based on a dedicated survey conducted in 2004.

This current study will utilize labor force data from the years 2020 to 2022, focusing solely on graduates possessing an intermediate diploma or higher. It aims to introduce field of specialization as a consistent variable that has not been previously addressed in any study. Additionally, numerous other variables will be incorporated, differing from those analyzed in previous studies. Furthermore, a machine learning technique will be employed to enhance accuracy in classification.

Chapter Three: The Data and Methodology

3.1 Introduction

This chapter is considered the basic cornerstone of our journey into the world of data exploration, as this chapter begins with a comprehensive exploration of data through a description of this data, then presenting it in graphical forms, and delves into the methods by which it was collected or obtained, and then the methods are discussed. Which is followed in cleaning the data and preparing it appropriately. This Chapter includes data preprocessing, often referred to as data preparation or data cleaning, revolves around the identification and rectification of errors and misleading information within a dataset. The depicted architecture in this research is presented in Figure 3.1 It outlines the steps involved in classifying the labor force among graduates using the Palestinian dataset acquired.

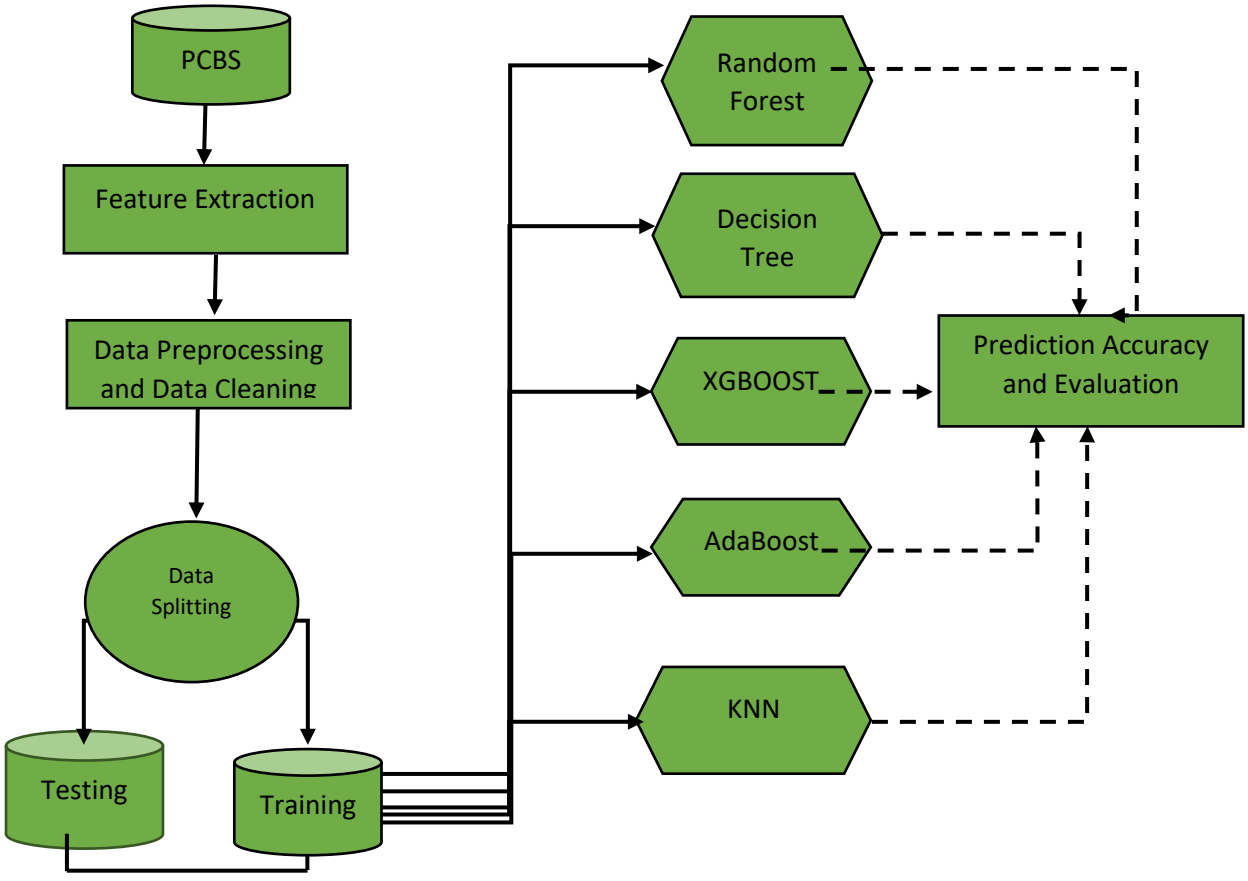


Figure 3.1 The proposed architecture for labor force among graduates classification model

3.2 Data Sources

The data for this study originates from the " Annual report on the Palestinian Labor Force Survey," which was collected by the Palestinian Central Bureau of Statistics (PCBS) for year 2020-2022 . The primary purpose of gathering data regarding the labor force and its various elements, encompassing employment, and unemployment, is to furnish fundamental insights into the dimensions and composition of the Palestinian labor force. Information gathered at various time intervals serves as a foundation for tracking contemporary patterns and shifts within the labor market and employment landscape. This data, when coupled with details concerning other economic facets, establishes a framework for assessing and scrutinizing macroeconomic policies.

The Features (independent variables) were selected based on a review of prior research and consultation with experts in this field

The characteristics and discrete values investigated for the study as delineated in the following Table 3.1:

Table 3.1 Description of data

Feature	**Type**	**Description**	**Value Label**
Year	Numeric	Year	2020-2022
Gender	Categorical	Gender	1. Male 2. Female
Refugee Status	Categorical	Refugee Status	1. Registered 2. Not Refugee
Age	Numeric	Age	20<Age<29
Higher Qualification	Categorical	Higher Qualification	1- Intermediate Diploma 2- BA\ BSc 3- Higher Qualification (Higher diploma,master degree,PHD)
Country of graduates	Categorical	Country of graduates	1- Palestine 2- Others
Graduation from a technical college	Categorical	Graduation from a technical college	1-Yes 2-No
Attendance training course	Categorical	Attendance training course during last 12 months	1-Yes 2-No

Difficulty	Categorical	IF/He/She have any difficulty in (vision, Hearing, Mobility and using finger, remembering and concentrating,	1-Yes 2-No
Marital Status	Categorical	Marital Status	1- Never Married 2- Married 3- Other
Governorate	Categorical	Governorate	1-Jenin 2-Tubas and Northern –valleys 3-Tulkarem 4-Nablus 5-Qalqilia 6-Salfeet 7-Ramallah & Al-Bireh 8-Jericho & AL Aghwar 9- Jerusalem 10-Bethlahim 11-Hebron 12- North of Gaza 13-Gaza City 14-Deir ALBalah 15-Khan Younis 16-Rafah
Type of Locality	Categorical	Type of Locality	1-Urban 2-Rural 3-Camp
Region	Categorical	Region	1. West Bank 2. Gaza Strip
Locality	Categorical	Locality	1- North of the West Bank

			2- Middle of the West Bank 3- South of the West Bank 4- Gaza Strip
Specialization	Categorical	Specialization of education	1- 'Education' 2- Language 3- 'Arts' and 'Humanities' 4- Social sciences, journalism and information 5- Business, administration and law 6- Natural sciences, mathematics and statistics 7- Information and Communication Technologies (ICTs) 8- Engineering, manufacturing and construction 9- Health and welfare 10- Others
Labor Force Status	Categorical	Labor Force Status	1- Employment 2- Unemployment

3.3 Sampling Design

Data was collected from the PCBS through the labor force survey. This survey is conducted periodically and quarterly on Palestinian families, with a focus on the years 2020 to 2022. These years were selected because they align with the International Standard Classification of Education (ISCED) [44] classification of university specializations established in 2013, which was applied to labor force data from 2020 onward.

The sample size for the labor force survey encompassed for 3 years was 100,000 households participated as respondents distributed across four quarters, the respondents are all individuals aged 15 years and above.

After filtering the data for this study, some criteria were selected, including selecting individuals between the ages of 20 years and 29 years old, holding an intermediate diploma

and having graduated from the university, and their working status as working or not working, so the final sample was 11,948 graduates.

The dataset comprises 16 variables, with a primary focus on the labor force status variable which will be as dependent variable, it has two binary categories, namely (Employment and Unemployment), and the remaining 15 variables serving as independents The objective is to identify the optimal model for characterizing the associations between various labor status types and independent variables, each of which can influence the respective dependent variable.

3.4 Data Preprocessing

This process entails the detection and correction of flawed, incomplete, or irrelevant data, ensuring the necessary adjustments are made. The preprocessing of data plays a crucial role in training AI models, particularly in machine learning models, as it is essential for ensuring accurate outcomes. It involves a range of tasks, such as data formatting and feature creation, tailored to the specific requirements of the AI project. This preparatory phase not only enhances data quality but also streamlines the modeling process, ultimately resulting in more dependable predictive models. Its significance in achieving meaningful results in machine learning cannot be overstated.

Data preprocessing encompasses a sequence of procedures that include:

3.4.1 Data Cleaning

Data cleaning encompasses identifying and correcting inaccuracies and inconsistencies within datasets., rendering them fit for analytical purposes. This practice empowers data experts to gain a more lucid insight into their organizations, furnish reliable analytics accessible to all users, and enhance operational efficiency within their organizations, while examining our data missing values present in any records during the preprocessing stage of this work. While processing the data, attention was focused on the 'specializations' variable due to its numerous entries. It became imperative to categorize these specializations into distinct groups for a more refined analysis and to eliminate data dispersion.

3.4.2 Selection of Data Fields

The dataset being extracted may contain irrelevant attributes, potentially impacting the extraction algorithm's performance negatively. Moreover, certain algorithms might not function optimally when dealing with a substantial number of attributes. Hence, it is crucial to implement the feature selection technique before deploying the extraction algorithm. Feature

selection aims to enhance model performance and prevent overfitting. During this phase, the feature testing technique was utilized, involving the gradual removal of features, and assessing their influence on the model's performance. The results revealed that all features are indispensable for achieving high classification performance, as removing any of them results in decreased accuracy.

3.4.3 Data Transformation

Following the aforementioned data preprocessing steps, the prepared data was imported into Anaconda Python to build models. To ensure compatibility with Python, the data file was transformed excel format.

3.4.4 Encoding Categorical Features

Encoding categorical features is a crucial critical step in preparing data for machine learning and statistical analysis. Categorical features are those features that represent categories, labels, or groups rather than continuous numerical values. To in work order with to process categorical data effectively efficiently, it must be transformed converted into a numerical format, as most machine learning algorithms and statistical techniques require numerical numeric input, this process is known called as encoding, and it involves mapping categorical values to numerical values. Several common methods exist for encoding categorical features. In this dataset, we will employ ordinal encoding, a suitable approach for ordered data in which categories possess a clear rank order. Ordinal encoding assigns numerical values to categories based on their inherent order, with the numerical values reflecting the relative rankings of the categories.

3.4.5 Data Normalization

Generally, feature scaling plays a crucial role in optimizing variable subset selection and improving the performance of machine learning models, especially when dealing with original data that exhibits varying ranges, the emphasis on larger variable values can lead to complex and distorted models. To address this issue, we implemented data normalization, which is a pre-processing technique used to standardize the data, ensuring an equitable contribution from each feature in machine learning. It involves scaling or transforming the data to a consistent scale or range. Implementing this approach notably boosts the precision and effectiveness of data analysis and machine learning algorithms [45].In order to transform the data onto a scale ranging from 0 to 1. This transformation is represented mathematically in equation (3.1) [46]:

$$X' = \frac{Xi - Xmin}{Xmax - Xmin} \tag{3.1}$$

X' : Normalized attribute

Xmin : The minimum value of the attribute

Xmax : The maximum value of the attribute

3.4.7 Data Splitting

Upon cleaning and processing the collected data, this study delved into the analysis of five prominent AI tools to determine the tool that exhibited the highest accuracy. This comparative assessment serves as a potential foundation for a national tool to assess the practical situation of graduates in Palestine. The study evaluated RF, DT, XGBOOST Classifier, KNN, AdaBoost classifier. Further explanation on these five AI tools is provided in the subsequent section.

The analysis was divided into 80% for training data and 20% for testing Data. Training data constitutes the largest portion of the dataset and is used to train machine learning models, the model learns from this data and adapts its parameters to make classification. Test data represents another independent subset that remains untouched during the model's training and hyperparameter optimization. Once the model is trained, its performance on test data is evaluated to gauge its ability to generalize to new, unseen data. This assessment provides a realistic measure of model performance.

3.5 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) involves summarizing data by highlighting their key characteristics and representing them visually [47]. It encompasses a broader overview of the data. Additionally, EDA entails swiftly outlining fundamental dataset details like the number of rows/columns, identifying missing data, understanding data types, and providing previews. It involves cleaning corrupted data, addressing missing and invalid data, rectifying incorrect values, and visualizing data distributions through methods like bar charts, histograms, and box plots.

3.5.1 Descriptive Statistics

The findings indicate that males comprise 44.2% of the sample of graduates from the LF, while females represent an equal percentage of 55.8% as in Figure 3.2

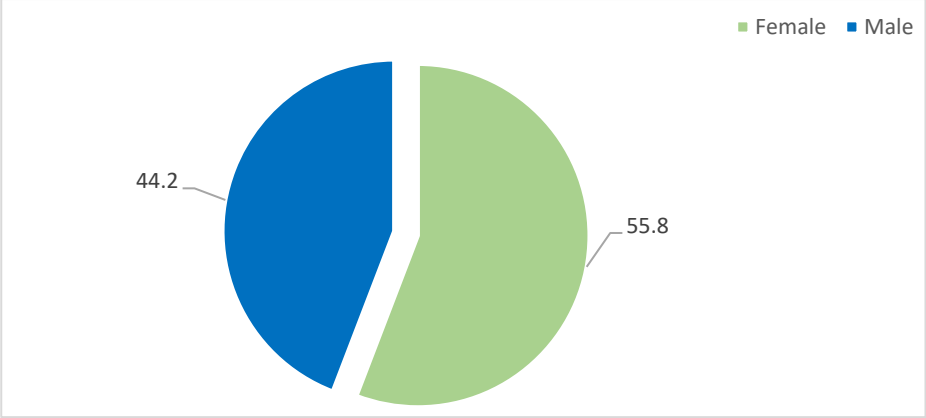


Figure 3.2 Distribution of participants by gender

Figure 3.3 illustrates that the highest labor force participation among gradations in this study was among individuals aged 27 years by 16.4%, meanwhile, for individuals aged 28 years, it was slightly lower at 15.1%. Additionally, 4.0% of participants were under 21 years old.

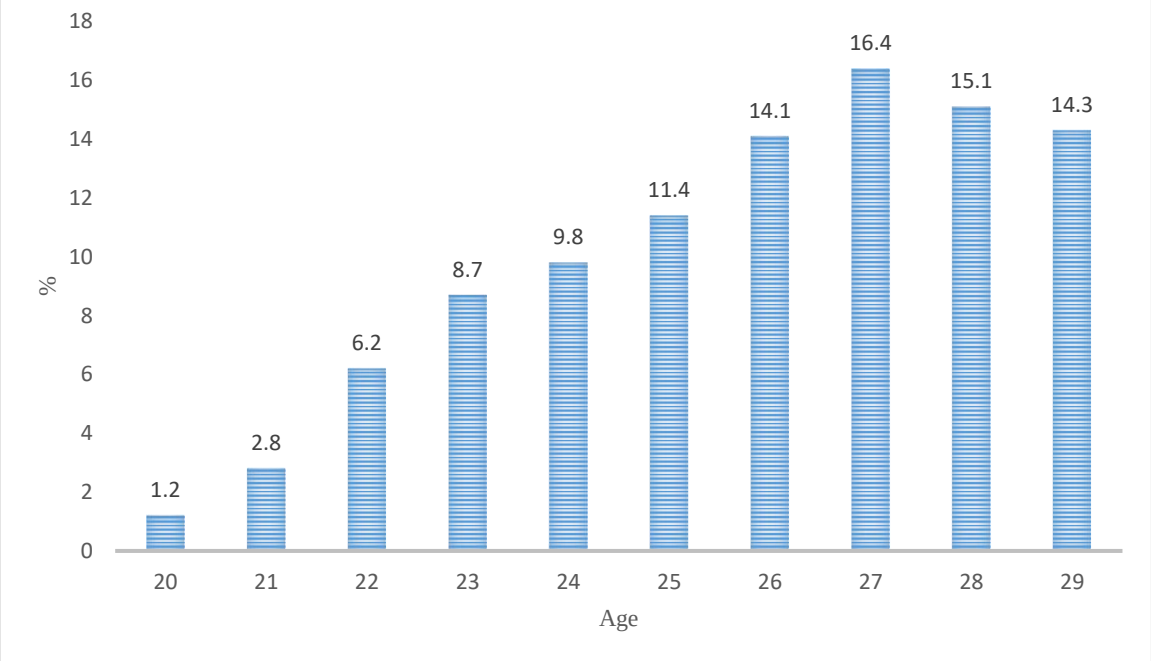


Figure 3.3 Distribution of participants by age

The data indicates in Figure 3.4 that over 65.7% of participants hold a bachelor's degree, while 21.8% possess an intermediate diploma. Conversely, less than 12.5% of individuals have either a higher diploma or a master's degree.

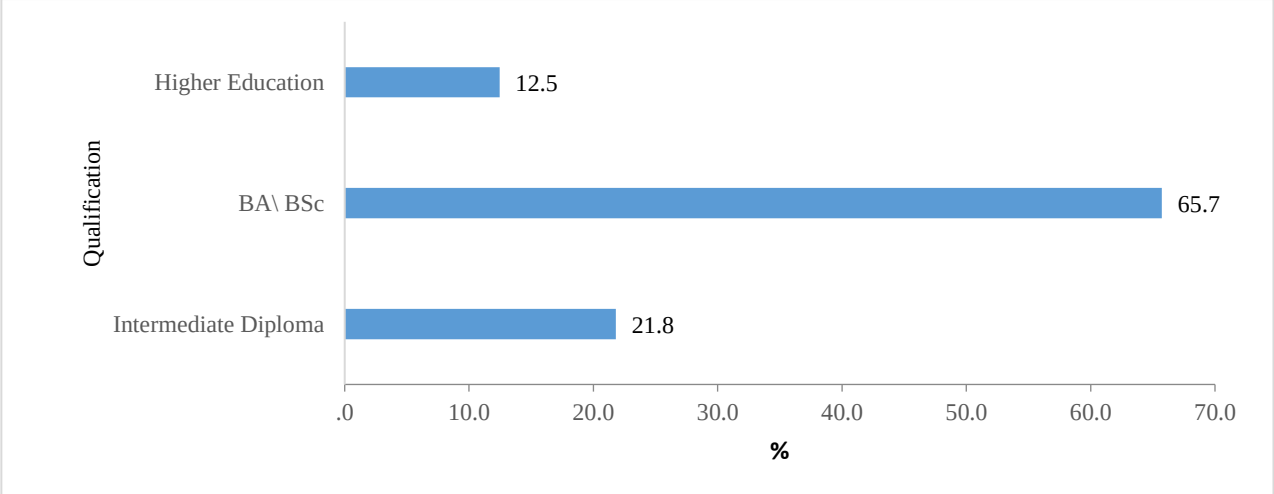


Figure 3.4 Distribution of participants by qualification

According to Figure 3.5, it illustrates that among the participants that 32.1% have Business and Administration and Law field as major academic field, while the field ‘health and welfare' constitutes 14.6% of the participants, and as field ‘Education’ for 11.7% of the participants.

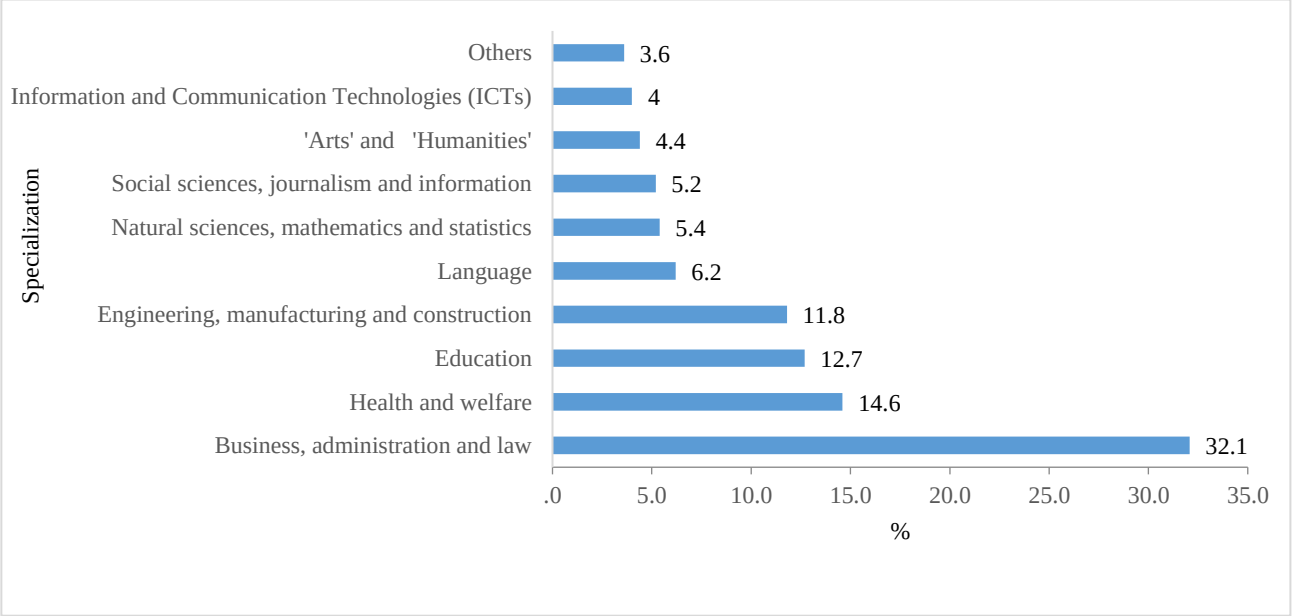


Figure 3.5 Distribution of participants by major academic (specialization)

As region, data indicates that 58.4% of participants from the West Bank, while 41.6% from Gaza Strip as in Figure 3.6

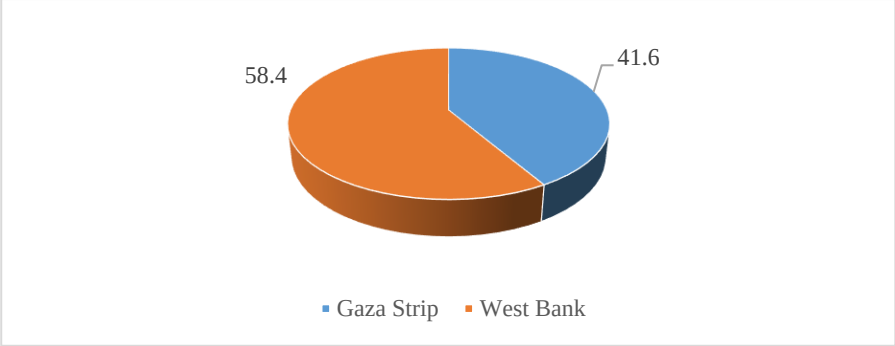


Figure 3.6 Distribution of participants by region

The Figure 3.7 illustrates that only 6.6 % of participant's attendance training course (participation in training courses overseen by the Ministry of Labor or the Qalandia Institute is required, and a certificate must be obtained upon completion of the training program)

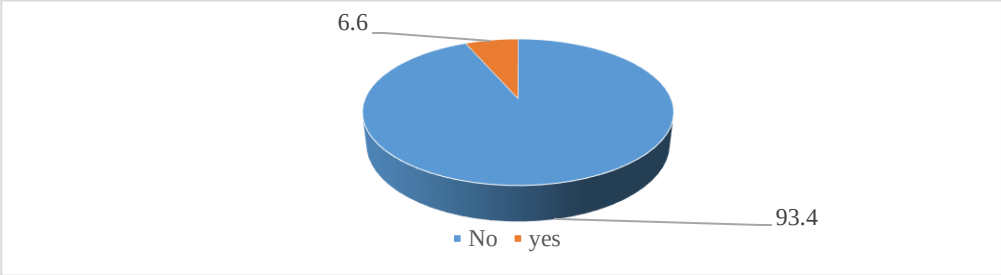


Figure 3.7 Distribution of participants by attendance of training course

Among youth graduates aged 20 to 29 holding an intermediate diploma or higher, the unemployment rate stands at 55.1%, while the employment rate corresponds to 44.9% as in Figure 3.8.

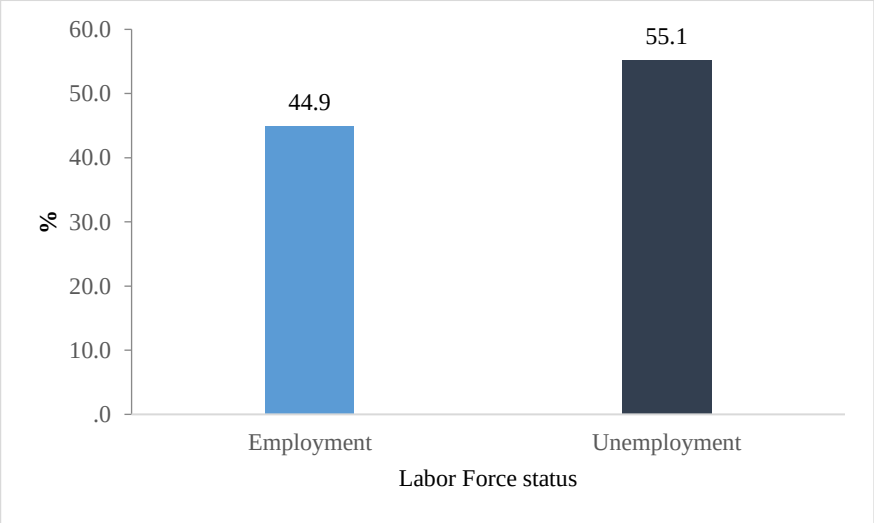


Figure 3.8 Distribution of participants by labor force status

From Figure 3.9, it is evident that the fields of Language, Education, Social Science, and ICT exhibit the highest unemployment rates, reaching 72.2%, 70.3%, 61.5%, and 57.4% respectively.

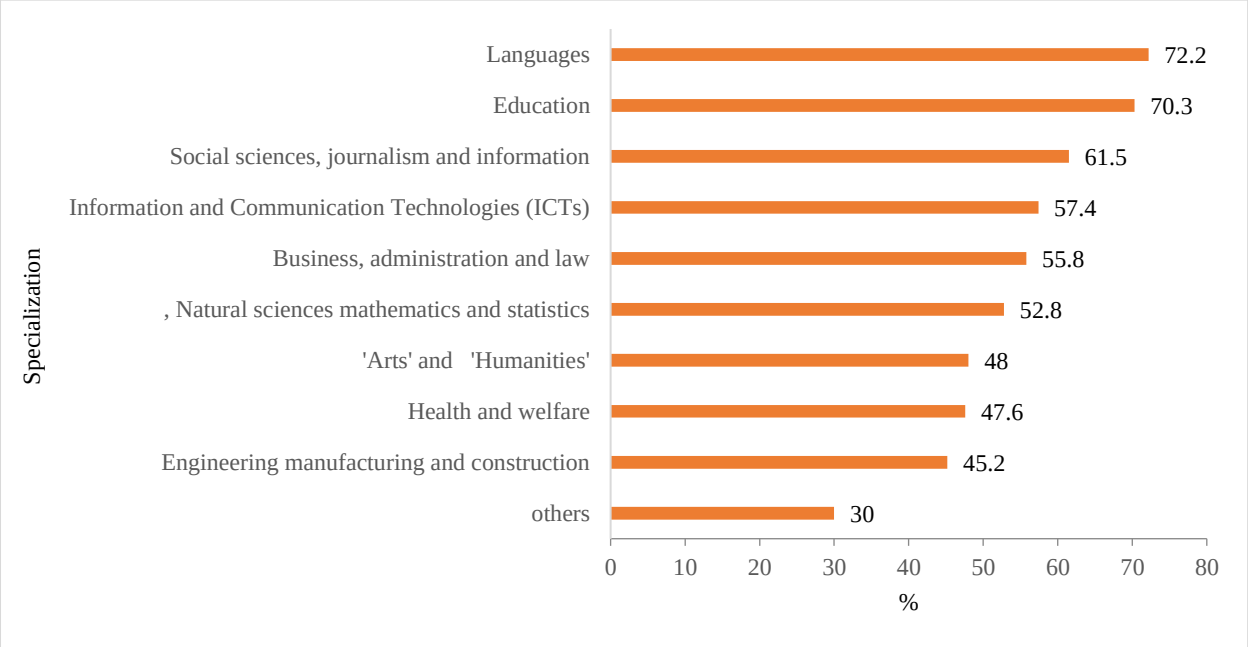


Figure 3.9 The unemployment rate among Palestinian individuals aged 20 to 29, holding either an intermediate diploma, bachelor's, or advanced degree, categorized by their field of education

Despite these notably high rates of unemployment indicated in Figure 3.9 ,depict a substantial influx of new students, graduate students, and enrolled students at the university. For instance, within the 'Education' as major university field, there are over 28,000 new students in the education field alone as in Figure 3.10.

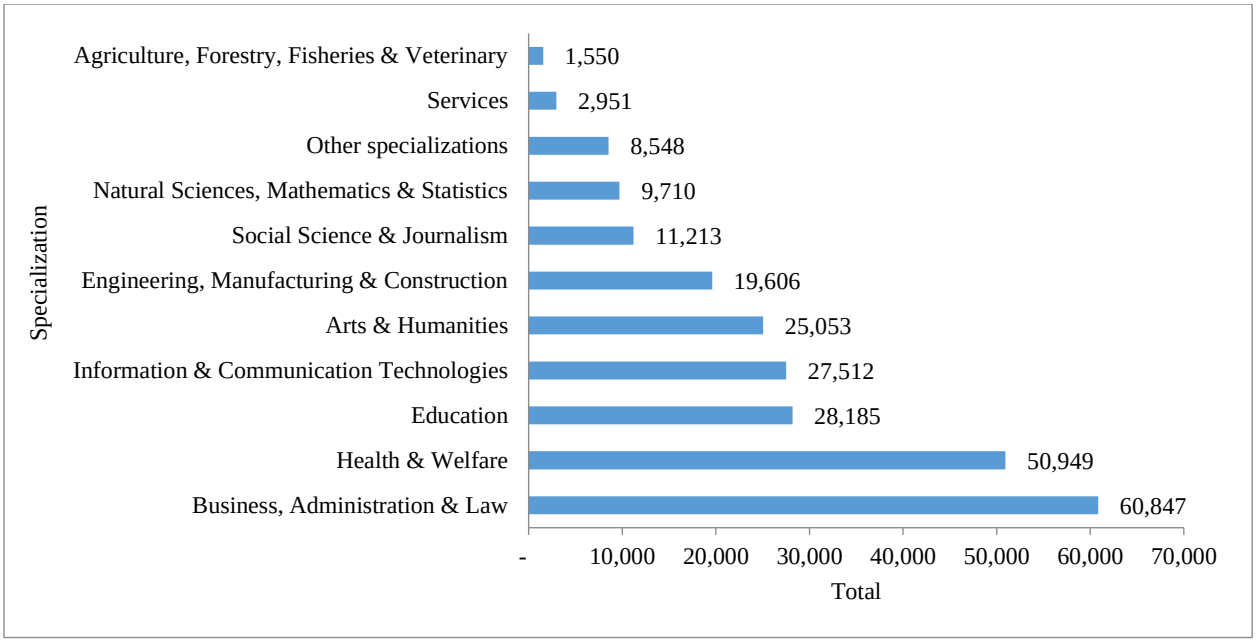


Figure 3.10 Total number of new students in Palestinian higher education institutions by specialization, 2019/2020- 2022/2023 [5]

According to Figure 3.11, there are 32,000 graduates in education and over 51,000 graduates in business, administration, and law.

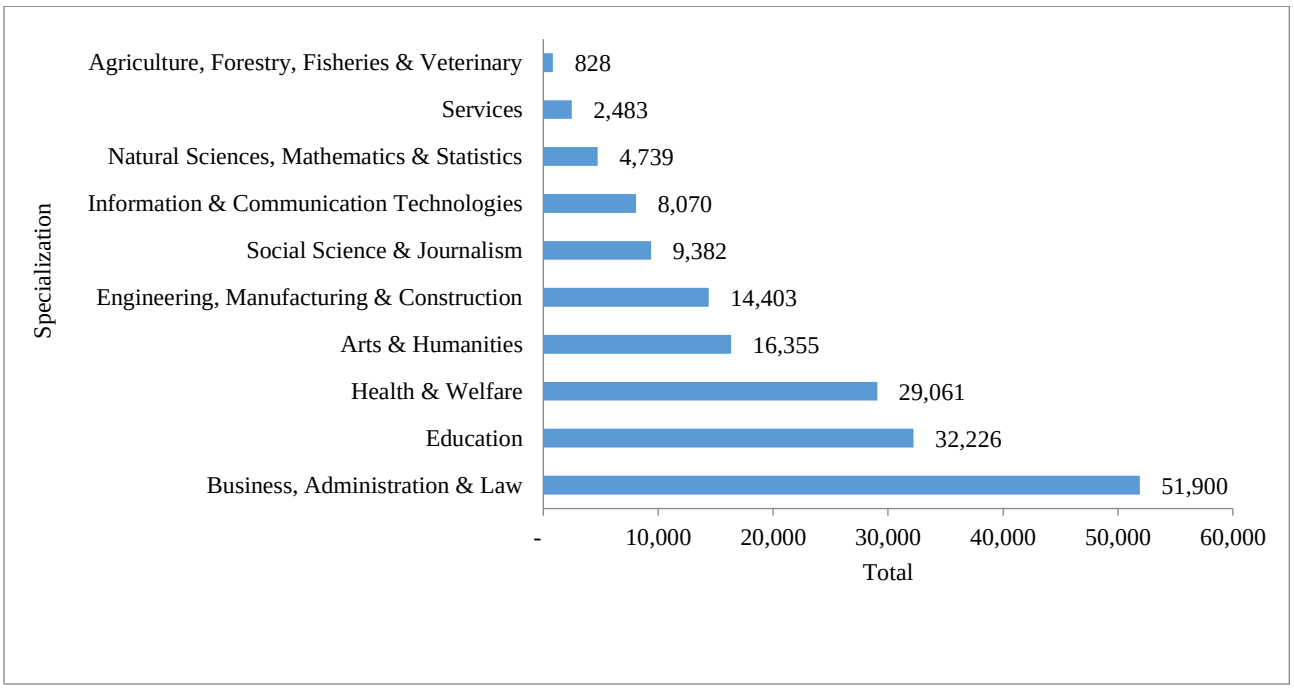


Figure 3.11 Total number of graduates (intermediate diploma or higher) from Palestinian Higher Education Institutions by specialization, 2019/2020- 2022/2023 [5]

In contrast to the high unemployment rates, university enrollment in education programs has surged to over 140,000 students, as highlighted by Figures 3.12.

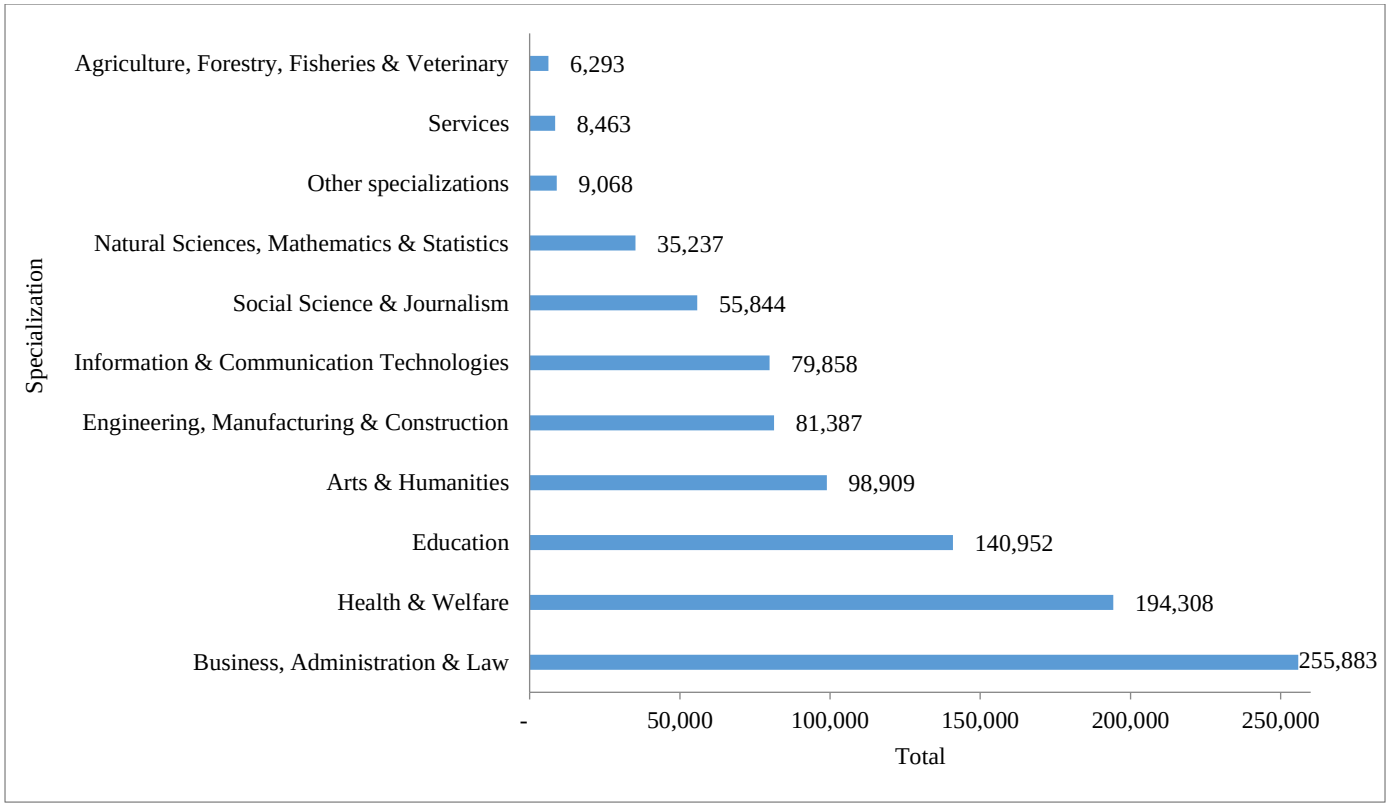


Figure 3.12 Total number of students enrolled in Palestinian Higher Education Institutions by specialization, 2019/2020- 2022/2023[5]

According to refugee status we can see that 60.3% of the graduates they work are non-refugees, while 39.7% are registered as in Figure 3.13.

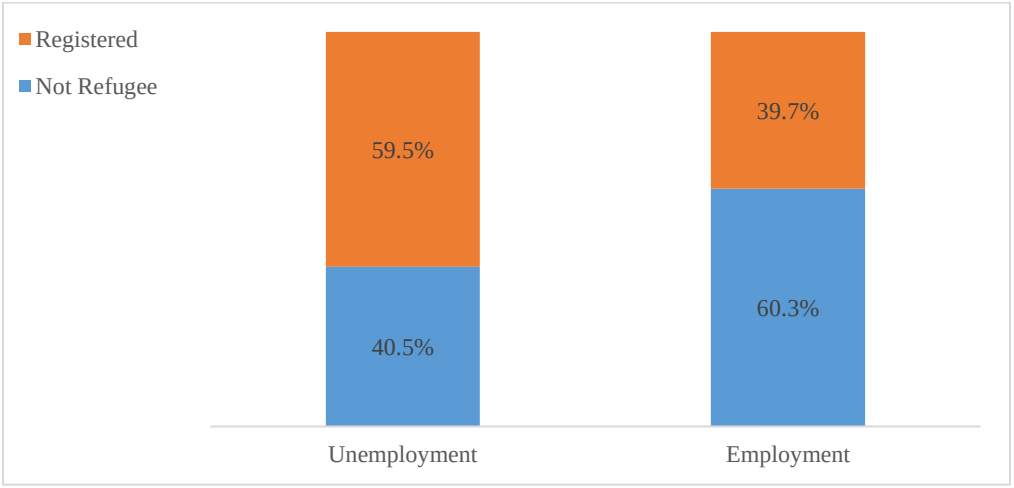


Figure 3.13 Distribution of labor force among graduates by refugee status

When categorized by gender, it's evident that 64.5% of unemployed individuals are female graduates as shown in Figure 3.14.

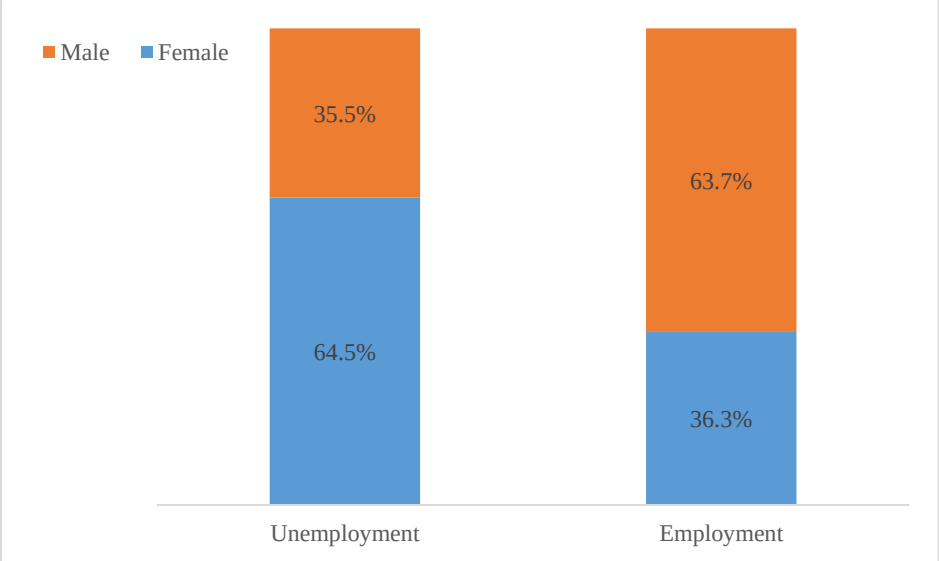


Figure 3.14 Distribution of labor force by gender

3.6 Machine Learning Algorithms for Classification

During this stage diverse machine learning techniques where applied to predict labor status as a binary independent variable: employment, unemployment; specifically: RF, DT, XGBoost, AdaBoost,and KNN. To boost the performance of these models in classification, a fine-tuning step is implemented, employing techniques such as grid search, aiming to achieve more precise and accurate results. Further information regarding these algorithms is introduced below.

3.6.1 Random Forest

The Random Forest machine learning algorithm is utilized as a supervised learning method for performing regression and classification tasks [48]. It relies on a multitude of decision trees. The term "random" pertains to two key principles. First, it randomly selects training data points to construct the trees. Second, it employs random feature subsets when making node splits. During training, every tree in the Random Forest learns from a randomly selected subset of the data, as depicted in Figure 3.15 [49] .Predictions are made by averaging the outputs from each individual tree. For the Random Forest to mimic human-like decision-making, it necessitates a substantial amount of training data. Once it has learned from this data, it can process information and classify new data into predefined categories [50]. It is recognized as one of the most successful machine learning methods due to its numerous advantages, including suitability for large datasets with fast model installation and evaluation, robustness in handling outliers, capability to address simple

linear and complex nonlinear relationships, and its ability to achieve competitive prediction accuracy in high-dimensional data [51]. In addition RF stands out as a remarkably convenient algorithm, primarily due to its default hyperparameters, which frequently yield accurate predictions. Comprehending these hyperparameters is relatively straightforward, considering their limited number [52].

Overfitting poses a significant challenge in machine learning; however, the random forest classifier effectively mitigates this concern. With a sufficient number of trees in the forest, overfitting becomes unlikely, ensuring the model's robustness[48].

These distinctive advantages have garnered increasing interest in Random Forest across various research domains.

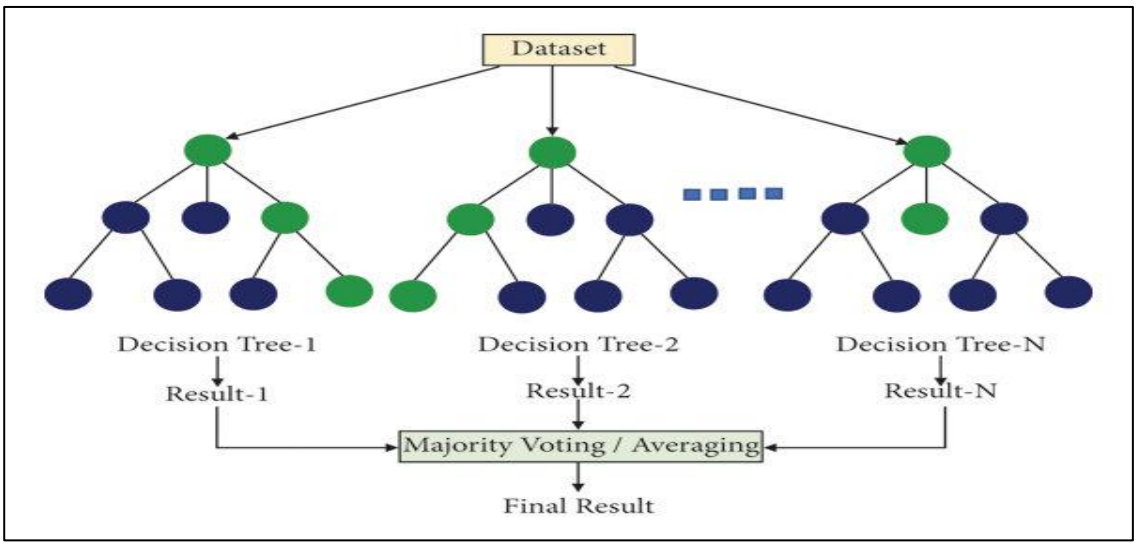


Figure 3.15 Structure of random forest classification algorithm [51]

3.6.2 Decision Tree

A decision tree is a supervised algorithm employed for regression and classification tasks. It serves as a decision support tool, having a tree-like structure. DT aim to construct a model capable of predicting or classifying the target variable by learning rules derived from the attributes present in the data, driven by a predefined objective [53]. Decision Trees can be divided into three main node types: the root node, internal nodes (also known as decision nodes), and leaf nodes. The root node represents the starting point with no incoming edges, while internal nodes correspond to variables with both incoming and outgoing edges. Leaf nodes, on the other hand, serve as the final decision points, and they

can be associated with classes for classification or numerical values for regression tasks as in Figure 3.16.

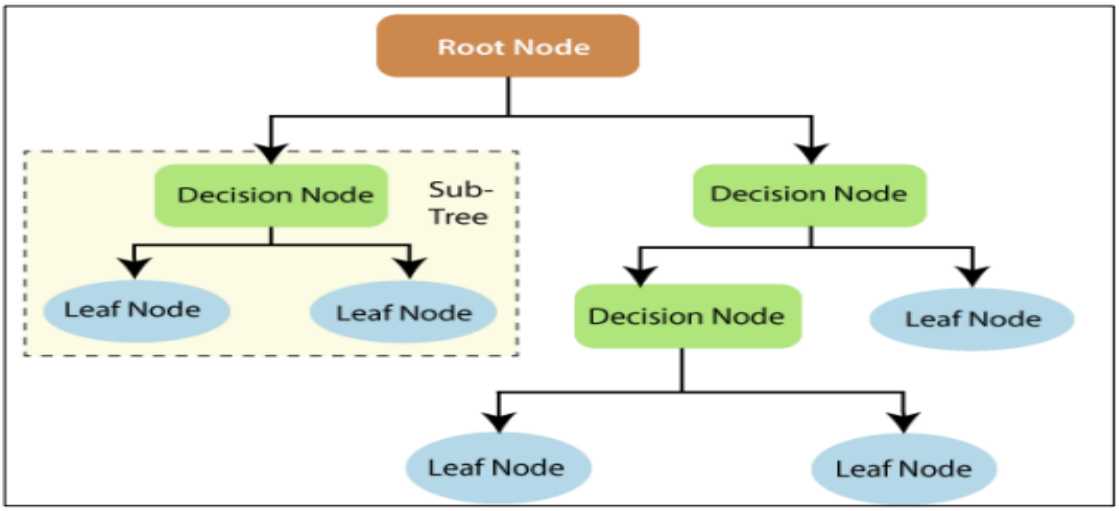


Figure 3.16 Structure of decision tree [54]

3.6.3 K- Nearest Neighbor

KNN represents a supervised machine learning algorithm. It's a memory-based method that classifies objects by examining their closest features [55], and KNN is known for its simplicity and widespread use as a classifier [56] its classification of new data points is rooted in the measurement of distances between the data and its nearest data points. It is classified by finding the nearest neighbor and applying Euclidean distance and cosine to distinguish between records in training and testing [57].

Typically, a set number of points, represented as K, is utilized, however, Choosing the right value for K is pivotal as it notably impacts the ultimate output[58]. Various methods can be employed to determine K, with one approach being to use the square root of 'n,' where 'n' represents the total number of points as shown in Figure 3.17.

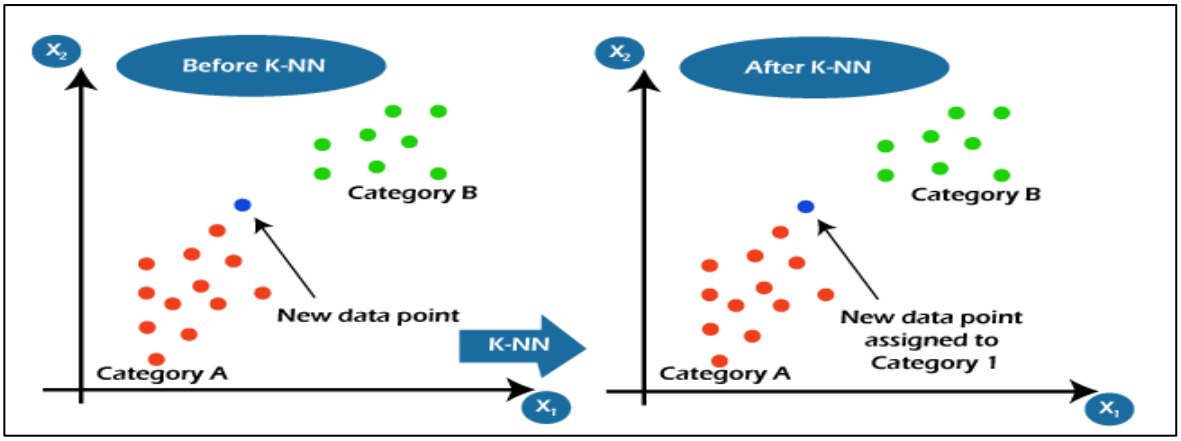


Figure 3.17 Structure of KNN [59]

3.6.4 Adaptive Boosting Classifier

AdaBoost, short for Adaptive Boosting, is a robust and widely employed algorithm in machine learning, particularly for classification tasks [60]. It is a supervised machine learning method proficient in effectively handling both classification and regression problem statements, that stands out as an exceptional Boosting algorithm renowned for its robust theoretical foundation and remarkable achievements in real-world scenarios. Its capability to elevate a weak learning algorithm marginally more accurate than random guessing—into a highly precise strong learning algorithm. This approach introduces a fresh methodology and innovative design concept to the realm of learning algorithm development [61].

3.6.5 Extreme Gradient Boosting

XGBoost, short for extreme Gradient Boosting, stands out as an optimized distributed gradient boosting system. It utilizes an iterative decision tree algorithm featuring multiple decision trees. Each tree learns from the residuals of preceding trees. Unlike Random Forest, which adopts majority voting for output results, XGBoost calculates its predicted output as the summation of all results, XGBoost consistently delivers state-of-the-art results in various machine learning challenges [62]. Its primary emphasis lies in achieving both computational speed and superior model performance. This non-parametric supervised learning model caters to regression and classification tasks and is specifically designed to handle large and intricate datasets [63].

3.7 Model Evaluation

To assess the effectiveness of the labor status classification model, a variety of evaluation techniques were employed. These included traditional metrics like confusion

matrix, accuracy, sensitivity, specificity, and F-score, as well as ROC analysis and calculation of the AUC. Experts will subsequently analyze and compare the performance of different models to determine the most promising ones.

3.7.1 Confusion Matrix

The confusion matrix serves as a vital tool in predictive analysis within machine learning. It is used to assess the performance of classification-based machine learning models, also provides a concise summary table indicating the count of both accurate and inaccurate predictions made by a classifier (or classification model) specifically designed for binary classification tasks as in Figure 3.18.

	Actual: Yes	**Actual: NO**
Predicted: Yes	True Positives (TP)	False Positives (FP)
Predicted: NO	False Negatives (FN)	True Negatives (TN)

Figure 3.18 Confusion matrix [64]

where:

- TN stands for the count of negative cases accurately classified.
- TP refers to the count of positive cases accurately classified.
- FN represents the count of positive cases inaccurately classified as negative.
- FP indicates the count of negative cases inaccurately classified as positive.

3.7.2 Accuracy

Accuracy gauges the number of correct predictions the model generated for the entire test dataset [61]. It serves as a fundamental metric, offering a reliable measure of the model's performance, where the mathematical representation as in equation (3.2):

$$\text{Accurcy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{3.2}$$

3.7.3 Precision

Precision serves as the metric employed to determine the accuracy of classification, it becomes particularly valuable in scenarios where False Positives hold greater significance than False Negatives, it is defined as follows in equation (3.3) [64]:

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})} \tag{3.3}$$

3.7.4 Recall

Recall indicates the proportion of actual positive cases that our model correctly predicted, the formula as flowing in equation (3.4) [64]:

$$\text{Recall} = \frac{(TP)}{(TP + FN)} \tag{3.4}$$

3.7.5 F1-Score

The F1-Score integrates both Precision and Recall, achieving its highest value when Precision equals Recall, the formula as in equation (3.5) [64]:

$$\text{F1 Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{3.5}$$

3.7.2 Roc and AUC Curve

The ROC curve illustrates the relationship between the true positive rate (recall) and the false positive rate. Whereas the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a powerful metric used to evaluate the performance of binary classification models. It provides a single, comprehensive measure of how well a model distinguishes between positive and negative cases across all possible classification thresholds [65] as in Figure 3.18.

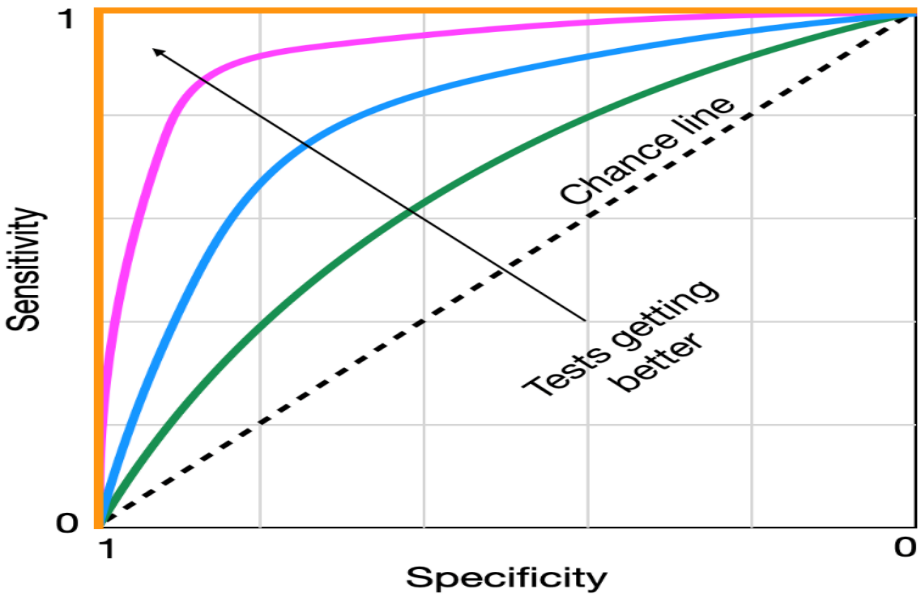


Figure 3.19 Roc and AUC curve [65]

3.8 Conclusion

Throughout this chapter, we examined the data intended for analysis in this research. The data was visualized using graphics, and both the dependent and independent variables were explored. Additionally, we delved into reviewing the key features of the algorithms that will be employed in classifying the data. Following this chapter, I will discuss the previously mentioned proposed models, which aim to classify graduates as employed or unemployed.

Chapter Four: Results and Discussion

4.1 Introduction

This chapter focuses on the construction of models and the experimental process conducted within the proposed framework to classify data based on 16 distinct features related to a person's employment status. This experiment and evaluation aim to demonstrate the practical implementation of the architecture introduced in chapter three and highlight selected machine learning models, including RF, DT, XGBoost, KNN, and AdaBoost, and this will be the first phase.

In the second phase, the focus shifts to classifying graduates into employment or unemployment (non-working) according to their overall major. To achieve this, three machine learning models-RF, DT, and KNN- will be employed. The analysis will specifically target data from four chosen majors: Business, Administration & Law; Engineering, Manufacturing & Construction; Education; and Health & Welfare.

The experiment entails interpreting the results obtained from the application of these models. Additionally, this section aims to present the outcomes of all machine learning models employed in this research, encompassing accuracy, precision, recall, f1 score, and AUC-ROC metrics.

This entire experiment and analysis are conducted using Anaconda Python programming and the Orange program for data mining [66].

4.2 Random Forest Model

In this section, to predict labor status (employed or unemployed), we constructed a classification model using the RF algorithm by using 15 independent variables. This algorithm excels at handling multiple independent variables and a binary dependent variable (0 for unemployed, 1 for employed). Python libraries associated with RF, were utilized in the construction of this model.

Hyper parameters applied to the RF, and selected best estimator and was used in training the algorithm, which included the Gini for criterion, while the max_depth,min_samples_leaf, min_samples_split, and n_estimators were (100),(1),(2),and (2000) respectively. The algorithm with these parameters has an accuracy of 95.9% using five-fold cross-validation.

4.2.1 Evaluating Model Performance

The RF model achieved an accuracy of 95.9%, indicating a highly commendable level of accuracy in its performance as in table below:

Table 4.1 Accuracy and classification report by random forest

Random Forest Classifier –Test accuracy 0.959				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.95**	**0.97**	**0.96**	**1324**
Employment	**0.97**	**0.94**	**0.95**	**1066**
Accuracy			**0.96**	**2390**
Macro avg	**0.96**	**0.96**	**0.96**	**2390**
Weighted avg	**0.96**	**0.96**	**0.96**	**2390**

From the classification table above, precision signifies the model's accuracy in making classifications. In our scenario, the precision values were 95% for unemployment and 97% for employment. Meanwhile, the recall evaluation metric gauges the model's capability to detect positive samples, resulting in 97% for unemployment and 94% for employment.

The RF model performed impressively, with outstanding prediction f1-score for both unemployment (96%) and employment (95%). This high performance is further underscored by the AUC-ROC curve which visually demonstrates the model's ability to distinguish between the two classes. The AUC of 0.99% signifies exceptional discrimination power, placing this model among the top performers in its category as in Figure 4.1.

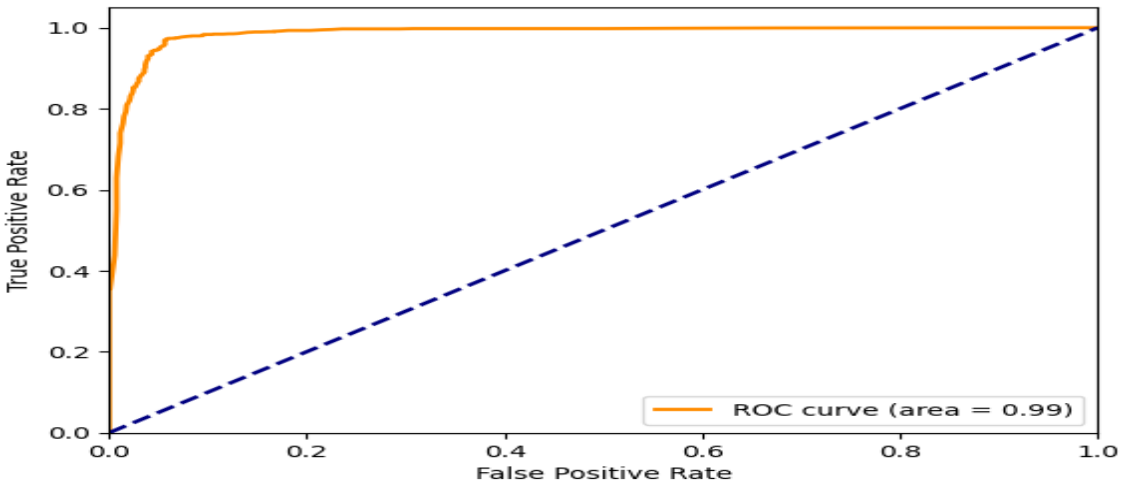


Figure 4.1 The AUC-ROC for random forest

4.2.2 Confusion Matrix

The confusion matrix, displayed below, captures the distribution of correct and incorrect predictions made by the model, this table highlights four key combinations of actual and predicted values:

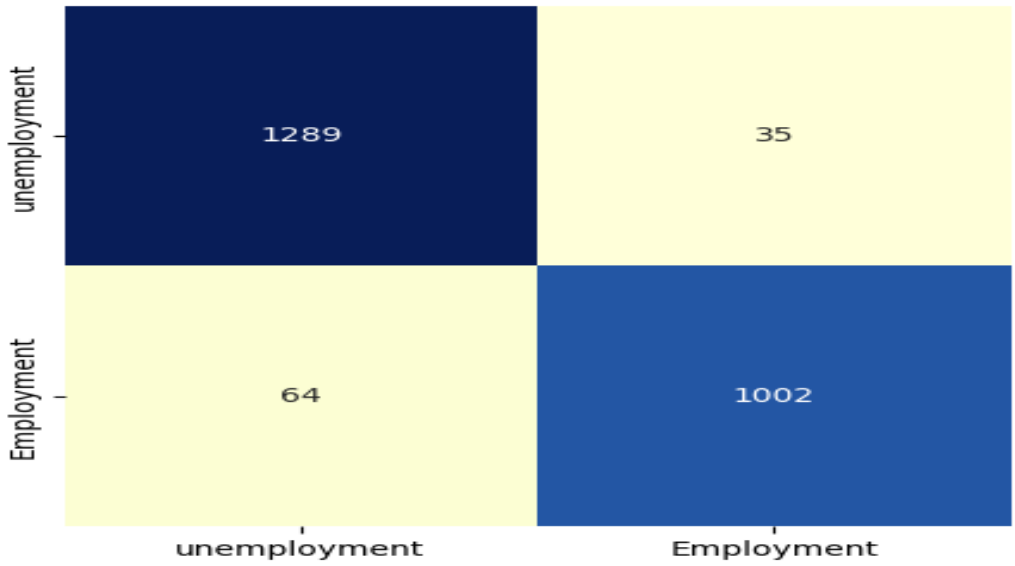


Figure 4.2 Confusion matrix result of random forest

Based on Figure 4.2, among a total of 2,390 individuals in the test dataset, 1,066 individuals are categorized as employed, while 1,324 individuals are classified as unemployed. Examining the confusion matrix results, it's evident that in the first row representing the unemployment class, 1,289 individuals are accurately classified as unemployed, but 35 individuals from the unemployment class are mistakenly classified as employment. Moving to the second row, 1,002 individuals are correctly classified as employed, while 64 individuals from the employment class are incorrectly labeled as unemployment.

4.3 Decision Tree

In this section, we elaborate on the performance evaluation and analysis of the machine learning approach applied to the labor force classification model using DT, by using the all of independent variables that we select it. In addition to optimize the performance, a hyper parameter tuning process was conducted for the DT model. This involved Gini for Criterion, while the max_depth, min_samples_leaf ,min_samples_split were 30,10 and 2 respectively, the DT model achieved an accuracy of 93.3% using 5-fold cross-validation beside that we focus on key metrics such as accuracy, the classification report encompassing precision, recall, and f1-score, as well as the confusion matrix.

4.3.1 Evaluating Model Performance

The DT model achieved an accuracy of 93.3%, indicating a highly commendable level of accuracy in its performance.

Table 4.2 Accuracy and classification report by decision tree

Decision Tree Classifier –Test accuracy 0.933				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.94**	**0.94**	**0.94**	**1324**
Employment	**0.93**	**0.92**	**0.93**	**1066**
Accuracy			**0.93**	**2390**
Macro avg	**0.93**	**0.93**	**0.93**	**2390**
Weighted avg	**0.93**	**0.93**	**0.93**	**2390**

From the classification table above, the precision, recall and f1-score were the same a values 94% for unemployment while for employment was (93%) for precision and f1-score, while it was 92% for recall. In addition, the ROC curves have been plotted and are displayed in the figure below. The AUC has been calculated for various classes of the dependent variable. The computed area under the curve is 93% as per the ROC curve design as in Figure 4.3.

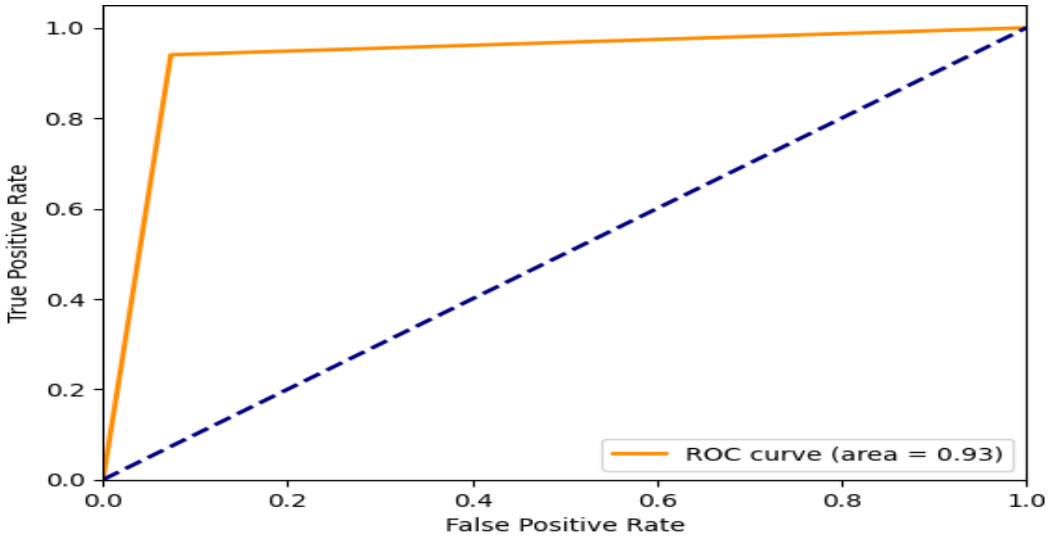


Figure 4.3 The AUC-ROC for decision tree

4.3.2 Confusion Matrix

The following confusion matrix showcases the distribution of accurate and inaccurate predictions performed by the model. It delineates four crucial combinations of actual and predicted values:

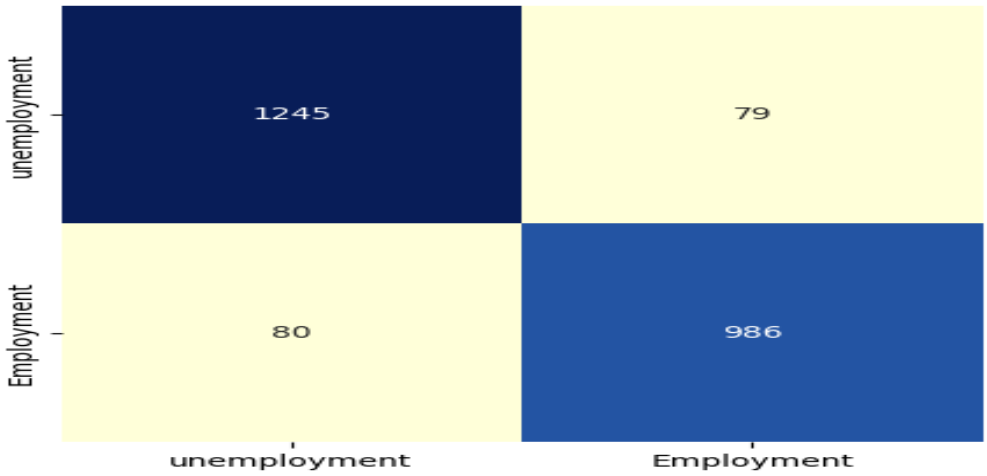


Figure 4.4 Confusion matrix result of decision tree

Figure 4.4 reveals the model's predictions for 2,390 test cases: 2,523 classified as employed and 1,324 as unemployed. A closer look at the confusion matrix shows accurate identification of 1,245 unemployed individuals, with only 79 misclassified as employed.

Moving to the second row, 986 employed instances are correctly classified, while 80 instances from the employed class are incorrectly predicted as unemployed.

4.4 XGBoost Classifier

To accurately classify labor status for graduates, we constructed a model using the XGBoost Classifier algorithm. The model leveraged 15 carefully selected independent variables to make the classification.

To optimize model performance, we employed Grid Search to identify the best hyperparameter settings. The optimal configuration, which yielded a 91.0% accuracy using 5-fold cross-validation, consisted of: Learning rate: (0.1), Max_depth: (3), N_estimators: (5000).

4.4.1 Evaluating Model Performance

The XGB Classifier model achieved an accuracy of 91%, indicating a highly commendable level of accuracy in its performance.

Table 4.3 Accuracy and classification report by XGBoost classifier

XGB Classifier –Test accuracy 0.91				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.90**	**0.94**	**0.92**	**1324**
Employment	**0.92**	**0.88**	**0.90**	**1066**
Accuracy			**0.91**	**2390**
Macro avg	**0.91**	**0.91**	**0.91**	**2390**
Weighted avg	**0.91**	**0.91**	**0.91**	**2390**

From the classification table above, the precision values were 90% for unemployment and 92% for employment. Meanwhile, the Recall evaluation metric resulting in 94% for unemployment and 88% for employment. While the F1-was 92% for unemployment and 90% for employment, beside that the Figure below showed that the computed area under the curve is 97% as per the ROC curve design as in Figure 4.5.

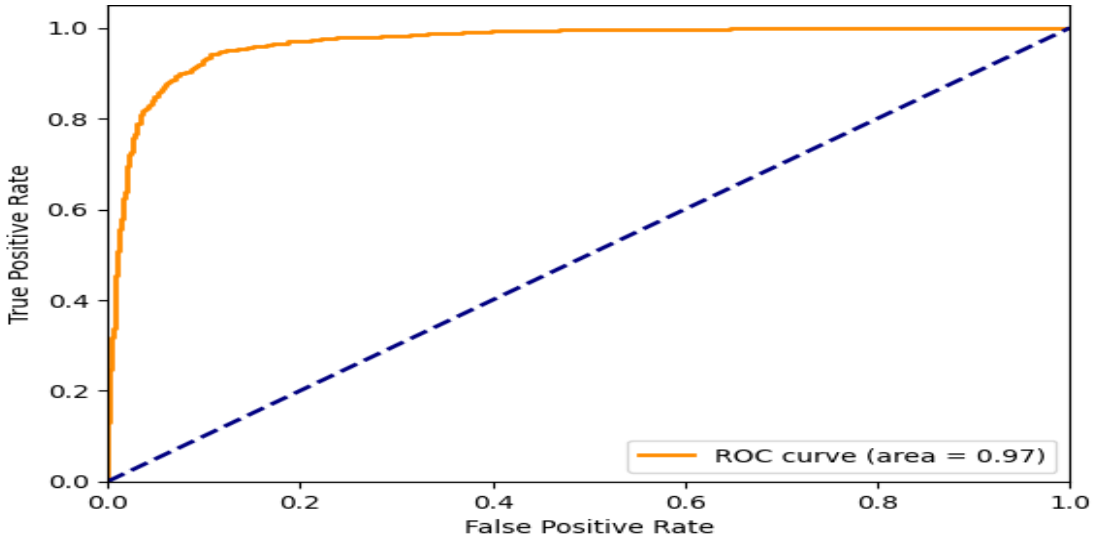


Figure 4.5 The AUC-ROC for XGBoost

4.4.2 Confusion Matrix

The resulting confusion matrix (presented in Figure 4.6) reveals patterns of both accurate and inaccurate predictions made by the model. This matrix outlines four significant combinations of real and predicted values, providing crucial insights into the accuracy of the model.

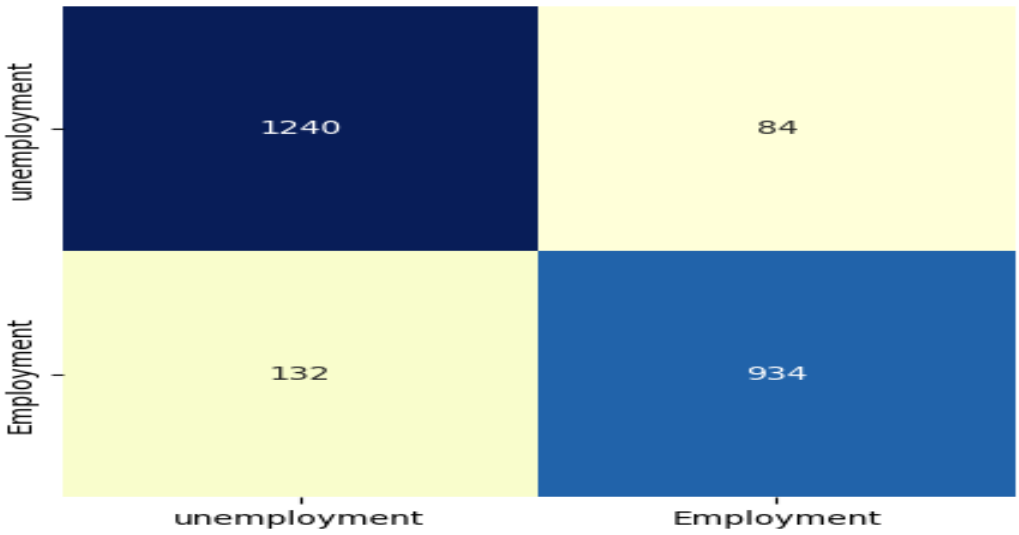


Figure 4.6 Confusion matrix result of XGBoost

In Figure 4.6, out of a total of 2,330 individuals within the test dataset, 1,066 individuals are identified as employed, while 1,324 individuals are categorized as unemployed. Upon analyzing the confusion matrix, the following observations can be made:

In the first row corresponding to the employed class, 1,240 individuals are correctly classified as unemployed, but 84 individuals from the unemployed class are erroneously identified as employed.

Shifting to the second row, 934 individuals are correctly classified as employed. However, 132 individuals from the employed class are inaccurately labeled as unemployed.

4.5 KNN

In this section, a labor status classification model was constructed utilizing the KNN algorithm. The KNN model was employed to classify the data based on a collection of independent variables, with the dependent variable restricted to two values (0 and 1).

4.5.1 Evaluating Model Performance

The KNN model shows a good performance with an 88% accuracy.

Table 4.4 Accuracy and classification report by KNN

KNN –Test accuracy 0.884				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.89**	**0.91**	**0.90**	**1324**
Employment	**0.88**	**0.85**	**0.87**	**1066**
Accuracy			**0.88**	**2390**
Macro avg	**0.88**	**0.88**	**0.88**	**2390**
Weighted avg	**0.88**	**0.88**	**0.88**	**2390**

From the classification table above, the precision values were 89% for unemployment, while 88% for employment, compared the Recall evaluation metric was 85% for employment and 91% for unemployment. In other side the f1-score values were 87% for employment and 90% for unemployment. On other side from the graph below the area under curve reached 94% as in Figure 4.7.

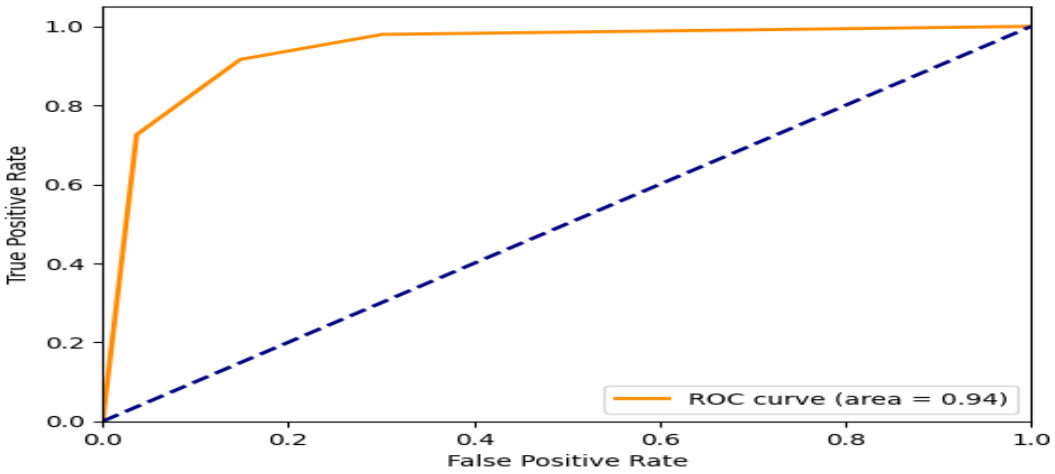


Figure 4.7 The AUC-ROC for KNN

4.5.2 Confusion Matrix

After running the above confusion matrix code, the following tabular form shown below consists of 4 different combinations of actual and predicted values.

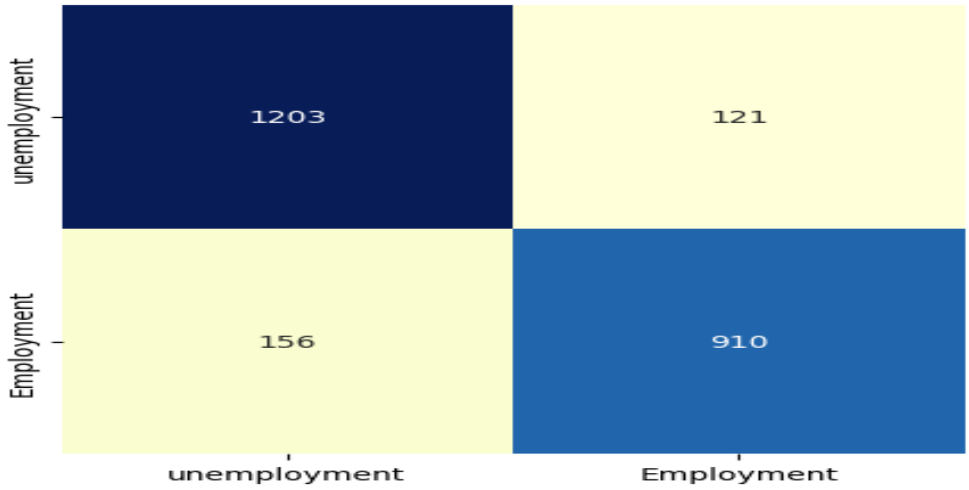


Figure 4.8 Confusion matrix result for KNN

Figure 4.8 reveals that in a test dataset of 2,330 individuals, 1,066 were categorized as employed and 1,324 as unemployed. Analyzing the confusion matrix. Examining the confusion matrix results, it's evident that in the first row representing the unemployed class, 1203 individuals are accurately classified as employed, but 121 individuals from the unemployed class are mistakenly classified as employment. Moving to the second row, 910 individuals are correctly classified as employment, while 156 individuals from the unemployment class are incorrectly labeled as unemployed.

Let's determine the optimal value for "k" to achieve the most accurate implementation of the KNN algorithm. This is accomplished by conducting a sweep across various values of "k". The outcomes are presented in Figure 4.9 below. Upon observing the figure, it's evident that the highest accuracy is achieved when *"k"* equals 3.

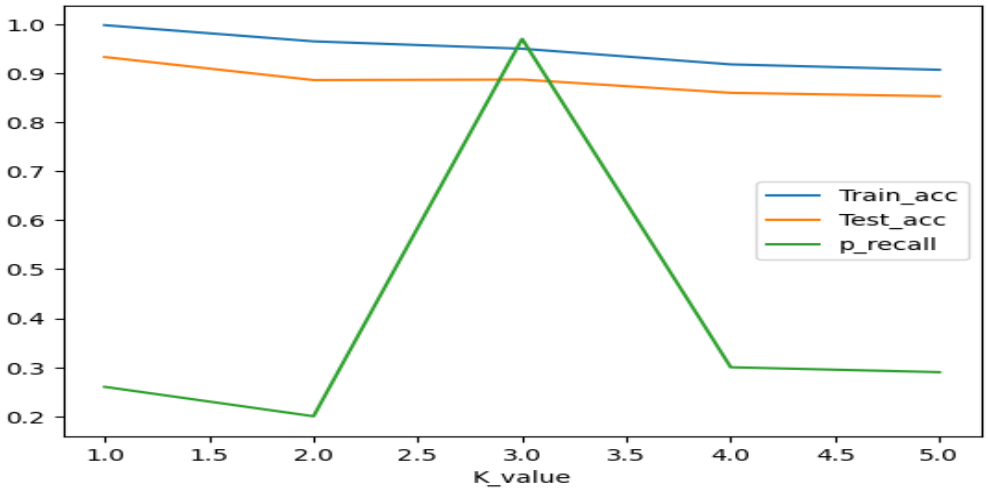


Figure 4.9 Testing and training accuracy with the positive recall comparison for KNN

4.6 ADABoost

In this section, a labor status classification model was constructed utilizing the ADABoost algorithm. The AdaBoost model was employed to classify the data based on a collection of 15 independent variables, with the dependent variable restricted to two values (0 and 1). To optimize model performance, we employed grid Search to identify the best hyperparameter settings. The optimal configuration, which yielded a 76.1% accuracy using 5-fold cross-validation, consisted of: n_features: (20), n_informative: (15), n_estimators: (1000)

4.6.1 Evaluating Model Performance

The AdaBoost model demonstrates a performance level achieving an accuracy of 76.1%.

Table 4.5 Accuracy and classification report by AdaBoost

AdaBosst Classifier- Test accuracy 0.761				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.79**	**0.77**	**0.78**	**1324**
Employment	**0.73**	**0.74**	**0.73**	**1066**
Accuracy			**0.76**	**2390**
Macro avg	**0.76**	**0.76**	**0.76**	**2390**
Weighted avg	**0.76**	**0.76**	**0.76**	**2390**

From the classification table above, the precision values were 79% for unemployment, while 73% for employment, compared the Recall evaluation metric was 74% for employment and 77% for unemployment. In other side the f1-score values were 73% for employment and 78% for unemployment. On other side from the graph below the area under curve reached 83% as shown in Figure 4.10

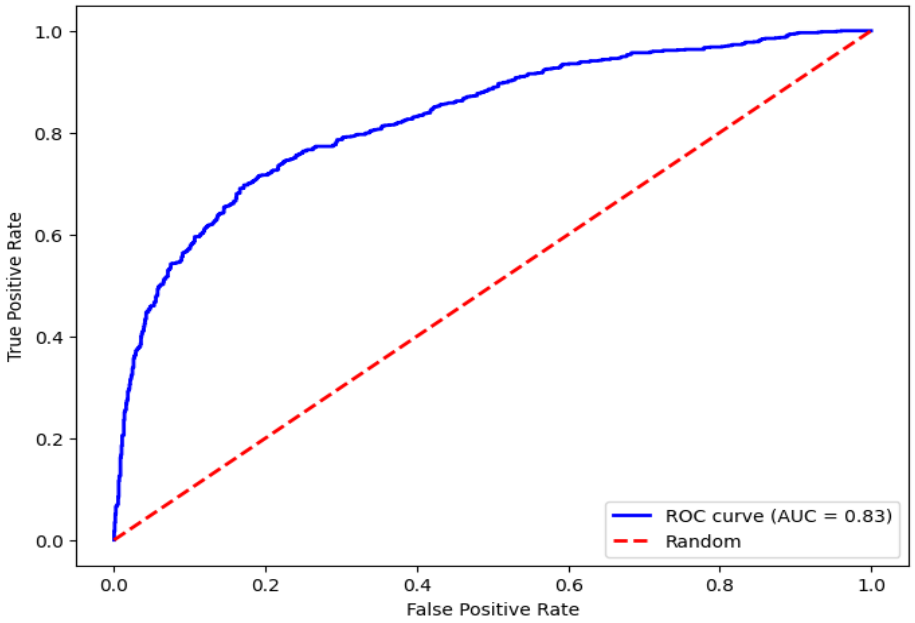


Figure 4.10 The AUC-ROC for AdaBoost

4.6.2 Confusion Matrix

The confusion matrix, displayed below Figure 4.11, captures the distribution of four crucial combinations of actual and predicted values.

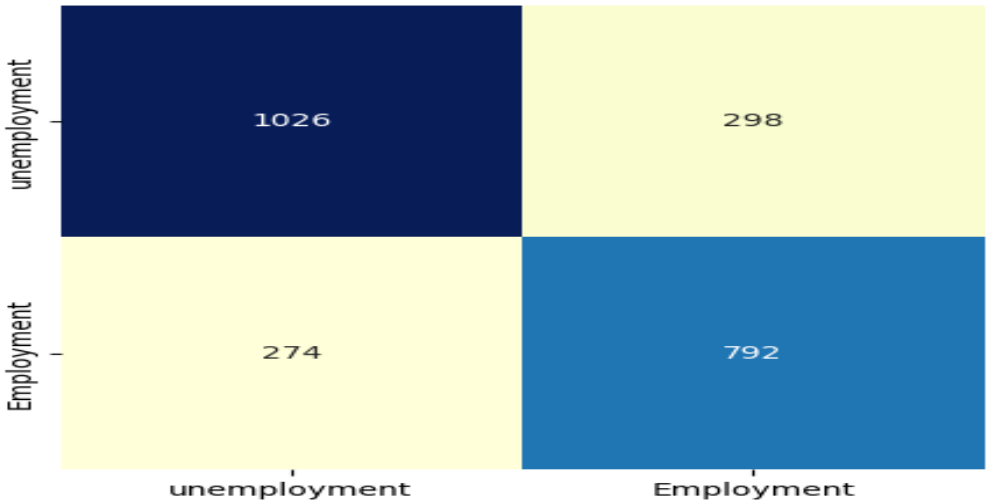


Figure 4.11 Confusion matrix result AdaBoost

Figure 4.11 reveals that in a test dataset of 2,330 individuals, 1,066 were categorized as employed and 1,324 as unemployed. Analyzing the confusion matrix. Examining the confusion matrix results, it's evident that in the first row representing the unemployed class, 1026 individuals are accurately classified as unemployed, but 298 individuals from the

unemployment class are mistakenly classified as employment. Moving to the second row, 792 individuals are correctly classified as employment, while 274 individuals from the employment class are incorrectly labeled as unemployed.

4.7 Evaluating Classification Algorithms

A comparative performance analysis was conducted among five machine learning algorithms, evaluating metrics such as accuracy, precision, recall, f1-score, and AUC Curve. According to this analysis, the RF classifier exhibited superior performance compared to Decision Tree, XGBoost, KNN, and AdaBoost models, as indicated in Figures 4.12 – 4.15.

In particular, Figure 4.12 depicted the accuracy of the proposed labor force status classification model using the five classifiers: RF, DT, XGBoost, KNN, and AdaBoost. Notably, Figure 4.12 illustrated that RF achieved the highest accuracy among all classifiers, with DT also displaying relatively strong accuracy compared to the other models.

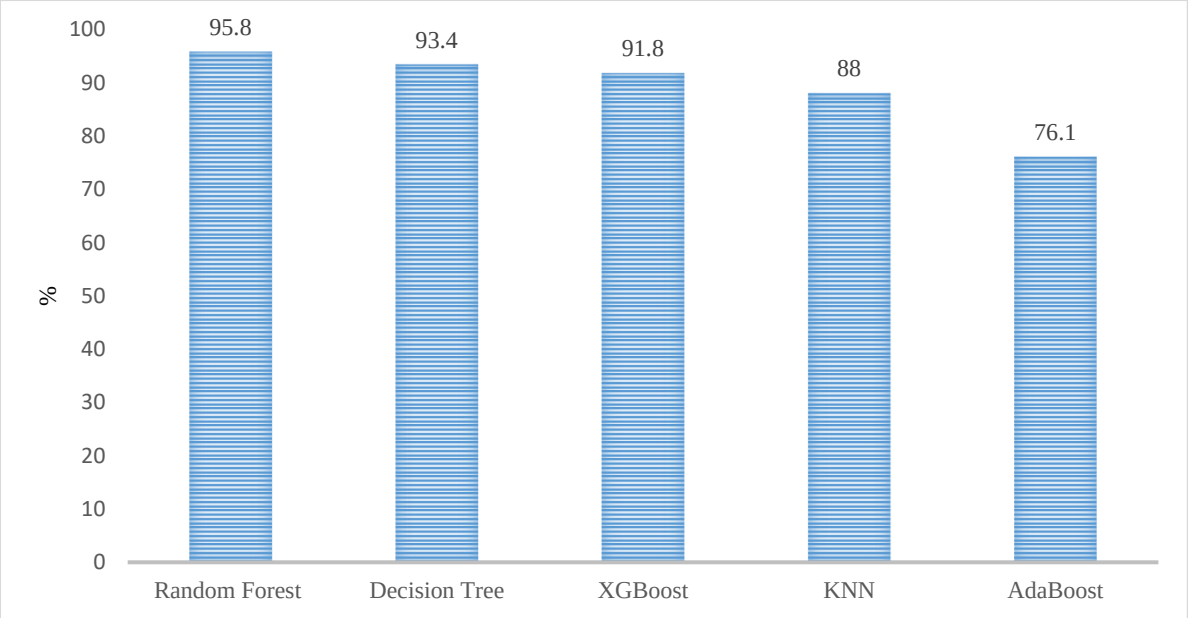


Figure 4.12 Accuracy of five classifiers

Figure 4.13 illustrates the precision of five machine learning models. Upon fine-tuning parameter values, it becomes evident that RF attains the highest precision, while the AdaBoost classifier exhibits the lowest precision specifically in predicting the employment class. Regarding the unemployment class, RF demonstrates a precision of 95%, outperforming other

models. Conversely, the AdaBoost classifier yields a lower precision of 79% in identifying instances of labor force status.

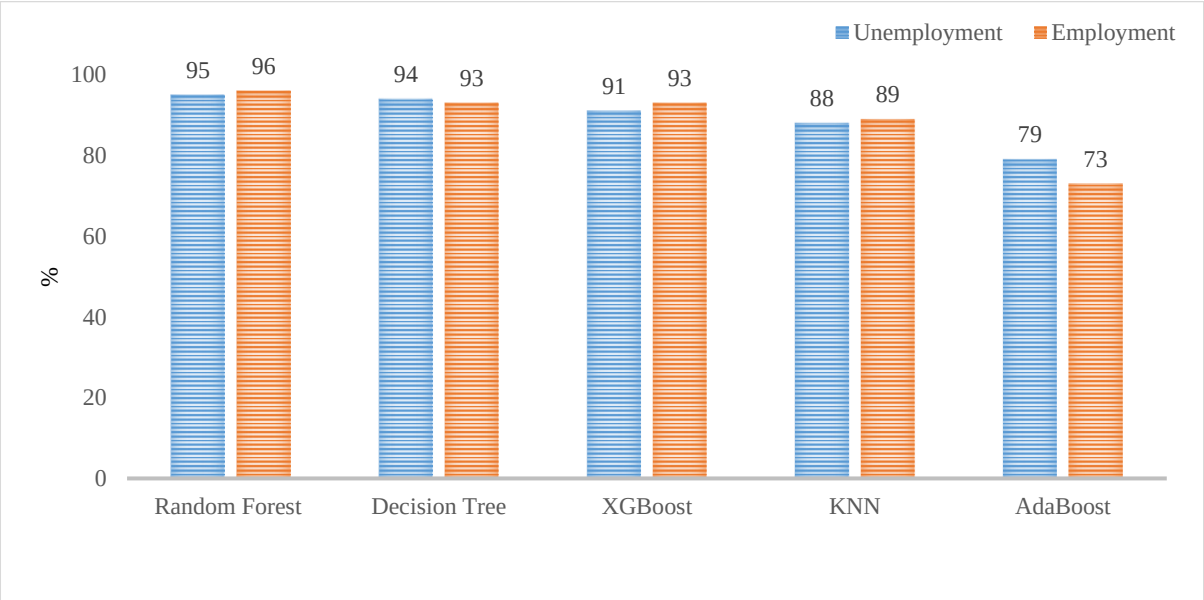


Figure 4.13 The precision of the five classifiers in labor status classes

In Figure 4.14, the recall of the five machine learning models (RF, DT, XGBoost, KNN, and AdaBoost) using the test dataset is depicted. It's apparent that the RF model outperforms the other models in both classes. Notably, the DT model and other models exhibit similar abilities in classification, particularly regarding the identification of unemployment cases. The recall percentages were notably close, with the DT achieving 94%, XGBoost at 95%, and KNN reaching 92% for identifying unemployment cases.

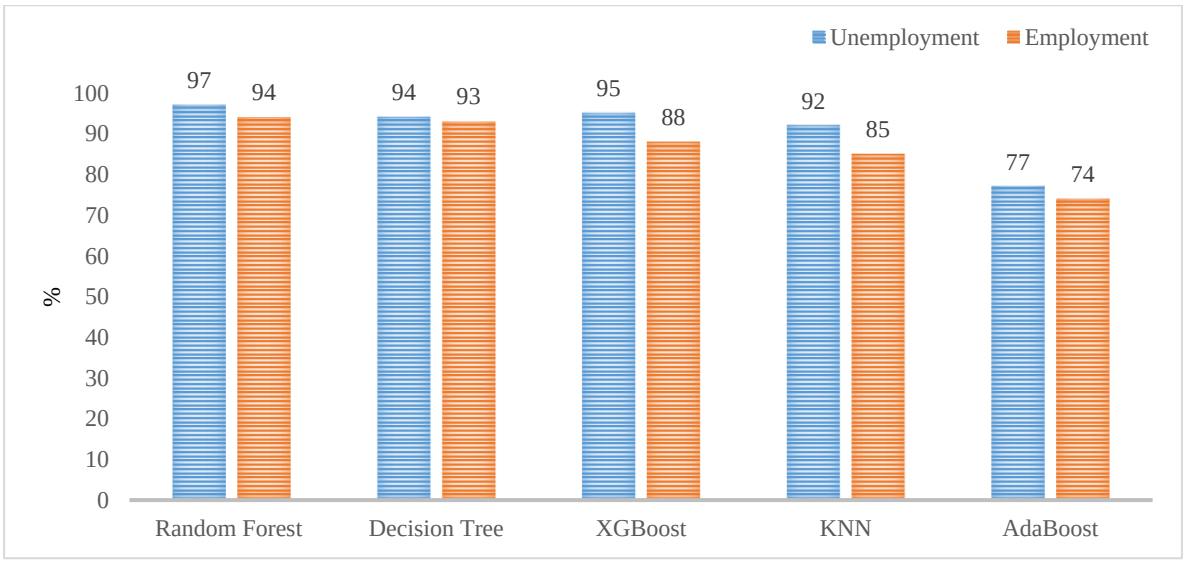


Figure 4.14 Recall of the five classifiers in labor status classes

In Figure 4.15, the f1-score performance of the five classifiers models (KNN, RF, DT, XGBoost, in addition to AdaBoost) is presented. Notably, RF achieves the highest f1-score in both classes compared to the other models, attaining values of (96%, 95%) for unemployment and employment, respectively. However, it's noteworthy that the f1 scores of the DT surpass those of XGBoost, KNN, and AdaBoost for both classes.

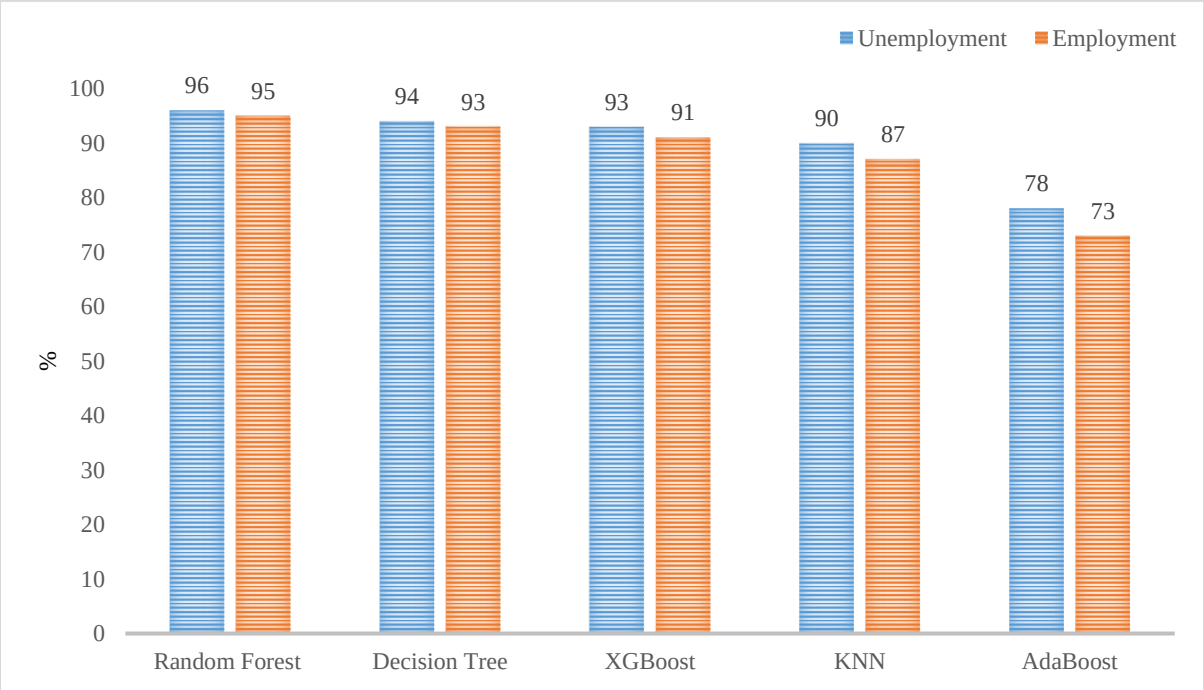


Figure 4.15 F1-Score of the five classifiers in labor status classes

4.8 Graduates by Majors

Introduction

This section aims to classify university graduates' work status (employed or unemployed) based on their specialization and various personal characteristics. To achieve this, three machine learning models (RF,DT, and KNN) will be applied to data from four select majors: Business, Administration & Law; Engineering, Manufacturing & Construction; Education; and Health & Welfare.

The selection of these four specializations was based on the significant sample size they represent, totaling more than 1,400 graduates. This ample sample size enables the selection of a representative subset for both training and testing purposes.

Fifteen independent variables, including gender, location, marital status, region, training courses, and university attended, will be used to assess the models' ability to predict graduates' work status. We will then analyze the accuracy, precision, recall, f1 score.

The entire experiment and analysis will be conducted using Anaconda Python programming.

4.8.1 University Major: Business, Administration and Law[3]

A total of 3,833 graduates, including both males and females, were included in the dataset for this university major. 20% of this data will be used for the test, while the remaining 80% will be used for training.

To analyze the graduates' data for this specific specialization, we'll employ three machine learning models: RF,DT, and KNN.

4.8.1.1 Random Forest Model

In this section, to predict labor status (employed or unemployed), we constructed a classification model using the RF algorithm by using 15 independent variables. This algorithm excels at handling multiple independent variables and a binary dependent variable (0 for unemployed, 1 for employed).

Hyper parameters applied to the RF, and selected best estimator and was used in training the algorithm, which including the Gini for criterion, while the max_depth,min_samples_leaf,

[3] Includes: (Business and administration), (Accounting and taxation),(Finance, banking and insurance), (Management and administration), (Marketing and advertising) and (Law).

min_samples_split, and n_estimators were (50),(1),(2),and (3000) respectively, the algorithm with these parameters has an accuracy of 94.9% using five-fold cross-validation.

4.8.1.2 Evaluating Model Performance

The Random Forest model achieved an accuracy of 94.9%, indicating a highly commendable level of accuracy in its performance as in table below:

Table 4.6 Accuracy and classification report by random forest for major (business, administration and law)

Random Forest Classifier –Test accuracy 0.949				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.96**	**0.96**	**0.96**	**446**
Employment	**0.94**	**0.94**	**0.94**	**321**
Accuracy			**0.96**	**767**
Macro avg	**0.95**	**0.95**	**0.95**	**767**
Weighted avg	**0.95**	**0.95**	**0.95**	**767**

From the classification table above, precision signifies the model's accuracy in making classifications. In our scenario, the precision values were 96% for unemployment and 94% for employment. Meanwhile, the Recall evaluation metric gauges the model's capability to detect positive samples, resulting gave the same of precision which was 96% for unemployment and 94% for employment. The RF model performed impressively, with outstanding prediction f1-score for both unemployment (96%) and employment (94%).

4.8.1.3 Confusion Matrix

The confusion matrix, displayed below, captures the distribution of correct and incorrect predictions made by the model, this table highlights four key combinations of actual and predicted values:

Figure 4.16 Confusion matrix result of random forest for major (business, administration and law)

Based on Figure 4.16, among a total of 767 individuals in the test dataset, 321 individuals are categorized as employed, while 446 individuals are classified as unemployed. Examining the confusion matrix results, it's evident that in the first row representing the unemployment class, 426 individuals are accurately classified as unemployed, but 20 individuals from the unemployment class are mistakenly classified as employment. Moving to the second row, 302 individuals are correctly classified as employed, while 19 individuals from the employment class are incorrectly labeled as unemployment.

4.8.1.4 Decision Tree

In this section, we elaborate on the performance evaluation and analysis of the machine learning approach applied to the labor force classification model using DT, by using the all of independent variables that we select it. In addition to optimize the performance, a hyper parameter tuning process was conducted for the DT model. This involved Gini for Criterion, while the max_depth, min_samples_leaf ,min_samples_split were (20),(2) and (2) respectively, the DT model achieved an accuracy of 93.3% using 5-fold cross-validation beside that we focus on key metrics such as accuracy, the classification report encompassing precision, recall, and f1-score, as well as the confusion matrix.

4.8.1.5 Evaluating Model Performance

The DT model achieved an accuracy of 94%, indicating a highly commendable level of accuracy in its performance.

Table 4.7 Accuracy and classification report by decision tree for major (business, administration and law)

Decision Tree Classifier –Test accuracy 0.94				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.95**	**0.95**	**0.95**	**446**
Employment	**0.93**	**0.93**	**0.93**	**321**
Accuracy			**0.94**	**767**
Macro avg	**0.94**	**0.94**	**0.94**	**767**
Weighted avg	**0.94**	**0.94**	**0.94**	**767**

From the classification table above, the precision, recall and f1-score were the same a values 95% for unemployment while for employment was (93%) the precision, recall and f1-score.

4.8.1.6 Confusion Matrix

The following confusion matrix showcases the distribution of accurate and inaccurate predictions performed by the model. It delineates four crucial combinations of actual and predicted values:

Figure 4.17 Confusion matrix result of decision tree for major (business, administration and law)

Figure 4.17 reveals the model's predictions for 767 test cases: 321 classified as employed and 446 as unemployed. A closer look at the confusion matrix shows accurate identification of 422 unemployed individuals, with only 24 misclassified as employed.

Moving to the second row, 299 employed instances are correctly classified, while 22 instances from the employed class are incorrectly predicted as unemployed.

4.8.1.7 KNN

In this section, a labor status classification model was constructed utilizing the KNN algorithm. The KNN model was employed to classify the data based on a collection of independent variables, with the dependent variable restricted to two values (0 and 1).

The k-value, a critical hyperparameter, significantly impacts model behavior and accuracy. After careful tuning, we found k=3 to be optimal, resulting in the highest achieved accuracy of 89.4%.

4.8.1.8 Evaluating Model Performance

The KNN model shows a good performance with an 88.4% accuracy.

Table 4.8 Accuracy and classification report by KNN for major (business, administration and law)

KNN Classifier –Test accuracy 0.884				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.89**	**0.93**	**0.91**	**446**
Employment	**0.89**	**0.85**	**0.87**	**321**
Accuracy			**0.89**	**767**
Macro avg	**0.89**	**0.89**	**0.89**	**767**
Weighted avg	**0.89**	**0.89**	**0.89**	**767**

From the classification table above, the precision values were 89% for unemployment and employment, compared the recall evaluation metric was 85% for employment and 93% for unemployment. In other side the f1-score values were 87% for employment and 91% for unemployment.

4.8.1.9 Confusion Matrix

After running the above confusion matrix code, the following tabular form shown below consists of 4 different combinations of actual and predicted values.

Figure 4.18 Confusion matrix result of KNN for major (business, administration and law)

Figure 4.18 reveals that in a test dataset of 767 individuals, 321 were categorized as employed and 446 as unemployed. Analyzing the confusion matrix. Examining the confusion matrix results, it's evident that in the first row representing the unemployed class, 419 individuals are accurately classified as employed, but 27 individuals from the unemployed class are mistakenly classified as employment. Moving to the second row, 298 individuals are correctly classified as employment, while 23 individuals from the unemployment class are incorrectly labeled as unemployed.

4.8.1.10 Evaluating Classification Algorithms

A comparative performance analysis was conducted among three machine learning algorithms, evaluating metrics such as accuracy, precision, recall, f1-score. According to this analysis, the RF classifier exhibited superior performance compared to DT, and KNN models, as indicated in Figures 4.19 – 4.22.

In particular, Figures 4.19 depicted the accuracy of the proposed labor force status classification model using the three classifiers: RF,DT, and KNN. Notably, the figure illustrated that RF achieved the highest accuracy among all classifiers, with Decision Tree also displaying relatively strong accuracy compared to the other models.

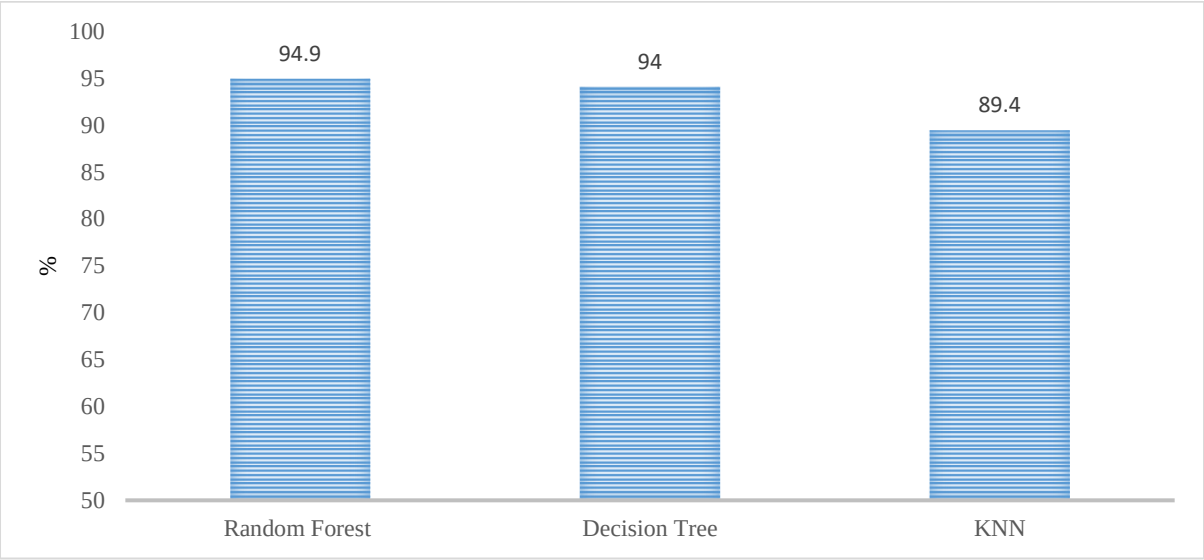


Figure 4.19 Accuracy of three classifiers for major (business, administration and law)

Figure 4.20 illustrates the precision of three machine learning models. Upon fine-tuning parameter values, it becomes evident that RF attains the highest precision, while the KNN

classifier exhibits the lowest precision specifically in predicting the employment class. Regarding the unemployment class, RF demonstrates a precision of 96%, outperforming other models. Conversely, the KNN classifier yields a lower precision of 89% in identifying instances of labor force status.

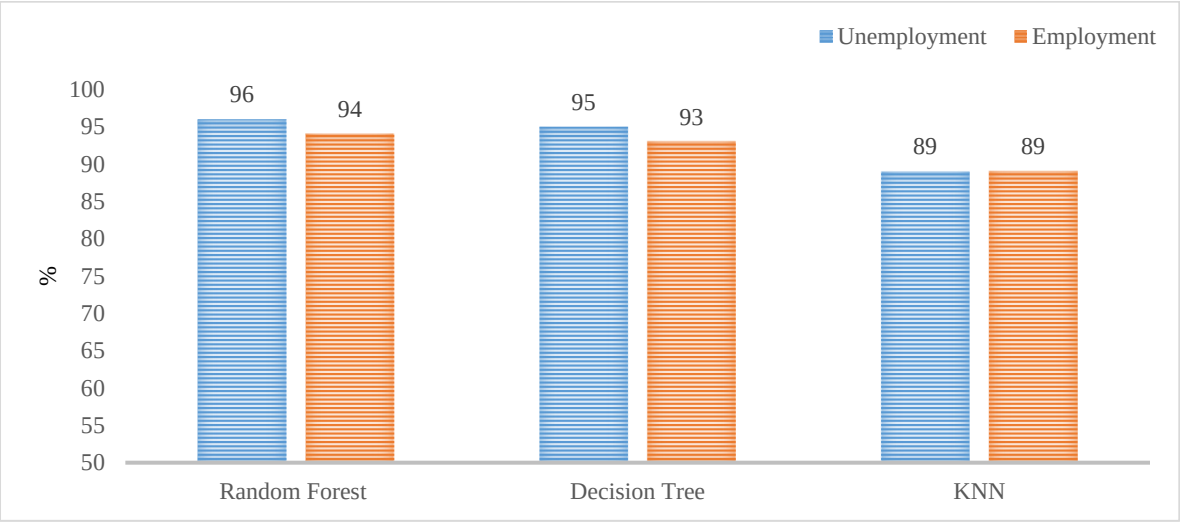


Figure 4.20 The precision of the three classifier in labor status classes for major (business, administration and law)

In Figure 4.21, the recall of the three machine learning models (RF, DT, and KNN) using the test dataset is depicted. It's apparent that the RF model outperforms the other models in both classes. Notably, the DT model and other models exhibit similar abilities in classification, particularly regarding the identification of employment cases. The recall percentages were notably close, with the DT achieving 95%, and KNN reaching 92% for identifying unemployment cases.

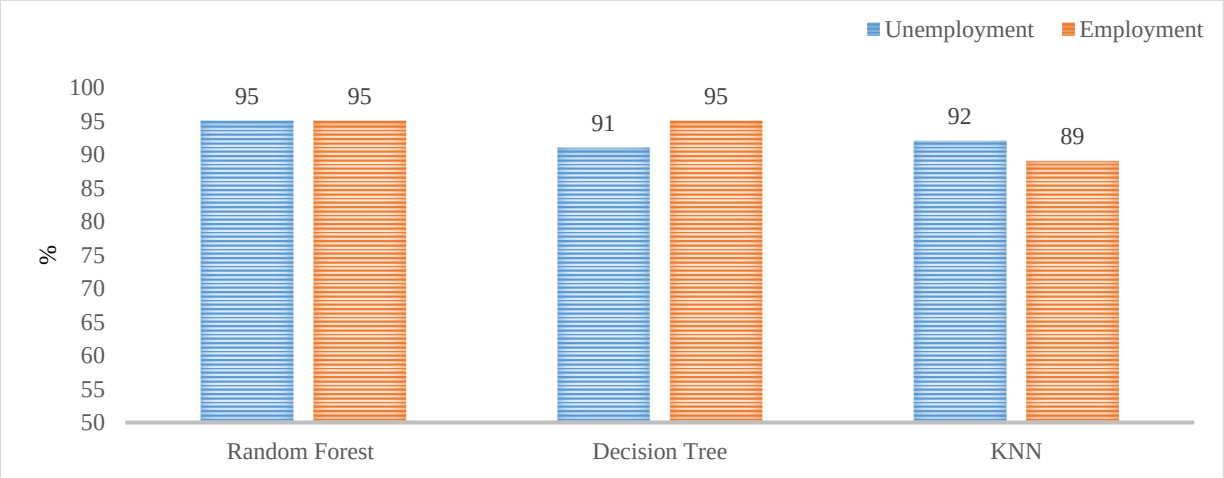


Figure 4.21 Recall of the three classifier in labor status classes for major (business, administration and law)

In Figure 4.22, the f1-score performance of the three classifiers models (RF,DT,KNN) is presented. Notably, RF achieves the highest f1-score in both classes compared to the other models, attaining values of (94%, 95%) for unemployment and employment, respectively. However, it's noteworthy that the f1 scores of the DT surpass KNN classifier.

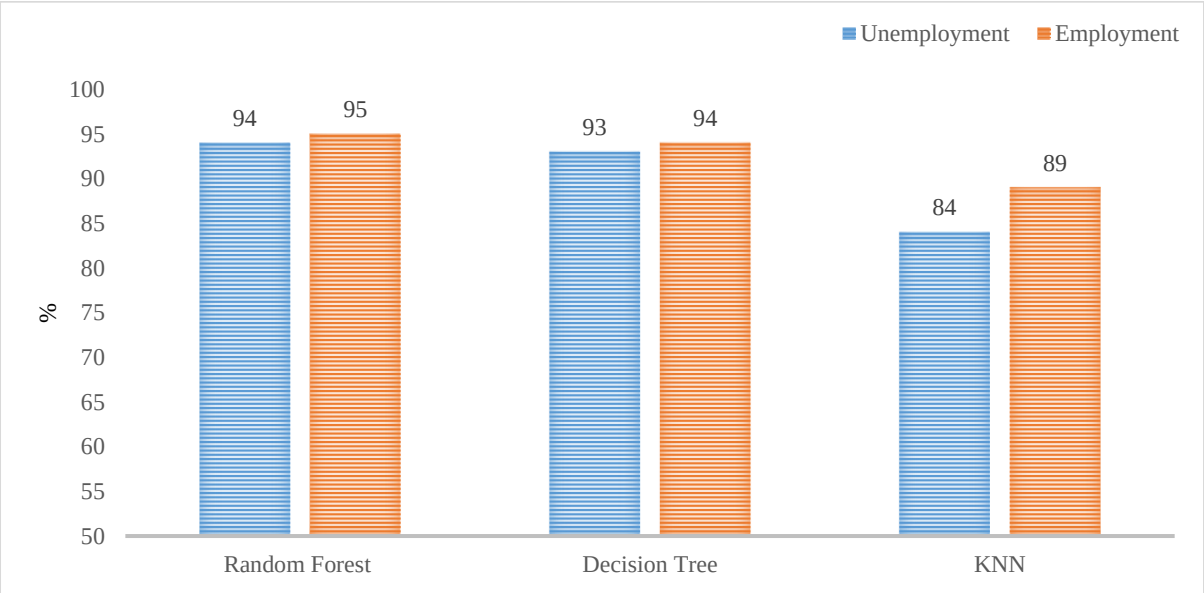


Figure 4.22 F1-Score of the three classifier in labor status classes for major (business, administration and law)

4.8.2 University Major: Engineering, Manufacturing and Construction[4]

A total of 1,407 graduates, including both males and females, were included in the dataset for this University Major. 20% of this data will be used for the test, while the remaining 80% will be used for training.

To analyze the graduates' data for this specific specialization, we'll employ three machine learning models: RF, DT, and KNN.

4.8.2.1 Random Forest Model

In this section, to predict labor status (employed or unemployed), we constructed a classification model using the RF algorithm by using 15 independent variables. This algorithm excels at handling multiple independent variables and a binary dependent variable (0 for unemployed, 1 for employed).

[4] Includes: (Engineering and engineering trades), (Chemical engineering and processe), (Environmental protection technology), (Electricity and energy),(Electronics and automation) and (Mechanics and metal trades).

Hyperparameters applied to the RF,and selected best estimator and was used in training the algorithm,which including the Gini for criterion,while the max_depth,min_samples_leaf, min_samples_split, and n_estimators were (100),(1),(2),and (3000) respectively, the algorithm with these parameters has an accuracy of 94.7% using five-fold cross-validation.

4.8.2.1 Evaluating Model Performance

The Random Forest model achieved an accuracy of 94.7%, indicating a highly commendable level of accuracy in its performance as in table below:

Table 4.9 Accuracy and classification report by random forest for major (engineering, manufacturing and construction)

Random Forest Classifier –Test accuracy 0.947				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.94**	**0.95**	**0.94**	**134**
Employment	**0.95**	**0.95**	**0.95**	**148**
Accuracy			**0.95**	**282**
Macro avg	**0.95**	**0.95**	**0.95**	**282**
Weighted avg	**0.95**	**0.95**	**0.95**	**282**

From the classification table above 4.9, precision signifies the model's accuracy in making classifications. In our scenario, the precision values were 94% for unemployment and 95% for employment. Meanwhile, the Recall evaluation metric gauges the model's capability to detect positive samples, resulting in 95% for unemployment and employment.

The RF model performed impressively, with outstanding prediction f1-score for both unemployment (94%) and employment (95%).

4.8.2.2 Confusion Matrix

The confusion matrix, displayed below, captures the distribution of correct and incorrect predictions made by the model, this table highlights four key combinations of actual and predicted values:

Figure 4.23 Confusion matrix result of random forest for major (engineering, manufacturing and construction)

Based on Figure 4.23, among a total of 282 individuals in the test dataset, 134 individuals are categorized as employed, while 148 individuals are classified as unemployed. Examining the confusion matrix results, it's evident that in the first row representing the unemployment class, 127 individuals are accurately classified as unemployed, but 7 individuals from the unemployment class are mistakenly classified as employment. Moving to the second row, 140 individuals are correctly classified as employed, while 8 individuals from the employment class are incorrectly labeled as unemployment.

4.8.2.3 Decision Tree

In this section, we elaborate on the performance evaluation and analysis of the machine learning approach applied to the labor force classification model using DT, by using the all of independent variables that we select it. In addition to optimize the performance, a hyper parameter tuning process was conducted for the DT model. This involved entropy for Criterion, while the max_depth, min_samples_leaf ,min_samples_split were (20),(1) and (2) respectively, the DT model achieved an accuracy of 93.3% using 5-fold cross-validation beside that we focus on key metrics such as accuracy, the classification report encompassing precision, recall, and f1-score, as well as the confusion matrix.

4.8.2.4 Evaluating Model Performance

The Decision Tree model achieved an accuracy of 93.3%, indicating a highly commendable level of accuracy in its performance.

Table 4.10 Accuracy and classification report by decision tree for major (engineering, manufacturing and construction)

Decision Tree Classifier –Test accuracy 0.933				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.95**	**0.91**	**0.93**	**134**
Employment	**0.92**	**0.95**	**0.94**	**148**
Accuracy			**0.93**	**282**
Macro avg	**0.93**	**0.93**	**0.93**	**282**
Weighted avg	**0.93**	**0.93**	**0.93**	**282**

From the classification table above, the precision, recall and f1-score were 95%,91%, and 93% respectively for unemployment while for employment was (92%) for precision and 95% for f1-score, while it was 94% for recall.

4.8.2.5 Confusion Matrix

The following confusion matrix showcases the distribution of accurate and inaccurate predictions performed by the model. It delineates four crucial combinations of actual and predicted values:

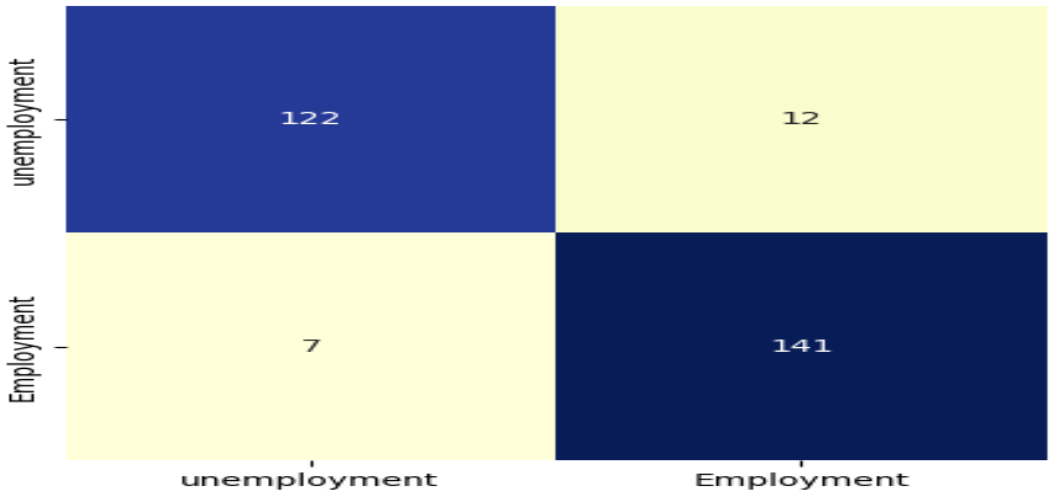


Figure 4.24 Confusion matrix result of decision tree for major (engineering, manufacturing and construction)

Figure 4.24 reveals the model's predictions for 2,390 test cases: 148 classified as employed and 134 as unemployed. A closer look at the confusion matrix shows accurate identification of 122 unemployed individuals, with only 12 misclassified as employed.

Moving to the second row, 141 employed instances are correctly classified, while 7 instances from the employed class are incorrectly predicted as unemployed.

4.8.2.6 KNN

In this section, a labor status classification model was constructed utilizing the KNN algorithm. The KNN model was employed to classify the data based on a collection of independent variables, with the dependent variable restricted to two values (0 and 1).

The crucial hyperparameter to tune is the k-value, which directly influences model behavior and accuracy, we used k=3 which achieved the best accuracy 85.5%.

4.8.2.7 Evaluating Model Performance

The KNN model shows a good performance with an 85.5% accuracy.

Table 4.11 Accuracy and classification report by KNN for major (engineering, manufacturing and construction)

KNN Classifier –Test accuracy 0.855				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.87**	**0.82**	**0.84**	**134**
Employment	**0.85**	**0.89**	**0.86**	**148**
Accuracy			**0.85**	**282**
Macro avg	**0.86**	**0.85**	**0.85**	**282**
Weighted avg	**0.86**	**0.85**	**0.85**	**282**

From the classification table above, the precision values were 87% for unemployment, while 85% for employment, compared the recall evaluation metric was 82% for employment and 89% for unemployment. In other side the f1-score values were 84% for employment and 86% for unemployment.

4.8.2.8 Confusion Matrix

After running the above confusion matrix code, the following tabular form shown below consists of 4 different combinations of actual and predicted values.

Figure 4.25 Confusion matrix result for KNN for major (engineering, manufacturing and construction)

Figure 4.25 reveals that in a test dataset of 282 individuals, 148 were categorized as employed and 134 as unemployed. Analyzing the confusion matrix. Examining the confusion matrix results, it's evident that in the first row representing the unemployed class, 110 individuals are accurately classified as employed, but 24 individuals from the unemployed class are mistakenly classified as employment. Moving to the second row, 131 individuals are correctly classified as employment, while 17 individuals from the unemployment class are incorrectly labeled as unemployed.

4.8.2.9 Evaluating Classification Algorithms

A comparative performance analysis was conducted among three machine learning algorithms, evaluating metrics such as accuracy, precision, recall, F1-score. According to this analysis, the RF classifier exhibited superior performance compared to DT, and KNN models, as indicated in Figures 4.26 – 4.29

In particular, Figures 4.26 depicted the accuracy of the proposed labor force status classification model using the three classifiers: RF, DT, and KNN. Notably, the Figure illustrated that RF achieved the highest accuracy among all classifiers, with DT also displaying relatively strong accuracy compared to the other models

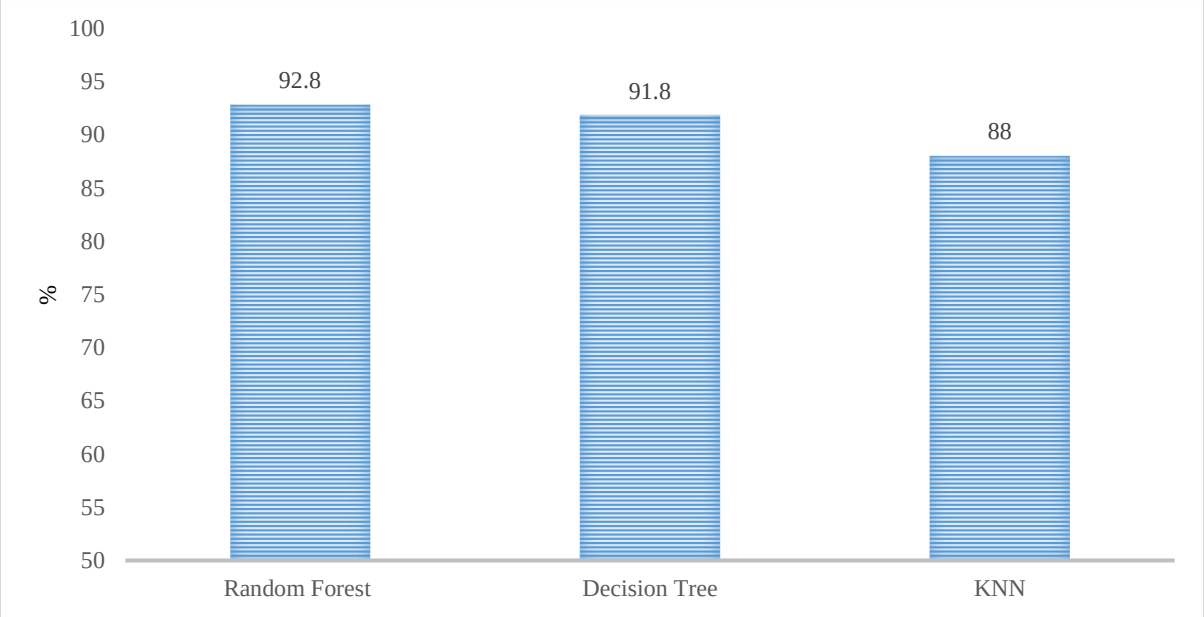


Figure 4.26 Accuracy of three classifiers for major (engineering, manufacturing and construction)

Figure 4.27 illustrates the precision of three machine learning models. Upon fine-tuning parameter values, it becomes evident that RF attains the highest precision, while the KNN classifier exhibits the lowest precision specifically in predicting the employment class. Regarding the unemployment class, RF demonstrates a precision of 92.8%, outperforming other models. Conversely, the KNN classifier yields a lower precision of 88% in identifying instances of labor force status.

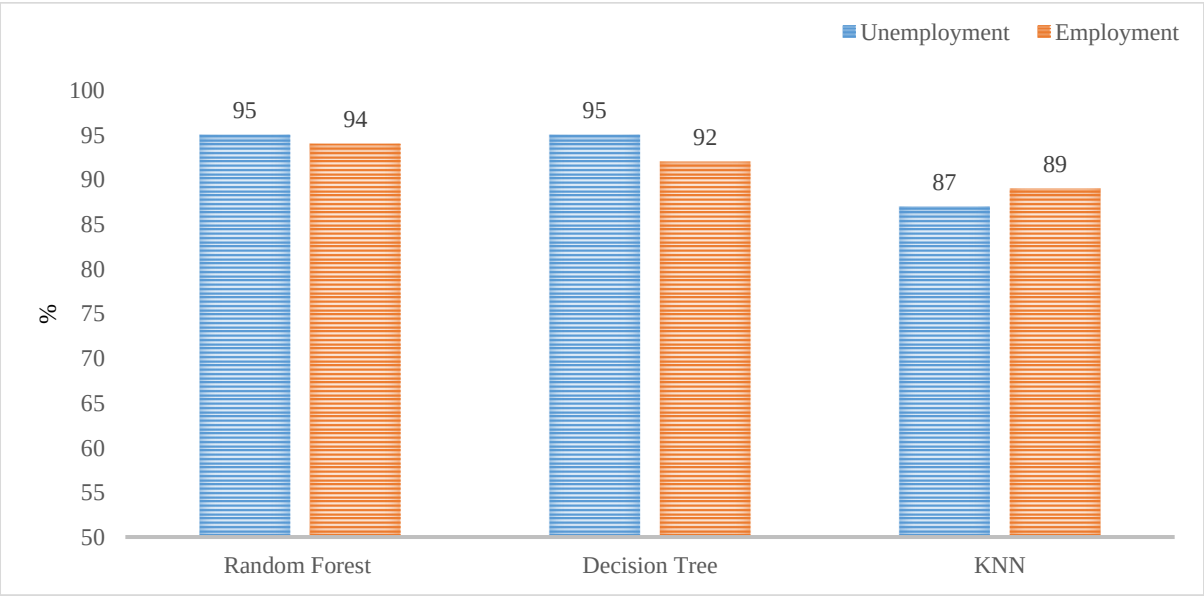


Figure 4.27 The precision of the three classifier in labor status classes for major (engineering, manufacturing and construction)

In Figure 4.28, the recall of the three machine learning models (RF, DT, and KNN) using the test dataset is depicted. It's apparent that the RF model outperforms the other models in both classes. Notably, the DT model and other models exhibit similar abilities in classification, particularly regarding the identification of employment cases. The recall percentages were notably close, with the DT achieving 95%, and KNN reaching 87% for identifying unemployment cases.

Figure 4.28 Recall of the three classifier in labor status classes for major (engineering, manufacturing and construction)

In Figure 4.29, the f1-score performance of the three classifiers models (RF,DT,KNN) is presented. Notably, RF achieves the highest f1-score in both classes compared to the other models, attaining values of (94%, 95%) for unemployment and employment, respectively. However, it's noteworthy that the f1 scores of the DT surpass KNN classifier.

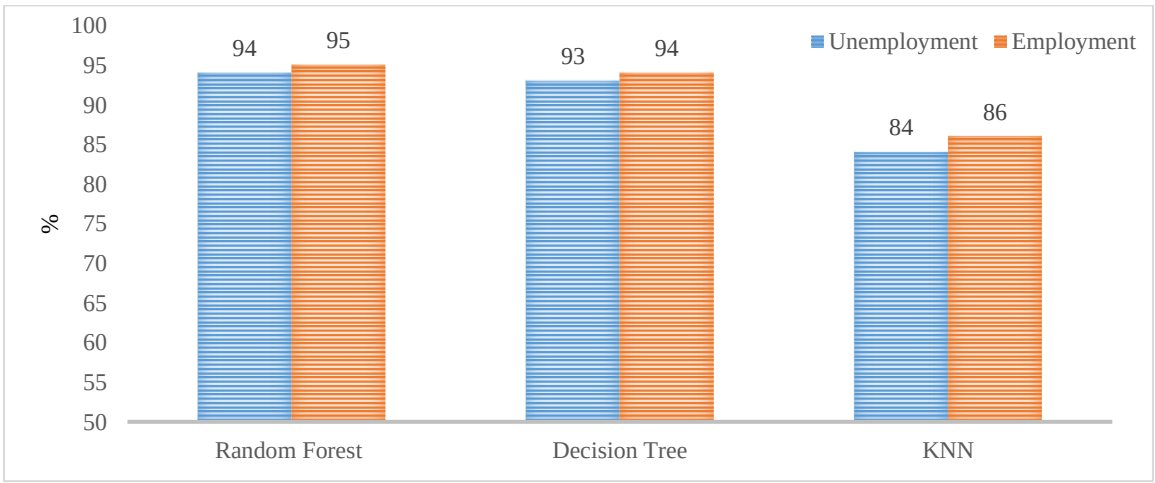


Figure 4.29 F1-Score of the three classifier in labor status classes for major (Engineering, manufacturing and construction)

4.8.3 University Major: Education[5]

A total of 1,521 graduates, including both males and females, were included in the dataset for this University Major. 20% of this data will be used for the test, while the remaining 80% will be used for training.

To analyze the graduates' data for this specific specialization, we'll employ three machine learning models: RF, DT, and KNN.

4.8.3.1 Random Forest Model

Hyper parameters applied to the RF, and selected best estimator and was used in training the algorithm, which including the Gini for criterion, while the max_depth,min_samples_leaf, min_samples_split, and n_estimators were (100),(1),(2),and (1000) respectively, the algorithm with these parameters has an accuracy of 94.1% using five-fold cross-validation.

4.8.3.2 Evaluating Model Performance

The RF model achieved an accuracy of 94.1%, indicating a highly commendable level of accuracy in its performance as in table below:

Table 4.12 Accuracy and classification report by random forest for major (education)

Random Forest Classifier –Test accuracy 0.941				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.93**	**0.95**	**0.96**	**200**
Employment	**0.96**	**0.87**	**0.91**	**104**
Accuracy			**0.94**	**304**
Macro avg	**0.95**	**0.92**	**0.93**	**304**
Weighted avg	**0.94**	**0.94**	**0.94**	**304**

From the classification table above, precision signifies the model's accuracy in making classifications. In our scenario, the precision values were 93% for unemployment and 96% for employment. Meanwhile, the Recall evaluation metric gauges the model's capability to detect positive samples, resulting in 98% for unemployment and 87% for employment.

[5] Includes: (Education science),(Training for pre-school teachers),(Teacher training with subject specialization) and (Teacher training without subject specialization).

The RF model performed impressively, with outstanding prediction f1-score for both unemployment (96%) and employment (91%).

4.8.3.2 Confusion Matrix

The confusion matrix, displayed in Figure 4.30 below, captures the distribution of correct and incorrect predictions made by the model, this table highlights four key combinations of actual and predicted values:

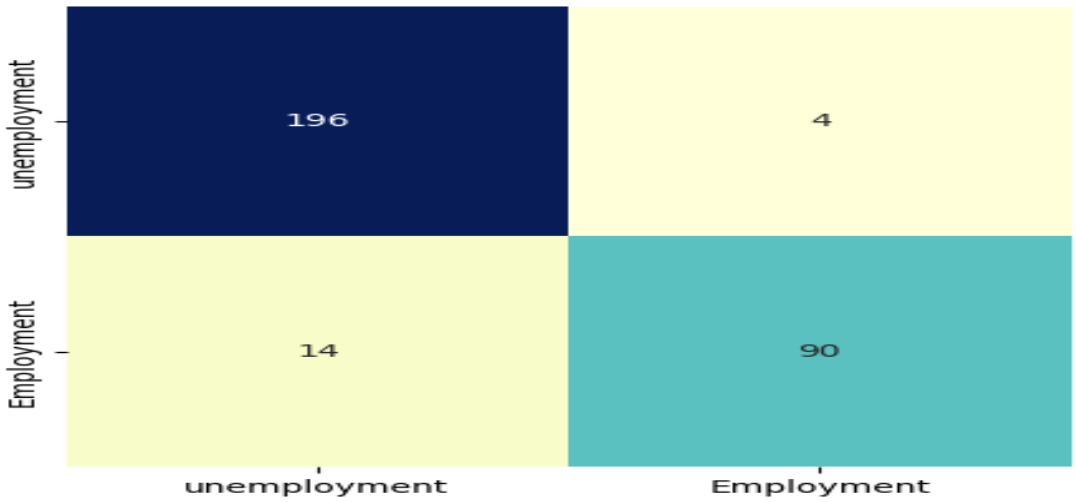


Figure 4.30 Confusion matrix result of random forest for major (education)

Based on Figure 430, among a total of 304 individuals in the test dataset, 104 individuals are categorized as employed, while 200 individuals are classified as unemployed. Examining the confusion matrix results, it's evident that in the first row representing the unemployment class, 196 individuals are accurately classified as unemployed, but 4 individuals from the unemployment class are mistakenly classified as employment. Moving to the second row, 90 individuals are correctly classified as employed, while 14 individuals from the employment class are incorrectly labeled as unemployment.

4.8.3.3 Decision Tree

In this section, we elaborate on the performance evaluation and analysis of the machine learning approach applied to the labor force classification model using DT, by using the all of independent variables that we select it. In addition to optimize the performance, a hyper parameter tuning process was conducted for the DT model. This involved Entropy for Criterion, while the max_depth, min_samples_leaf ,min_samples_split were (20),(1) and (2) respectively, the DT model achieved an accuracy of 92.8% using 5-fold cross-validation beside

that we focus on key metrics such as accuracy, the classification report encompassing precision, recall, and f1-score, as well as the confusion matrix.

4.8.3.4 Evaluating Model Performance

The DT model achieved an accuracy of 92.8%, indicating a highly commendable level of accuracy in its performance.

Table 4.13 Accuracy and classification report by decision tree for major (education)

Decision Tree Classifier –Test accuracy 0.928				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.93**	**0.96**	**0.95**	**200**
Employment	**0.93**	**0.86**	**0.89**	**104**
Accuracy			**0.93**	**304**
Macro avg	**0.93**	**0.91**	**0.92**	**304**
Weighted avg	**0.93**	**0.93**	**0.93**	**304**

From the classification table above, the precision, recall and f1-score were 93%,96%, and 95% respectively for unemployment while for employment was (93%) for precision and 86% for f1-score, while it was 89% for recall.

4.8.3.4 Confusion Matrix

The following confusion matrix showcases the distribution of accurate and inaccurate predictions performed by the model. It delineates four crucial combinations of actual and predicted values:

Figure 4.31 Confusion matrix result of decision tree for major (education)

Figure 4.31 reveals the model's predictions for 304 test cases: 104 classified as employed and 200 as unemployed. A closer look at the confusion matrix shows accurate identification 193 unemployed individuals, with only 7 misclassified as employed.

Moving to the second row, 89 employed instances are correctly classified, while 15 instances from the employed class are incorrectly predicted as unemployed.

4.8.3.6 KNN

In this section, a labor status classification model was constructed utilizing the KNN algorithm. The KNN model was employed to classify the data based on a collection of independent variables, with the dependent variable restricted to two values (0 and 1).
Through tuning, k=3 was found to be the optimal value, leading to the highest accuracy of 86.8%.

4.8.3.7 Evaluating Model Performance

The KNN model shows a good performance with an 86.8% accuracy.

Table 4.14 Accuracy and classification report by KNN for major (education)

KNN Classifier –Test accuracy 0.868				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.93**	**0.96**	**0.95**	**200**
Employment	**0.93**	**0.86**	**0.89**	**104**
Accuracy			**0.87**	**304**
Macro avg	**0.87**	**0.83**	**0.85**	**304**
Weighted avg	**0.87**	**0.87**	**0.87**	**304**

From the classification Table 4.14 above , the precision values were 87% for unemployment and for employment, compared the Recall evaluation metric was 72% for employment and 94% for unemployment. In other side the f1-score values were 90% for unemployment and 79% for employment.

4.8.3.8 Confusion Matrix

After running the above confusion matrix code, the following tabular form shown below consists of 4 different combinations of actual and predicted values.

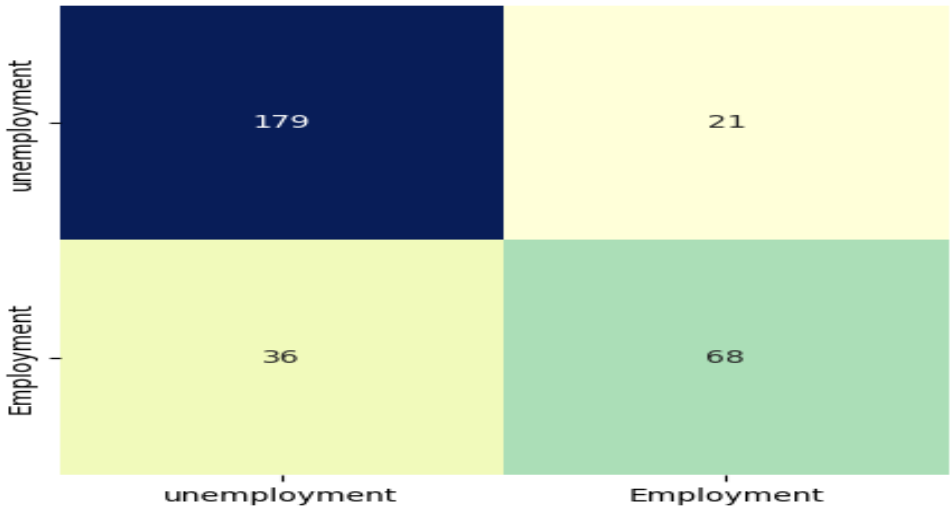


Figure 4.32 Confusion matrix result for KNN for major (education)

Figure 4.32 reveals that in a test dataset of 304 individuals, 104 were categorized as employed and 200 as unemployed. Analyzing the confusion matrix. Examining the confusion matrix results, it's evident that in the first row representing the unemployed class, 179 individuals are accurately classified as employed, but 21 individuals from the unemployed class are mistakenly classified as employment. Moving to the second row, 68 individuals are correctly classified as employment, while 36 individuals from the unemployment class are incorrectly labeled as unemployed.

4.8.3.9 Evaluating Classification Algorithms

A comparative performance analysis was conducted among three machine learning algorithms, evaluating metrics such as accuracy, precision, recall, f1-score. According to this analysis, the RF classifier exhibited superior performance compared to DT, and KNN models, as indicated in Figures 4.33 – 4.36. In particular, Figures 4.33 depicted the accuracy of the proposed labor force status classification model using the three classifiers: RF, DT, and KNN. Notably, the Figure below showed that RF achieved the highest accuracy among all classifiers, with DT also displaying relatively strong accuracy compared to the other models.

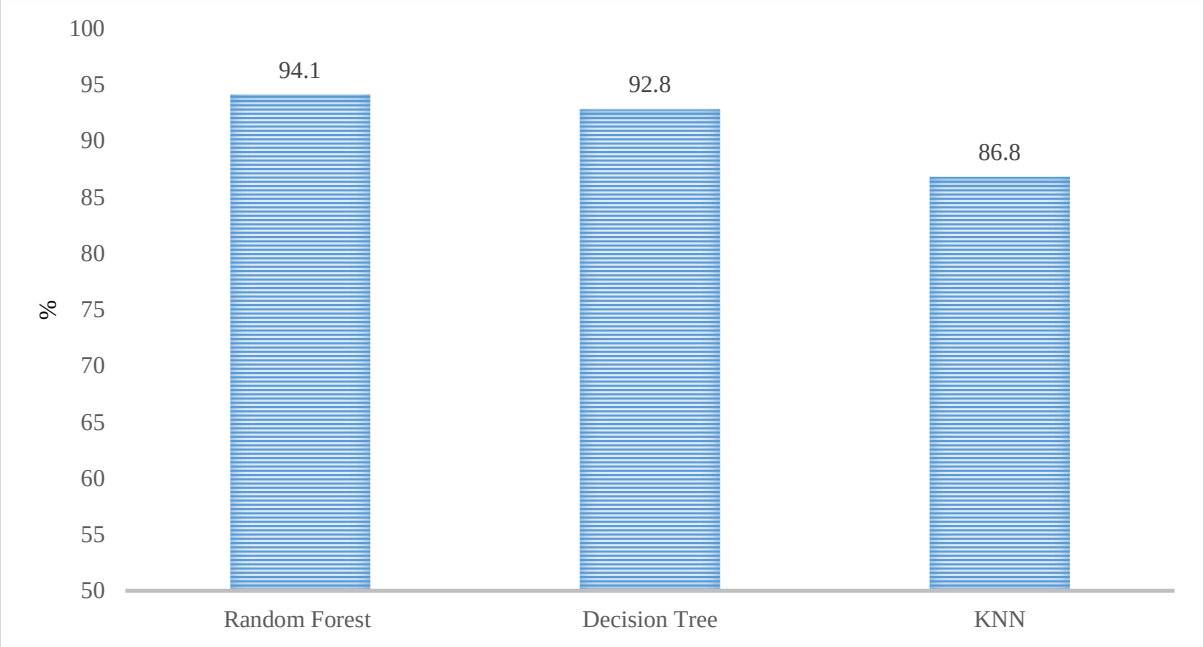


Figure 4.33 Accuracy of three classifiers for major (education)

Figure 4.34 illustrates the precision of three machine learning models. Upon fine-tuning parameter values, it becomes evident that RF attains the highest precision, while the KNN classifier exhibits the lowest precision specifically in predicting the employment and unemployment class. Regarding the unemployment class, RF and DT demonstrates a precision of 95%, outperforming KNN models.

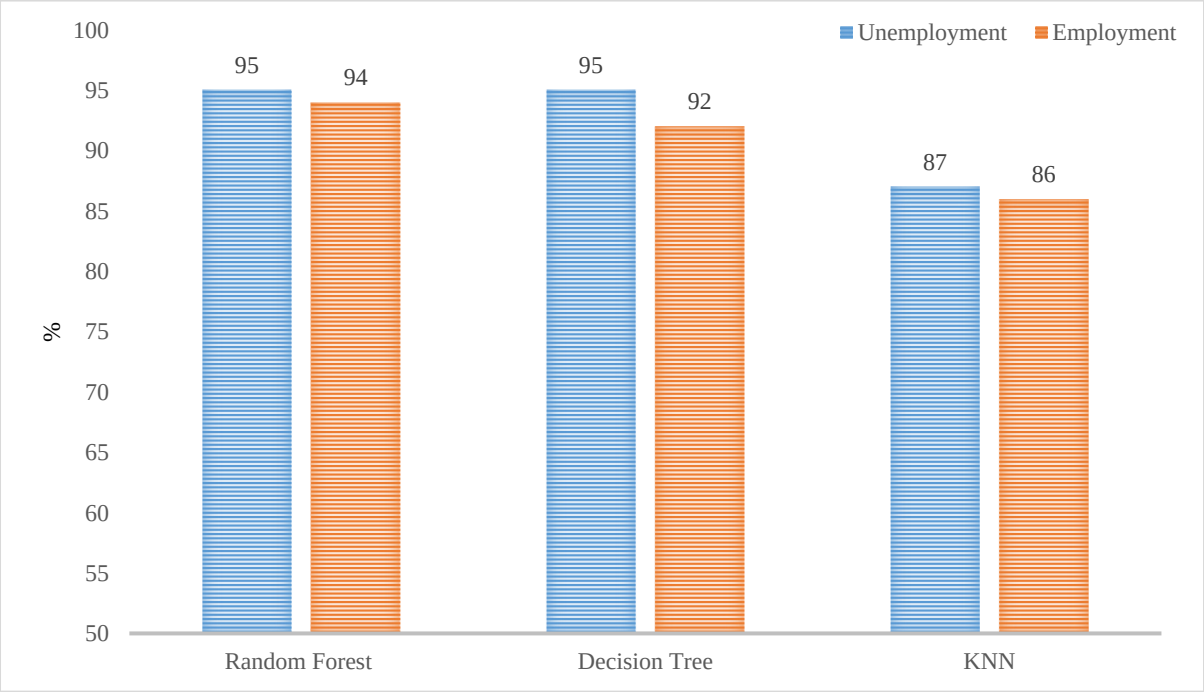


Figure 4.34 The precision of the three classifier in labor status classes for major (education)

In Figure 4.35, the recall of the three machine learning models (RF, DT, and KNN) using the test dataset is depicted. It's apparent that the RF model outperforms the other models in both classes. Notably, the DT model and other models exhibit similar abilities in classification, particularly regarding the identification of employment cases. The recall percentages were notably close, with the DT achieving 95%, and KNN reaching 92% for identifying unemployment cases.

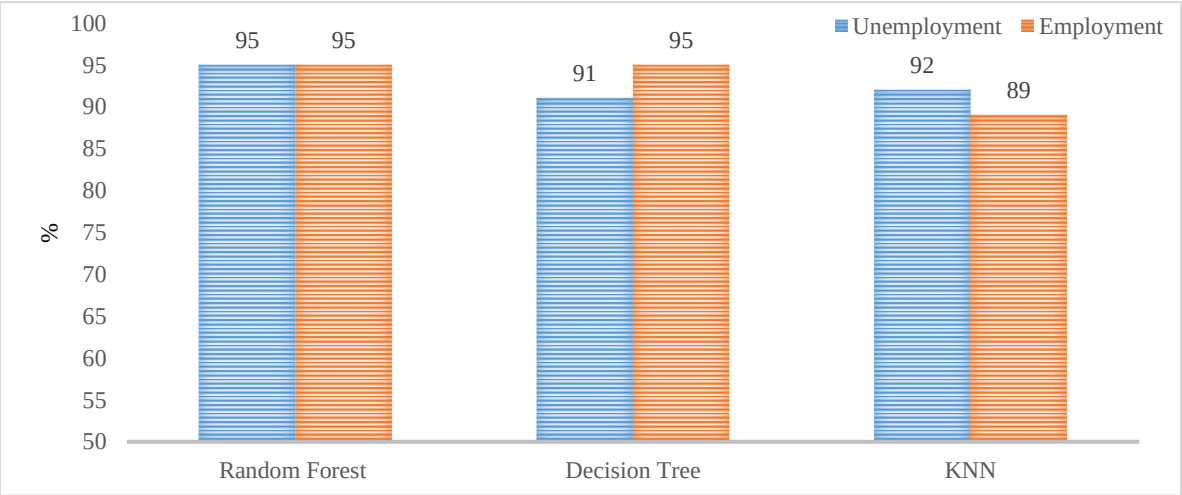


Figure 4.35 Recall of the three classifier in labor status classes for major (education)

In Figure 4.36, the recall of the three machine learning models (RF, DT, and KNN) using the test dataset is depicted. It's apparent that the RF model outperforms the other models in both classes with 94% compared with 93% and 84% for DT and KNN Respectively for unemployment classification while it was 95%, 94% and 86% for RF, DT, and KNN respectively for classification employment.

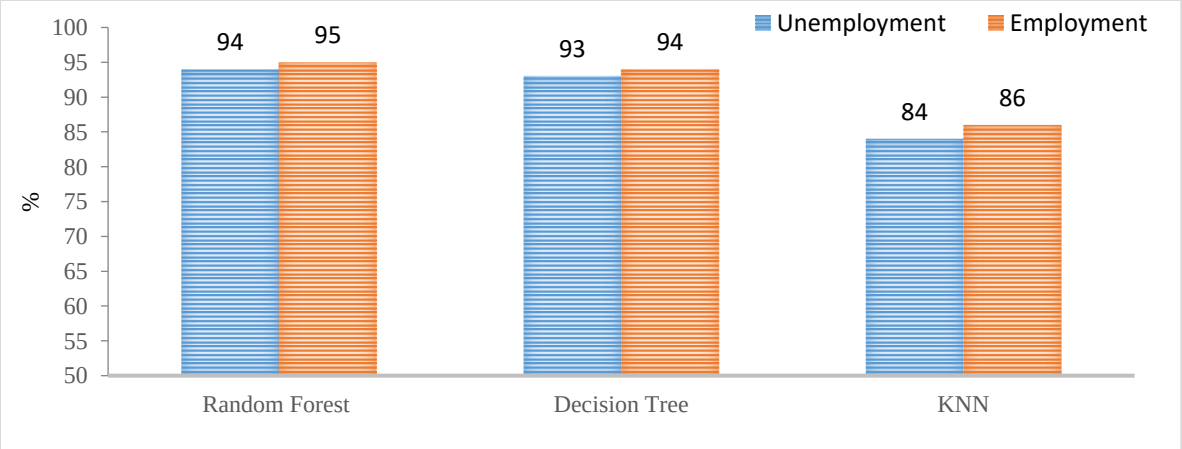


Figure 4.36 F1-Score of the three classifier in labor status classes for major (education)

4.8.4 University Major: Health and Welfare[6]

A total of 1,748 graduates, including both males and females, were included in the dataset for this University Major. 20% of this data will be used for the test, while the remaining 80% will be used for training.

To analyze the graduates' data for this specific specialization, we'll employ three machine learning models: RF, DT, and KNN.

4.8.4.1 Random Forest Model

Hyper parameters applied to the RF, and selected best estimator and was used in training the algorithm, which including the Gini for criterion, while the max_depth,min_samples_leaf, min_samples_split, and n_estimators were (100),(1),(2),and (3000) respectively, the algorithm with these parameters has an accuracy of 94% using five-fold cross-validation.

4.8.4.2 Evaluating Model Performance

The RF model achieved an accuracy of 94%, indicating a highly commendable level of accuracy in its performance as in table below:

Table 4.15 Accuracy and classification report by random forest for major (health and welfare)

Random Forest Classifier –Test accuracy 0.94				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.93**	**0.94**	**0.93**	**158**
Employment	**0.95**	**0.93**	**0.95**	**192**
Accuracy			**0.94**	**350**
Macro avg	**0.94**	**0.94**	**0.94**	**350**
Weighted avg	**0.94**	**0.94**	**0.94**	**350**

From the classification table above, precision signifies the model's accuracy in making classifications. In our scenario, the precision values were 93% for unemployment and 95% for employment. Meanwhile, the recall evaluation metric gauges the model's capability to detect positive samples, resulting in 94% for unemployment and employment.

The RF model performed impressively, with outstanding prediction f1-score for both unemployment (93%) and employment (95%).

[6] Includes: (Dental studies),(Medicine),(Nursing and midwifery),(Medical diagnostic and treatment technology) ,(Therapy and rehabilitation), (Pharmacy) and (Welfare)

4.8.4.2 Confusion Matrix

The confusion matrix, displayed below, captures the distribution of correct and incorrect predictions made by the model, this table highlights four key combinations of actual and predicted values:

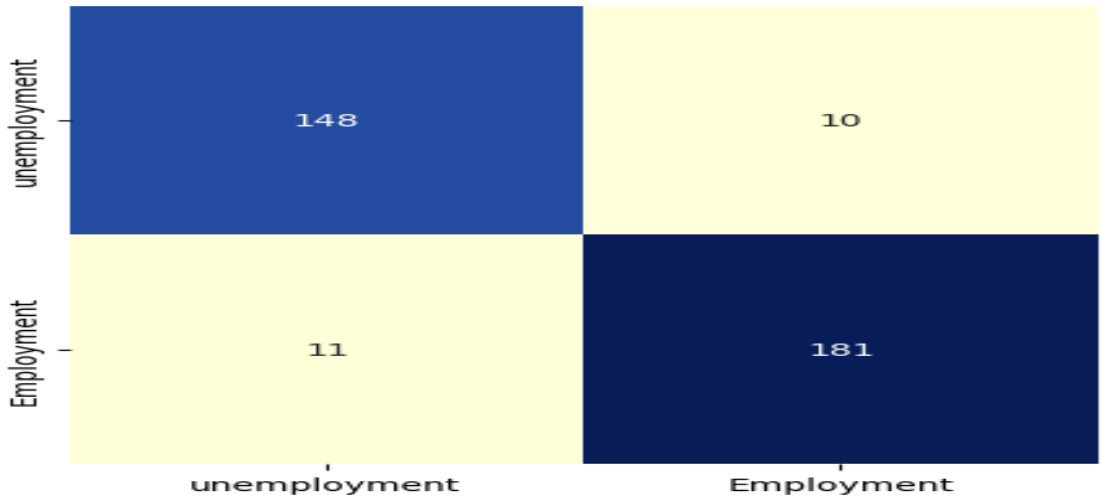


Figure 4.37 Confusion matrix result of random forest for major (health and welfare)

Based on Figure 4.37, among a total of 350 individuals in the test dataset, 192 individuals are categorized as employed, while 158 individuals are classified as unemployed. Examining the confusion matrix results, it's evident that in the first row representing the unemployment class, 148 individuals are accurately classified as unemployed , but 10 individuals from the unemployment class are mistakenly classified as employment . Moving to the second row, 181 individuals are correctly classified as employed, while 11 individuals from the employment class are incorrectly labeled as unemployment.

4.8.4.3 Decision Tree

In this section, we elaborate on the performance evaluation and analysis of the machine learning approach applied to the labor force classification model using DT, by using the all of independent variables that we select it. In addition to optimize the performance, a hyper parameter tuning process was conducted for the DT model. This involved *Gini for Criterion*, while the max_depth, min_samples_leaf ,min_samples_split were (50),(1) and (2) respectively, the DT model achieved an accuracy of 92% using *5-fold cross-validation* beside that we focus on key metrics such as accuracy, the classification report encompassing precision, recall, and f1-score, as well as the confusion matrix.

4.8.4.4 Evaluating Model Performance

The DT model achieved an accuracy of 92%, indicating a highly commendable level of accuracy in its performance.

Table 4.16 accuracy and classification report by decision tree for major (health and welfare)

Decision Tree Classifier –Test accuracy 0.92				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.92**	**0.90**	**0.91**	**158**
Employment	**0.92**	**0.94**	**0.93**	**192**
Accuracy			**0.92**	**350**
Macro avg	**0.92**	**0.92**	**0.92**	**350**
Weighted avg	**0.92**	**0.92**	**0.92**	**350**

From the classification table above, the precision, recall and f1-score were 92%,90%, and 91% respectively for unemployment while for employment was 92% for precision and 94% for f1-score, while it was 93% for recall.

4.8.4.5 Confusion Matrix

The following confusion matrix showcases the distribution of accurate and inaccurate predictions performed by the model. It delineates four crucial combinations of actual and predicted values:

Figure 4.38 Confusion matrix result of decision tree for major (health and welfare)

Figure 4.38 reveals the model's predictions for 350 test cases: 192 classified as employed and 158 as unemployed. A closer look at the confusion matrix shows accurate identification 142 unemployed individuals, with only 16 misclassified as employed. Moving to the second row,

180 employed instances are correctly classified, while 12 instances from the employed class are incorrectly predicted as unemployed.

4.8.4.6 KNN

In this section, a labor status classification model was constructed utilizing the KNN algorithm. The KNN model was employed to classify the data based on a collection of independent variables, with the dependent variable restricted to two values (0 and 1).

4.8.4.7 Evaluating Model Performance

The KNN model shows a good performance with an 85.4% accuracy.

Table 4.17 Accuracy and classification report by KNN for major (health and welfare)

KNN Classifier –Test accuracy 0.854				
	Precision	**Recall**	**F1-Score**	**Support**
Unemployment	**0.82**	**0.87**	**0.84**	**158**
Employment	**0.89**	**0.84**	**0.86**	**192**
Accuracy			**0.85**	**350**
Macro avg	**0.85**	**0.86**	**0.85**	**350**
Weighted avg	**0.86**	**0.85**	**0.85**	**350**

From the classification table above, the precision values were 82% for unemployment, while 89% for employment, compared the recall evaluation metric was 87% for employment and 84% for unemployment. In other side the f1-score values were 86% for employment and 84% for unemployment.

4.8.4.8 Confusion Matrix

After running the above confusion matrix code, the following tabular form shown below consists of 4 different combinations of actual and predicted values.

Figure 4.39 Confusion matrix result for KNN for major (health and welfare)

Figure 4.39 reveals that in a test dataset of 350 individuals, 192 were categorized as employed and 158 as unemployed. Analyzing the confusion matrix. Examining the confusion matrix results, it's evident that in the first row representing the unemployed class, 142 individuals are accurately classified as employed, but 16 individuals from the unemployed class are mistakenly classified as employment. Moving to the second row, 180 individuals are correctly classified as employment, while 12 individuals from the unemployment class are incorrectly labeled as unemployed.

4.8.4.9 Evaluating Classification Algorithms

A comparative performance analysis was conducted among three machine learning algorithms, evaluating metrics such as accuracy, precision, recall, f1-score, according to this analysis, the RF classifier exhibited superior performance compared to DT, and KNN models, as indicated in Figures 4.40- 4.43.

In particular, Figures 4.40 depicted the accuracy of the proposed labor force status classification model using the three classifiers: RF, DT, and KNN. Notably, RF achieved the highest accuracy among all classifiers, with DT also displaying relatively strong accuracy compared to the other models.

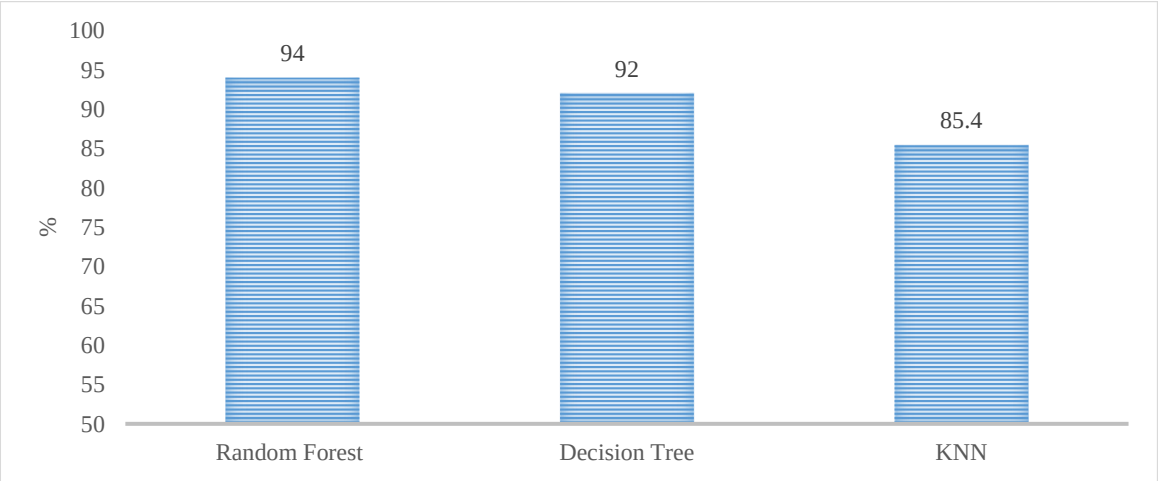


Figure 4.40 Accuracy of three classifiers for major (health and welfare)

Figure 4.41 illustrates the precision of three machine learning models. Upon fine-tuning parameter values, it becomes evident that RF attains the highest precision, while the KNN classifier exhibits the lowest precision specifically in predicting the employment and unemployment class. Regarding the unemployment class, RF and DT demonstrates a precision of 95%, outperforming KNN models.

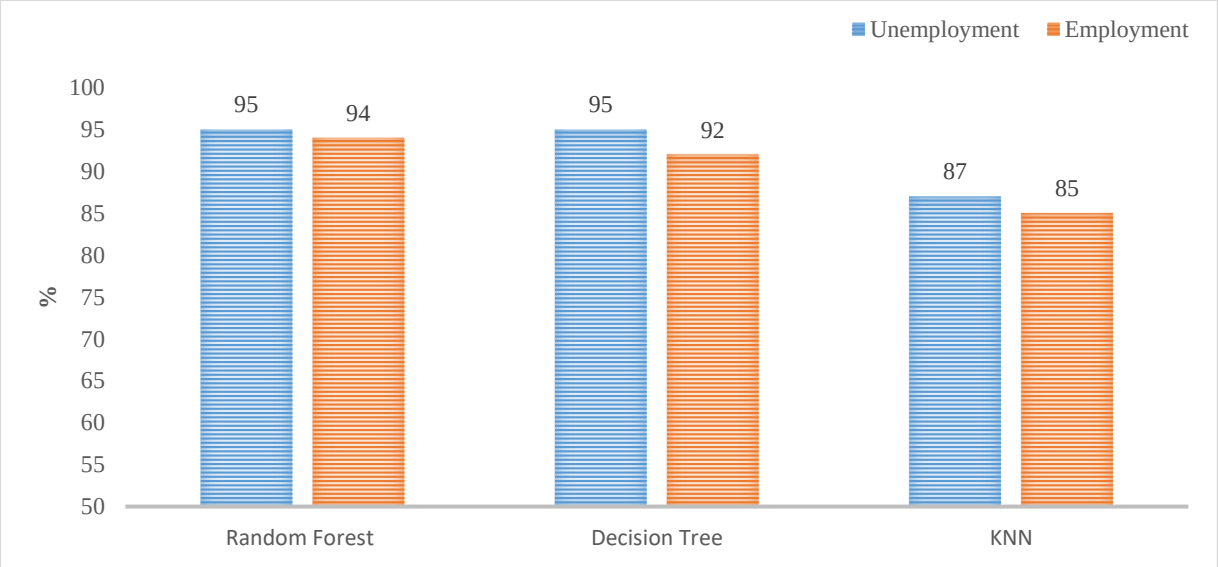


Figure 4.41 The precision of the three classifier in labor status classes for major (health and welfare)

In Figure 4.42, the recall of the three machine learning models (RF, DT, and KNN) using the test dataset is depicted. It's apparent that the RF and DT models outperforms the KNN model for employment classification while it was 95%, %91 and 92% for RF, DT, and KNN respectively for classification unemployment.

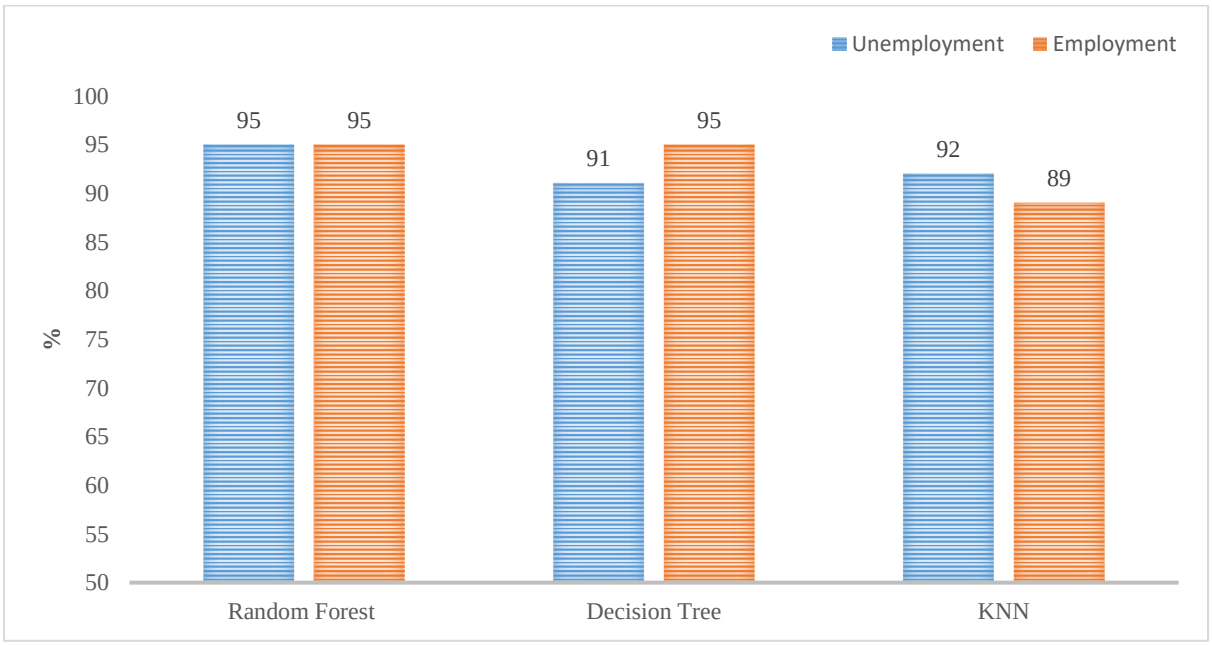


Figure 4.42 Recall of the three classifier in labor status classes for major (health and welfare)

In Figure 4.43, the f1-score performance of the three classifiers models (RF,DT,KNN) is presented. Notably, RF achieves the highest f1-score in both classes compared to the other models, attaining values of (94%, 95%) for unemployment and employment, respectively. However, it's noteworthy that the f1 scores of the DT surpass KNN classifier.

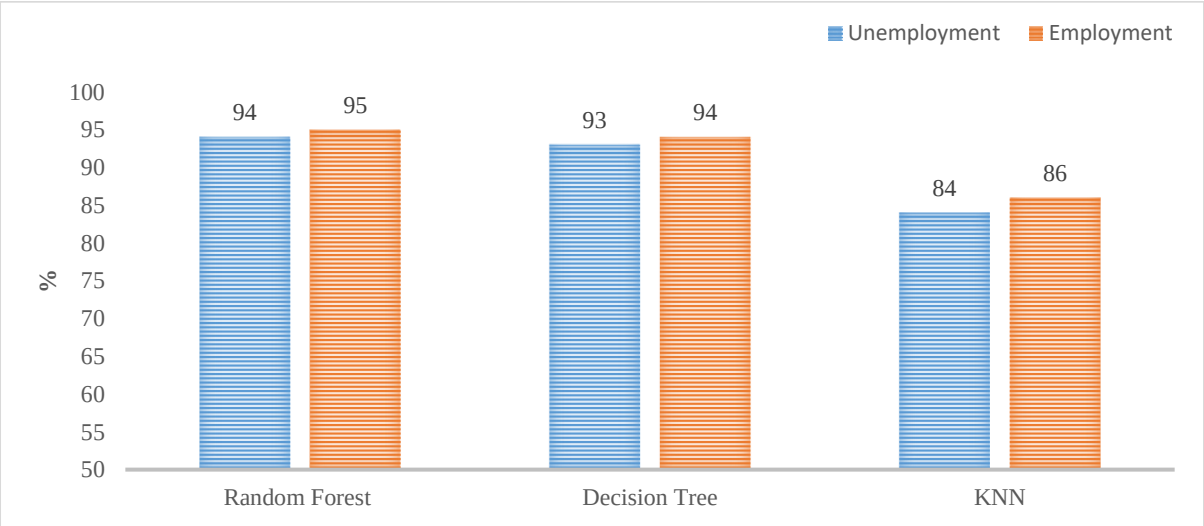


Figure 4.43 F1-Score of the three classifier in labor status classes for major (health and welfare)

4.10 Discussion of Results

The main objective of this study was to create a model for classifying the labor force among graduates in Palestine. The experiments were conducted using 80% (9558 samples) for training data, while 20% (2390 samples) for testing data. Algorithms (RF, DT, XGBoost, KNN, and AdaBoost) each experienced separate testing and training phases on shared datasets to gauge their individual performance.

While RF emerged with a stellar 95.8% accuracy on the test data, DT put up a strong showing with 93.4%. XGBoost, KNN, and AdaBoost followed at 91.8%, 88%, and 76.1%, respectively. Regarding recall and F-1 results, where it scored as (96%, 95%) for unemployment and employment classes to RF, followed by DT (94%, 93%), XGBoost (93%, 91%), KNN (90%, 84%), and AdaBoost (78%, 73%), for unemployment and employment classes, respectively. The RF model demonstrated superior performance across both classes.

Based on the experimental outcomes, the RF model emerges as the recommended choice for determining labor force status due to its superior performance across multiple evaluation metrics, particularly accuracy and recall, in classifying both unemployment and employment categories.

Through this study, among the 18 factors/characteristics that were chosen to pinpoint determine the factors that influence the determination of unemployment factors in Palestine, there were some characteristics in many studies that were important, such as (gender, marital status, specialization, age, region, status of labor) as it is in studies [16] [17] [19] [20],[21],[22],[23],[24],[29]

On the other hand, there were many studies that recommended using a RF tree to classify employment status (unemployed, employed) as studies [27],[32],[34],[38], and it is consistent with our current study, although each study has its own case that depends on the nature of the independent variables it relies on and the sample it relies on.

However, when contrasted with the studies [20], [22], [23], which categorized the labor force state into employment and unemployment using dependent and independent variables—some of which align with our study—the classification outcomes yielded an accuracy of no more than 83.5% through methods like DA, MRL, and ANN. Nevertheless, employing RF provided more precise results along with the utilization of additional derived variables.

4.11 Challenges and Limitations

One of the challenges encountered in this study is the insufficient availability of data at a granular level for each field of specialization. Consequently, several specializations were amalgamated. For instance, scientific disciplines within the College of Science (such as mathematics, biological, environmental studies, physical sciences, and statistics)) were consolidated under the broader category of "Natural Sciences, mathematics, and statistics." Likewise, all engineering specializations were grouped under "Engineering, manufacturing and construction" Additionally, certain specializations were merged due to limited graduate cases.

Furthermore, detailed data regarding graduates is lacking, as the conducted survey by the PCBS gathers extensive information about employment aspects but lacks some specific details.
Another major obstacle to understanding the situation of Palestinian graduates is the absence of specialized surveys, compounded by the weaknesses of the Ministry of Education's tracking system.

In addition the insufficient data for graduates in certain specializations, we couldn't categorize this data based on specialization.

Chapter Five: Conclusion, Recommendations, and Future working

5.1 Introduction

Despite boasting top-tier literacy rates (around 3%), Palestine grapples with a paradox: university graduates struggle to find jobs. This disconnect reveals a pressing need to adapt education to bridge the gap between learning and career success.

A significant segment of the population holds university degrees, spanning diverse specializations from intermediate diploma bachelor's to doctoral levels. However, many students find themselves jobless upon graduation, with estimates suggesting an unemployment rate approaching 50%. This stark reality poses a grave threat to individual futures, causing frustration and burdening families who have invested heavily in their children's education.

Ministry of Education and Higher Education diligently tracks student enrollment, graduation, and specialization details, while the PCBS conducts regular the survey of labor force. This wealth of information presents a valuable opportunity for proactive intervention.

Here's where machine learning steps in. By applying algorithms to this extensive data, researchers can extract crucial insights into the factors driving graduate unemployment. This knowledge empowers policymakers to develop targeted solutions, addressing systemic issues and guiding students towards in-demand fields with promising job prospects.

In essence, leveraging machine learning to navigate the complex issue of graduate unemployment can significantly benefit both individual families and the Palestinian economy at large. Equipping young graduates with vital skills and navigating them towards well-matched career paths isn't just policy, it's unlocking the promise of Palestinian education. This paves the way for a vibrant future for all, especially for our demographic cornerstone - the young graduates.

This chapter includes the results of the study, which includes three main sections: conclusions and recommendations. In addition to future working.

5.2 Conclusion

During the study, crucial data was sourced from the labor force survey conducted by the Central Agency on a quarterly basis, yielding a total dataset of 11,948 records encompassing both male and female graduates. Subsequently, essential preprocessing measures were applied to refine this dataset, ultimately preparing 11,948 records for experimentation. The main objective of the experiment was to create a classification model utilizing five different algorithms (RF, DT, XGBoost, KNN, and AdaBoost) for all graduates, Another classification of data is based on the major (specialization) by applying three algorithms: RF, DT, and KNN.

In Precision and sensitivity took center stage in the model's evaluation, meticulously measured using a confusion matrix. This rigorous assessment was enabled by Python's versatility within the Anaconda environment.

Five algorithms were tested on data split between training and testing to build a model for classifying Palestinian graduates in the labor force. RF achieved the highest accuracy (95.8%) with strong recall and F-1 scores for both unemployment and employment classes, outperforming DT (93.4%), XGBoost (91.8%), KNN (88%), and AdaBoost (76.1%). Notably, RF excelled in both accuracy and classification of employment status, making it the most effective model for this task.

Another test in this study focused on classifying graduates based on their university majors. The classification involved determining the work status (employed or unemployed) of university graduates, considering their specialization and various personal characteristics. Three machine learning models RF, decision tree, and KNN were applied to data from four selected majors: Business, Administration & Law; Engineering, Manufacturing & Construction; Education; and Health & Welfare. RF surpassed both the DT and KNN, achieving the highest accuracy results of 94.9%, 94.7%, 94.1%, and 94% for each specialty, respectively.

The results obtained from this experimental undertaking established that machine learning algorithms can proficiently classify the employment status of graduates. These algorithms assist in crafting predictive models that demonstrate a satisfactory level of accuracy, especially regarding specific attributes. Consequently, universities and unions in Palestine could adopt machine learning models to identify practical instances of graduates based on specific characteristics, notably their university specialization.

However, considering the limitations posed by the quality and scale of the dataset utilized, it becomes imperative to expand the dataset's size and augment attribute diversity. This necessity arises due to the strong association between unemployment and a country's macroeconomic indicators. Addressing unemployment necessitates boosting metrics such as national domestic product, foreign investment, and other macroeconomic indicators.

The researcher firmly believes that the model's comprehensiveness could have been augmented by incorporating an additional dataset with a larger size and diverse characteristics. The expansion of the feature set through the utilization of a supplementary dataset encompassing a wider range of variables has the potential to significantly improve model performance. The subsequent section presents the recommendations derived from the study's findings.

5.3 Recommendation

It is advisable to gather additional specialized surveys focusing on graduates, encompassing their characteristics and the specifics of their specializations. Moreover, there is a need to establish an updated and easily accessible database containing information about graduates and their respective specializations.

In addition its important to develop a new specializations in vocational education that meet the market need in cooperation with the Ministry of Labor, Palestinian Employment Fund, in cooperation with the Ministry of Higher Education. It is also the responsibility of Palestinian universities must assess graduates' realities, including their quantity and the unemployment rates specific to each field of study, as amassing a high volume of graduates without securing employment is a significant concern.

Integrating twenty-first century skills into higher education curricula is crucial, to empower Palestinian students for success in the dynamic global and local job market, this will foster active learning and enhances their employability across both general and higher education sectors.

It is very important to activate the student tracking system to leverage extensive data and enhance understanding of the labor market and its attributes by the ministry of higher Education.

5.4 Future Working

In the future, efforts will focus on integrating labor force data from PCBS regarding graduate unemployment with the Ministry of Education's databases on graduate and student numbers, using ID numbers for cross-referencing, through this comprehensive linkage, we aim to extract unemployment rates across different specializations. Furthermore, it will enable us to correlate this data with chosen dependent variables in surveys, thus creating a recommendation system beneficial to policymakers in universities, colleges, relevant ministries, and professional unions across various fields. Ultimately, this integrated platform will serve as a foundational resource for incoming university students.

Through another project implemented by the Ministry of Education to track graduates, we will try to link it to the labor force survey data, to create a comprehensive, modern, and timely database on the Palestinian labor market. This resource will facilitate more effective planning of government policies and programs.

References

[1] Palestinian Central Bureau of Statistics. (2023). The situation of youth in Palestine, 2023. [online] Available at: https://www.pcbs.gov.ps/post.aspx?lang=en&ItemID=4561 (Accessed 20 March 2024).

[2] Ministry of Higher Education & Scientific Research. (2023). Higher education statistical yearbook for scholastic year 2022/2023. [online] Available at: https://www.mohe.pna.ps/services/statistics(Accessed 01 March 2024).

[3] Davidescu, A.A., Apostu, S.A. and Paul, A. (2021). Comparative analysis of different univariate forecasting methods in modelling and predicting the Romanian unemployment rate for the period 2021–2022. Entropy, 23(3), pp.325-356.

[4] Palestinian Central Bureau of Statistics. (2022). The results of the labour force survey,2022 [online] Available at: https://pcbs.gov.ps/site/512/default.aspx?lang=en&ItemID=4421 (Accessed 01 March 2024).

[5] Palestinian Central Bureau of Statistics. (2023). The Palestinian economy during 2022, as well as the economic forecasts for the year 2023. [online] Available at: https://www.pcbs.gov.ps/post.aspx?lang=en&ItemID=4394) (Accessed 12 March 2024).

[6] Palestine Economic Policy Research Institute (MAS). (2022). Prospects for development in palestine 2021: weathering the storm, mobilizing together. [online] Available at: https://mas.ps/cached_uploads/download/2022/04/22/development-report-eng-1650653647.pdf (Accessed 02 March 2024).

[7] Husin, N.A., Rasli, S., Kumar, M.S.G. and Suppiah, G. (2021). Unemployment crisis among fresh graduates. American International Journal of Social Science Research, 10(1), pp.1-14.

[8] Oniore, J.O., Bernard, A.O. and Gyang, E.J. (2015). Macroeconomic determinants of unemployment in Nigeria. International Journal of Economics, Commerce and Management, 3(10), pp.215-230.

[9] Fernando, R. L. S., Kularathna, E. A. I., & Kumarasinghe, I. D. C. D. (2022). Employability of management graduates of state universities in Sri Lanka: problems and prospects. International Journal of Governance and Public Policy Analysis, 4(01), pp.66-90.

[10] Asif, A., Faraz, A.A. and Makhdoom, W. (2022). A statistical study on unemployment rate and its resultant consequences in fresh graduates of Punjab university. Journal of Statistics, 26, pp.34-43

[11] Huu, A., Nhat, T., Thanh, T. and Hoàng, G. (2022). The reason why the unemployment rate of college graduates is increasing: Case study in Ho Chi Minh City, Vietnam. International Journal of Multidisciplinary Research and Development, 9(1), pp.19-25.

[12] Hossain, M.I., Yagamaran, K.S.A., Afrin, T., Limon, N., Nasiruzzaman, M. and Karim, A.M. (2018). Factors influencing unemployment among fresh graduates: A case study in Klang Valley, Malaysia. International Journal of Academic Research in Business and Social Sciences, 8(9), pp.1494-1507.

[13] Mohd, J., Kadir, A., Naghavi, N. and Abdul, N.A.A. (2020). Unemployment among graduates-Is there a mismatch. International Journal of Asian Social Science, 10(10), 583-592.

[14] Dănăcică, D.E., Babucea, A.G., Paliu-Popa, L., Bușan, G. and Chirtoc, I.E. (2023). The nexus between higher education and unemployment—evidence from Romania. Sustainability, 15(4), pp.3641-3661.

[15] Leopold, T.A., Ratcheva, V. and Zahidi, S. (2017). The future of jobs and skills in Middle East and North Africa: Preparing the region for fourth industrial revolution. In World Economic Forum.

[16] Dimova, R., Elder, S. and Stephan, K. (2016). Labour market transitions of young women and men in the Middle East and North Africa. International Labour Office.

[17] Cordón-Lagares, E., García-Ordaz, F. and García-del-Hoyo, J.J. (2022). The transition from higher education to first employment in Spain. European Journal of Education, 57(4), pp.659-673.

[18] Cordón-Lagares, E., García-Ordaz, F. and García-del-Hoyo, J.J. (2022). The transition from higher education to first employment in Spain. European Journal of Education, 57(4), pp.659-673.

[19] Arslan, H. and Senturk, I. (2018). Individual determinants of unemployment duration in Turkey. Eskisehir Osmangazi Universitesi Iibf Dergisi-Eskisehir Osmangazi University Journal of Economics and Administrative Sciences, 13(1), pp.113-127.

[20] Niragire, F. and Nshimyiryo, A. (2017). Determinants of increasing duration of first unemployment among first degree holders in Rwanda: a logistic regression analysis. Journal of Education and Work, 30(3), pp.235-248.

[21] Shekh Eid,F.S. (2014). Using discriminant analysis and multinomial logistic regression in the classifying the employment status of graduates in Palestine. Master's Thesis. Al Azhar University-Gaza

[22] Migdad M.(2011). Unemployment among the graduates of the education system in Palestine. Journal of Al Azhar University Gaza (Human Sciences),13(1),pp.25- 50.

[23] Abdalla, M. and Salama, H. (2015). Using Classification Methods in Identifying the Labor Force Categories in Palestine. Journal of Al Azhar University-Gaza (Natural Sciences), 2015, 17(1), pp.1-22.

[24] Mohamed, A.M., Abdel-Fattah, M.A. and Aldirawi, A.S.M. (2021). A comparison between classification statistical models and neural networks with application on Palestine data. J. Math. Comput. Sci., 11(4), pp.3916-3926.

[25] Hajjaj, O. (2018). The economic intervention strategy of the international non-governmental organizations in palestine. a case study: the short-term employment projects. Uniwersytet śląski. Edited by G. Balint et al., 7(1), pp. 343–354.

[26] AlKhayyat, S.L. and Alghamdi, H.M. (2022). A comparison between artificial neural networks and logistic regression to determine the general factors that effect on unemployment for Saudi women. Applied Mathematical Sciences, 16(1), pp.27-35.

[27] Azziaty, Nor & Tan, Kian & Lim, Chen Kim. (2017). Predictive analysis and data mining among the employment of fresh graduate students in HEI. AIP Conference Proceedings. 1891(1),pp. 020007-1 – 020007-6.

[28] Casuat, C.D. and Festijo, E.D. (2019), December. Predicting students' employability using machine learning approach. IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pp. 1-5.

[29] Pauzi, W.N.D.W., Hasan, H. and Mahmud, Z. (2021), November. Supervised and unsupervised data mining techniques on employability of public higher learning institute graduates in Malaysia. Journal of Physics: Conference Series, 2084(1), pp. 1-16.

[30] Mohamed Saouabi and Abdellah Ezzati. (2019). Prediction Model for Employability in Morocco Using Data Mining Techniques. Journal of Engineering and Applied Sciences, 14(4), pp.1690-1694.

[31] Alheadary, W.G. (2023). Controlling employability issues of computing graduates through machine learning-based detection and identification. Engineering, Technology & Applied Science Research, 13(3), pp.10888-10894.

[32] Jantawan, B. and Tsai, C.F. (2013), October. The application of data mining to build classification model for predicting graduate employment. International Journal of Computer Science and Information Security, 11(10), pp. 1-7.

[33] Maaliw, R.R., Quing, K.A.C., Lagman, A.C., Ugalde, B.H., Ballera, M.A. and Ligayo, M.A.D. (2022), January. Employability prediction of engineering graduates using ensemble classification modeling. IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 288-294

[34] Panityakul, T., Suriyaamorn, W., Chinram, R. and Kongchouy, N. (2022). Classification models for employability of statistics and related fields graduates from Thailand universities. Computer Science, 17(1), pp.365-375.

[35] He, S., Li, X. and Chen, J.(2021), May. Application of data mining in predicting college graduates employment. 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 65-69.

[36] W Aziz, M.T.R.A. and Yusof, Y. (2016), August. Graduates employment classification using data mining approach. AIP conference proceedings,1761(1).

[37] Piad, K.C. (2018). Determining the Dominant Attributes of Information Technology Graduates Employability Prediction Using Data Mining Classification Techniques. Journal of Theoretical & Applied Information Technology, 96(12), pp.3780-3790.

[38] Jantawan, B. and Tsai, C. (2014). A classification model on graduate employability using bayesian approaches: a comparison. International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization), 3297(6), pp.4584-4588.

[39] Casuat, C.D., Festijo, E.D. and Alon, A.S. (2020). Predicting students' employability using support vector machine: a smote-optimized machine learning system. International Journal, 8(5), pp.2101-2106.

[40] Crisostomo, A.S., Encarnacion, R. and Al Balushi, S. (2023), May. A data mining approach to construct classification model for predicting tourism graduates employability. 2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA),pp. 162-168.

[41] Moumen, A., Bouchama, E.H. and EL IDIRISSI, Y.E.B. (2020), December. Data mining techniques for employability: Systematic literature review. 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS),pp. 1-5.

[42] Moumen, A. and Mejjad, N. (2021). Graduates employability: an exploratory literature review. SHS Web of Conferences, 119, pp. 1-8.

[43] Kumar, M.S. and Babu, G.P. (2019). Comparative study of various supervised machine learning algorithms for an early effective prediction of the employability of students. Journal of Engineering Sciences, 10(10), pp.240-251.

[44] UNESCO. (2015). International Standard Classification of Education Fields of Education and Training 2013 (ISCED-F 2013): Detailed Field Descriptions. UNESCO Institute for Statistics, Available online: http://dx.doi.org/10.15220/978-92-9189-179-5-en (Accessed 03 March 2024).

[45] Singh, D. and Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, pp.105524-05547

[46] Polatgil M. (2022). Investigation of the effect of normalization methods on ANFIS success: Forestfire and Diabets Datasets'.International Journal of Information Technology and Computer Science, 14(1), pp. 1–8.

[47] Sahoo, K., Samal, A.K., Pramanik, J. and Pani, S.K. (2019). Exploratory data analysis using python. International Journal of Innovative Technology and Exploring Engineering, 8(12), pp.4727-4735

[48] Yeşilkanat, C.M. (2020). Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos, Solitons & Fractals, 140.

[49] Sharma, A. Decision Tree vs. Random Forest—Which Algorithm Should You Use? Analytics Vidhya, (12 May 2020). [online] Available at: online: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ (Accessed 07 December 2023).

[50] Gupta, I., Sharma, V., Kaur, S. and Singh, A.K. (2022). PCA-RF: an efficient Parkinson's disease prediction model based on random forest classification. arXiv preprint arXiv:2203.11287

[51] Resende, P.A.A. and Drummond, A.C. (2018). A survey of random forest based methods for intrusion detection systems. ACM Computing Surveys (CSUR), 51(3), pp.1-36.

[52] Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R. and Kundu, S. (2018). Improved random forest for classification. IEEE Transactions on Image Processing, 27(8), pp.4012-4024.

[53] Tarik, A., Aissa, H. and Yousef, F. (2021). Artificial intelligence and machine learning to predict student performance during the COVID-19. Procedia Computer Science, 184, pp.835-840.

[54] Charbuty, B. and Abdulazeez, A. (2021). Classification based on decision tree algorithm for machi learning. Journal of Applied Science and Technology Trends, 2(01), pp.20-28

[55] Thaher, T., Zaguia, A., Al Azwari, S., Mafarja, M., Chantar, H., Abuhamdah, A., Turabieh, H., Mirjalili, S. and Sheta, A. (2021). An enhanced evolutionary student performance prediction model using whale optimization algorithm boosted with sine-cosine mechanism. Applied Sciences, 11(21), pp.10237-10272.

[56] Abu Alfeilat, H.A., Hassanat, A.B., Lasassmeh, O., Tarawneh, A.S., Alhasanat, M.B., Eyal Salman, H.S. and Prasath, V.S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big data, 7(4), pp.221-248.

[57] Sneha, N. and Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data, 6(1), pp.1-19.

[58] Cunningham, P. and Delany, S.J. (2021). K-nearest neighbour classifiers-a tutorial. ACM computing surveys (CSUR), 54(6), pp.1-25.

[59] Khan, M.A.R., Paul, A.R., Rahman, F., Akter, J., Sultana, Z. and Rahman, M. (2023). Appropriate Job Selection Using Machine Learning Techniques.

[60] Kaur, H. and Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, 18(1/2), pp.90-100.

[61] Ying, C., Qi-Guang, M., Jia-Chen, L. and Lin, G. (2013). Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica, 39(6), pp.745-758.

[62] Wang, Y., Pan, Z., Zheng, J., Qian, L. and Li, M. (2019). A hybrid ensemble method for pulsar candidate classification. Astrophysics and Space Science, 364(4), pp.1-13.

[63] Chen, T. and Guestrin, C. (2016), August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[64] Xu, J., Zhang, Y. and Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. Information Sciences, 507, pp.772-794.

[65] Bowers, A.J. and Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. Journal of Education for Students Placed at Risk (JESPAR), 24(1), pp.20-46.

[66] Ishak, A., Siregar, K., Ginting, R. and Afif, M. (2020), December. Orange software usage in data mining classification method on the dataset lenses. In IOP Conference Series: Materials Science and Engineering,1003(1).

الملخص

تعاني فلسطين، كمعظم الدول النامية والعربية، من مشكلة البطالة. وترتبط هذه المشكلة في الإقتصاد الفلسطيني بالعديد من المتغيرات الإقتصادية والإجتماعية، وفي مقدمتها النمو. بلغ معدل البطالة بين المشاركين في القوى العاملة في فلسطين 24.4% عام 2022، مقارنة مع 23.0% عام 2015. علاوة على ذلك، تواجه نسبة كبيرة من الشباب الفلسطيني والخريجين فرص عمل محدودة، مما يؤدي إلى الإحباط وخيبة الأمل. لا تؤدي المستويات المرتفعة من البطالة بين الشباب إلى تعريض الفرص الفردية للخطر فحسب، بل تشكل أيضاً تحدياً مجتمعياً أوسع نطاقاً، مما قد يساهم في الاضطرابات الاجتماعية. وفقاً للأرقام الصادرة عن الجهاز المركزي للإحصاء الفلسطيني، فقد بلغ معدل البطالة للأفراد 20–29 عاماً الحاصلين على الدبلوم المتوسط أو البكالوريوس فأعلى أكثر من 48%. وهذا يعني أن نصف الخريجين عاطلون عن العمل.

ونظراً لأهمية قضية البطالة بين الخريجين في فلسطين، تعرفنا في هذا البحث على واقع البطالة بين الخريجين، وعملنا على تطبيق تقنيات التعلم الآلي لتصنيف بيانات القوى العاملة لتحديد واقع البطالة بين الخريجين في فلسطين من حيث حالة العمل (عامل، عاطل عن العمل) وتحديد أثر بعض المتغيرات على الحالة الوظيفية مثل (الجنس، العمر، التخصص الجامعي، المنطقة، المحافظة،...)، بالإضافة إلى ذلك ومن اجل معرفة أهمية التخصص الجامعي تم تطبيق تقنيات التعلم الآلي على بيانات خاصة ببعض التخصصات الجامعية، حيث تم اختيار اربع تخصصات جامعية. حيث تستحق البطالة اهتماماً خاصاً لدى خريجي الجامعات ومجال دراستهم.

تم إنشاء نموذج تصنيف باستخدام خمس خوارزميات مختلفة – RF وDT وXGBoost وKNN وAdaBoost لبيانات الخريجين بشكل عام، وتم تطبيق ثلاث خوازميات لتصنيف البيانات حسب التخصص الجامعي.

أظهرت الغابة العشوائية (RF) دقة ممتازة بلغت 95.8% في بيانات الاختبار، وقدمت شجرة القرار (DT) دقة عالية ايضاً بنسبة 93.4%. تليها XGBoost، وKNN، وAdaBoost بنسبة 91.8%، و88%، و76.1% على التوالي لبيانات القوى العامة الخاصة بالخريجين. أما على مستوى تصنيف الخريجين حسب التخصص فقد اظهرت الغابة العشوائية دقة عالية في تصنيف حالة الخريجين وذلك للأربع تخصصات التي تم اختيارها (الأعمال والإدارة والقانون)، (الهندسة

والتصنيع والبناء)، (والتعليم)، (والصحة والرفاه). حيث تجاوزت RF كلاً من شجرة القرار و KNN، وحققت أعلى نتائج دقة بلغت 94.9%، و94.7%، و94.1%، و94% لكل تخصص على التوالي.