UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

Career prediction Model for computing college Students in Kenya

BY:

SAMWEL NAI

P52/35554/2019

SUPERVISOR:

DR. WANJIKU NG'ANG'A

A project report submitted in partial fulfillment of the requirements of the Degree of Masters of Science in Computational Intelligence of the University of Nairobi

MAY 2022

Declaration

Student declaration

The project report is my unique effort submitted to University of Nairobi as a requirement of Master of Science in Computational Intelligence and has not been used for education purpose in any other institution of higher learning.

SIGN:……………………………………. DATE: 26-07-2022**………………………………………**

SAMWEL NAI, UNIVERSITY OF NAIROBI

Registration Number**:** P52/35554/2019

Supervisor Declaration

The project report was submitted with my consent as the University of Nairobi supervisor.

SIGN ………………………………….. DATE03 August 2022**…………………………………………**

DR. WANJIKU NG'ANG'A, LECTURER - UNIVERSITY OF NAIROBI

Dedication

This report is devoted to my lovely wife and kinfolk, who have always been there for me when I needed them and who continue to love, support, and encourage me. I dedicate it, above all, to God, who has blessed me abundantly.

Acknowledgements

My appreciation to the All-Powerful God for having placed me here. You have done so much for me, Lord. I shouldn't be surprised to feel relieved! Amen! I'm singing with joy!

Second, I want to sincerely thank and appreciate Dr. Wanjiku Ng'ang'a, my supervisor, for her advice and criticism that was constructive.

Last but not least, I express my thankfulness to my coworkers as well as colleague Nicholus Mutua towards their unwavering support as well as inspiration ever since I was accepted into the postgraduate program.

Abstract

Choosing an acceptable professional career route is one of the most essential decisions that students must make in our society today. The increasing number of alternative jobs and prospects in computing, makes this decision more challenging. The goal of this study was to identify computing career parameters, compare and contrast Random Forest and Naive Bayes (NB) supervised machine learning algorithms and then develop a prototype. This objectives were accomplished through CRISP-DM using the Kaggle repository data 4 ver1 dataset. Computing professional parameters for prediction include professional skills and abilities, CPGA, communication skills, analytical skills, team player, personal interest and professional experience.

The algorithms for predicting careers have been thoroughly examined. Due to their excellent prediction accuracy, Random forest and Nave Bayes were identified for career prediction system. The model was developed using five, ten, fifteen and nineteen attributes. This study examined the percentage F1 score, recall, precision and accuracy of these two cutting-edge supervised learning approaches. Testing and training of both algorithms was done using the same datasets with varied number of attributes. The findings demonstrated that the Random Forest algorithm outperformed the Naïve Bayes algorithm that had an accuracy of 89.885% as well as higher recall, precision in addition F1 score. There was the gradual increase in all performance metrics using different number of attributes until it reaches the point of stagnation.

A prototype based on the Random Forest method was developed. The prototype developed was evaluated for accuracy in career prediction for computing college students. This prototype can be used by computer graduates and Human Resource managers to make more accurate and consistent prediction on computing careers.

Table of Contents

List of Figures

List of Tables

List of Abbreviation

CRISP-DM: Cross Industry Standard Process for Data Mining

CSV - Comma Separated Values

NB - Naïve Bayes

FN - False Negative

FP - False Positive

CDDQ - Decision-Making Difficulties Questionnaires

RF - Random Forest

ANOVA - Analysis of Variance

LDA - Linear Discriminant Analysis

HTML - HyperText Markup Language

GUI - Graphical User interface

KNN – K-Nearest Neighbor

CGPA - Cumulative Grade Point Average

1.0 CHAPTER ONE: INTRODUCTION

1.1 Background Information

Computing careers have grown in the past decade. Most Kenyan universities are now offering information technology(IT), computer information systems(CIS) as well as computer science(CS) as a degrees. The path to career specialization after taking these degrees is not well defined. College students undertaking degree programs will require guidance on the careers that best suit them .

College education defines the career path of a student. According to (Mualuko, 2007), 47% of primary schools pupils are able to join secondary level and 27% of secondary school students get access to post-secondary education (technical colleges and universities). Computing college students who take the right career path will significantly contribute in the economic development and Psychosocial well-being of its citizens (Kabari & Agaba, 2019). Economically, there is an increase in the real gross domestic product per capita. Psychosocial well-being is improving the quality of life which involves living healthier and longer by balancing the emotional factors , social and physical components (Eiroa-Orosa, 2020). Today's market demands computing experts in all pillars of Kenya's vision 2030. This has led to creation of job opportunities in the computing field. Career guidance is required in order to produce employees who will fit in the ten key sectors of vision 2030 strategies. The workforce in Kenya is below standard as expected and aspiration towards Kenya's vision 2030. Therefore, Career guidance should be balanced and skewed towards majority of computing students without any discrimination. Human potential can only be unleashed through proper career advice in the education process. In as much as college students are mature to understand the career path, there is a discrepancy on their expectations and available jobs in the market (Macharia, 2019).

In digital technology, it is not easy to define specialty in computing because technology is dynamic and continues to redefine itself with new improvements. A specialty in information technology and computing, however, can be categorized into Database Administrators, Information Security Analysts, Solution Architects, Computer Systems Analysts, Computer Network Architects, Computer Programmers and Developers, Design and UX, E-Commerce

Analysts, and Network Security Engineers, per the US Bureau of Labor Statistics. (Wong & Kemp, 2017).

In order to create career counseling models, many algorithms have been utilized. Some of these algorithms are Decision Tree(DT), Linear Regression(LR), K-Nearest Neighbor(KNN) and Naive Bayes(NB) (Gerhana, Fallah, Zulfikar, Maylawati, & Ramdhani, 2019).

The decision on one’s career is a key element in a person’s life. There exist many challenges encountered by college students, from wrong career selection as it opens the door for lifelong consequences. Reliance on human beings as the guiders and counselors whose reasoning is based on human experience might vary from one individual to another (Kazi Afaq, Sharif, & Ahmad, 2017).

According to (Subahi, 2018), one of the most crucial choices a graduate must make is selecting the first suitable job in any field linked to computing. Making this decision has become difficult for graduates due to the latest surge in profession routes and work prospects in computer interrelated sectors. The productivity of graduates in their first work will be significantly impacted and diminished as a result of choosing improper employment that does not match their talents and knowledge. This makes it essential to have a tool that can help graduates assess their skills after earning a degree in computers, allowing them to select the best entry-level positions.

In Kenya, college students are left to be guided by career counselors. Through counseling, students are counseled on how to make tough decisions, plan their career and understand their abilities. The main challenge in counseling is to engage the students in the process. According to a UK research, 45% of those over 14 received inadequate or insufficient employment guidance (Macharia, 2019). Most parents and students lack adequate information and therefore choose their career through perception of the ideal computing career (Kazi Afaq, Sharif, & Ahmad, 2017).

1.2 Statement of the problem

Long term work progression prediction is not well explored due to diversity in career trajectory and career planning. Developed models cannot support current job seekers in planning their long term career paths. The models focus on immediate steps of career movement and are informed by theoretical guidance (Michiharu, Yunqi, Thanh, Yongfeng, & Dongwon, 2022).

Discrepancy between what computing students are capable of and what is on the ground is created by counselors who focus on advising students on the career depending only on results rather than facts and potentials of the students (Nunsina & Situmorang, 2020).

According to (Nunsina & Situmorang, 2020), the algorithm used was not optimally reliable in predicting the specific fields of computing specialties due to minimal number of speciality areas considered. The three major areas considered include Software engineering, multimedia and Computer Engineering network. In a research by (Lagman, 2019; Razaque, 2017) in academic performance and career, they state that existing computing career prediction models could not assist computing college students in selecting the right career appropriately because they do not consider social and environmental factors but only academic factors.

1.3 Objectives

1.3.1 General objective

Development of a model based on Machine Learning for career prediction for college graduate students in computing specialties

1.3.2 Specific Objectives

1. To find fundamental career prediction parameters for college students in computing.
2. To identify the appropriate machine learning algorithm for career prediction by computing college students.
3. To develop an automated career prediction model based on the appropriate algorithm for college students in computing.
4. To evaluate the proposed career prediction model for accurate prediction.

1.4 Research questions

1. What are the fundamental career prediction parameters for college students in computing?
2. What are the appropriate machine learning algorithms for career prediction by computing college students?
3. How to develop an automated career prediction model based on the appropriate algorithm for college students in computing?
4. How to evaluate the proposed career prediction model for college students in computing?

1.5 Significance of the study

The project was meant toward assisting college students select a career based on their capabilities, interests, background information and performance. It is important for computing college students to have a guide on their career paths based on factual concepts. The research contributes to industrial human resource management where shortlisting and engagement of computing graduates will be guided by the model developed.

1.6 Justification

Lack of a computing specialty model for college students with high accuracy, robust and efficiency has led to computing students pursuing careers that are not of their interest, leading to psychosocial problems. This research will act as a guide to computing student where reliable career guidance will be done based on different parameters.

Career prediction in the computational job market using Artificial Intelligence, can greatly help not only to job seekers, in understanding and planning their future pathway, but also for recruiters in finding talented computing employees (Michiharu, Yunqi, Thanh, Yongfeng, & Dongwon, 2022). In human resource departments, manual shortlisting of applicants to specific computer related vacancies are used. The research will assist the department in doing the shortlisting of applicants based on their performance, interests, certification, workshops and background information. As a guiding principle, the research will enhance good time management and selection of competent computing employees for the vacancies.

Random Forest as well as Naïve Bayes algorithms have proven accuracy, robust and efficiency in prediction therefore are common. They provide an easy and fast class prediction for both the training and testing data. Where the parameters are independent and in cases of categorical input variables, the Naïve Bayes algorithm and Random Forest perform better than other algorithms (Gerhana et. at., 2019).

1.7 Scope of the study

The model will accept average cumulative grade point of the computing student and background information to predict computing students specialty.

All computing students from University, colleges and tertiary institutions will be able to use the model to predict their specialty, as well as industrial human resource managers, when selecting and shortlisting applicants for employment in computing.

1.8 Limitation of the study

Owing to limited period as well as insufficient resources, the research employed secondary data rather than gathering primary data from graduate college students.

This model will not be able to address all specialties in computing as technology is dynamic and changing every day.

The model will not be able to cater to other students in colleges who are undertaking courses that are not related to computing.

2.0 CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The section introduces fundamental categories in computing: Computer science, Computer Information Systems, and Information Technology leading to different careers in computing world. The existing prediction techniques are reviewed with major focus on the models and algorithms. A detailed analysis of predictions based on Naïve Bayes and Random forest is done which has created a research gap in addressing career prediction accurately and efficiently.

2.2 Career in computer and information technology

Computer occupations may be divided into three categories: careers in CS, IT and CIS. Careers in CS are centered on applying the mathematical concepts used to develop programs to make sure systems are operating efficiently and successfully achieving their objectives. The Information technology careers focus on addressing the installation, improvement and maintenance of computer systems, networks and databases (Wong & Kemp, 2017).

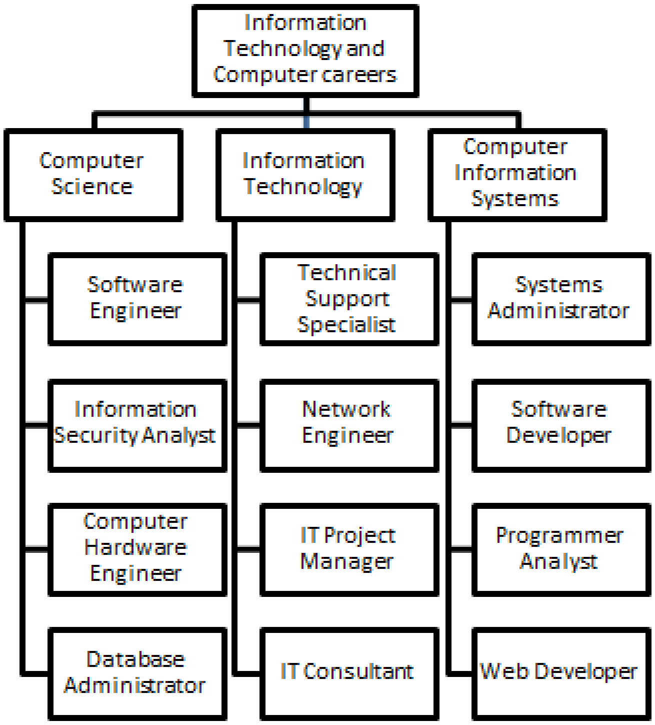


Figure 1 Computing career classification by Wong & Kemp, 2017

There are many new career options due to the gradual increase in the fields of research and technology. In computing, there are many roles or career paths like data scientist, UI developers, Business Process Analyst, database administrator, software testing, Network analyst, Network managers and so on. Each role requires basic knowledge in the required fields (Reddy, 2021).

The need for qualified IT professionals seeking effective professions in a variety of computer career disciplines increases due to the fast evolution of business innovation and technology. CS, IT , SE, IS, and management information systems (MIS) are just a few of the computing degrees that equip future professionals to enter the workforce and maintain their competency in the productiveness. Alumnae in computing programs possess wide range of hand-on-skills as well as understanding in a range of computer and information technology areas. In addition, business knowledge sets, non-practical abilities are also necessary owing to the enormous influence that Information Technology has on commercial procedures and setups (Subahi, 2018).

2.3 Career Prediction

Job competition and the emergence of post-industrial revolution, has led to career path choice becoming complex for computing college students. During the old days, family affairs played a major role in children choosing their careers. That is, the son of a blacksmith would automatically become a blacksmith. Today, one has to do career research to be able to make choices that are relevant to them and be satisfied in their workplace. Currently, most university graduates have less or no information about career opportunities to assist them in making good career choices. Career path is determined by many factors that include interests, personality, role models, cultural identity, socialization, environment they live and financial resources. These factors can be intrinsic or extrinsic or even both. (Bandura, Caprara, Barbaranelli, Pastorelli, & Regalia, 2001).

Accuracy in career pathway prediction models should emphasize on reasoning instead of similarity matching for career prediction and job recommendation. The cognitive reasoning nature of job transition enhances accuracy in optimal career prediction. The selection of a data analyst or a computer engineer is informed by experience or background in programming and logic reasoning (Michiharu, Yunqi, Thanh, Yongfeng, & Dongwon, 2022)

According to VidyaShreeram & Muthukumaravel, 2021, a student's choice of career has a significant impact on his/her life. Student career prediction is difficult because of the complications of every student's dreams and aims. Students' data provide the suggestions basing on behavioral aspects to predict the career path. Traditional methods have been used to predict students' career but take longer time to give prediction results. Therefore, student career prediction using machine learning concepts is faster and accurate. The prediction research helps students to identify their career path to adopt and human resource managers to select suitable candidates for employment.

The created Intuitive Career System uses a variety of questions to gauge a student's aptitude in addition to inquiries about their background. Based on their social media accounts, the students' personalities are evaluated using the Facebook Graph API. The model, which predicts which career best matches the student's skills and personality, is then given as the prediction. The model has an average accuracy of 77.41% in predicting aptitude, 75.41% in predicting personality and 60.09% in predicting background knowledge. Because it takes into account both personality and ability, which have an impact on professional job choices, this is a realistic approach to counseling (Rangnekar, Suratwala, Krishna, & Dhage, 2018).

According to (Ojenge & Muchemi, 2008), there is a discrepancy between what college students do as career from their natural interests. As graduates continue pursuing their career paths, it is very important for them to understand their capabilities and their interests so that they can be able to identify their relevant careers. Students' capabilities are assessed from many factors that include interests, academics, sports, skills and background information (Roy, Roopkanth, Bhavana, & Priyanka, 2018). One of the main tasks of adolescents is the choice of a career in the future. Career exploration and commitments help one develop a firm understanding of their objectives, abilities and interests. Thus, career guidance is essential to college students to help them make such decisions. This has led to the establishments of career guidance centers (NIE, et al., 2018).

The ability to reason beyond similarity matching in developed models, is important for career pathway prediction. The identification of reasoning patterns in different algorithms can help in increasing the accuracy of career prediction and job recommendations (Michiharu, Yunqi, Thanh, Yongfeng, & Dongwon, 2022).

2.4 Prediction Models

According to (NIE, et al., 2018), Decision-Making Difficulties Questionnaires (CDDQ) was used to find out decisions of student's career. It was found out that students' lack of information for career choices was the major reason. In order to make occupational decisions, students should identify their personalities and interests.

According to (Ojenge & Muchemi, 2008), expert system for career guidance, personality of the student and Aptitude test were used to predict student careers. The system proposed first analyze personality and end with the aptitude test. The personality module used rules based on Myers-Briggs Typology Indicator model. The four possible human natures used were Idealist, Rational, Artisan and Guardian. Parameters used were personality, outlook, temperament and lifestyle. For college admission, the model used grades scored in selected subjects. The prototype analysis achieved a 71% career satisfaction level.

Existing prediction models use three techniques in their predictions. They include prediction using grades, prediction using yes/no or the hybrid system which combines grade and yes/no questions (Reddy, 2021). The researcher used personality and aptitude to predict the career. For aptitude, six topics were used which include, science, Mathematics, Information Technology, History, Biology and English. The Myers-Briggs personality test was employed. According to the findings, personality and aptitude did not lead to the same career. This makes it much more important to take the two criteria into account when predicting careers. When employing aptitude, KNN and Stochastic Gradient Descent both provided accuracy figures of 73.74% and 81.035%, respectively. Stochastic Gradient Descent and KNN both exhibited accuracy rates of 73.75% and 81%, respectively, for personality categorization.

According to (Nunsina & Situmorang, 2020), where analysis of optimizing K-Nearest Neighbor in determining student careers through certainty factors was done, and an accuracy of 93.83% was obtained. With most focus on determining the student's career from analyzing the student's grades. The algorithm was not optimally reliable in predicting the specific fields of computing specialties because the researcher chose to determine three major computing specialties (programmer, computer technician and graphic design). The model used the values of selected subjects that include: software engineering, object-oriented programming, network technology, database programming, basic visual design and basic computer network. Computing being a broad and dynamic field, has many different specialties that a computing student can specialize in.

The comparison of models developed from different baseline methods was done in evaluating the Mean Average Precision(MAP) for job tittle pathway indicating the length of career prediction pathways. MAP for Linear Regression was 0.03402, RF was 0.04966, XGBoost was 0.05546, Long short-term memory (LSTM) was 0.14782, NEighborhood based Multi-Omics clustering(NEMO) was 0.15298, look-ahead(backtracking) was 0.19672 and NAOMI was 0.19672. Based on the above, the NAOMI model was developed that predicted the next N steps of career a movement to propose a new task or a future career pathway prediction using large scale real-world dataset (Michiharu, Yunqi, Thanh, Yongfeng, & Dongwon, 2022).

K-Nearest Neighbor has been used to predict candidates for fit in proper tests. K-Nearest Neighbor is a predicting method based on data by measuring close distance with a successful or unsuccessful sample. The model used grade, education, competency, assessment, talent score and past position in line with the new position. From the findings, when K=1 was more precise and accurate as compared to K=3 and K=5. Comparing True positive and False positive, K=5 had better results than K=3 and K=1 (Sampurno, Hediyantama, & Widiyati, 2019).

Decision tree have been used for classification and prediction. Decision trees generate a hierarchy chart with parent nodes and leave nodes. Algorithms are used to make decisions on different paths to be taken. C4.5 algorithm has been used for classification that is used to generate decision trees in data mining. According to (Gerhana, Fallah, Zulfikar, Maylawati, & Ramdhani, 2019), NB algorithm was 88% accurate as compared to C4.5 algorithm which was 87% accurate.

Random forest has been used to predict student career. According to (VidyaShreeram & Muthukumaravel, 2021), In comparison to support vector machines and decision trees, the RF classifier's accuracy was 93%. Python programming language was used to implement machine learning classifiers.

Naïve Bayes has been used to identify students' academic achievements and personality. The model used energy orientation, dimensions of making decisions, lifestyle and management of information. Based on personality, study habits of pupils were determined using training data. Testing data were utilized to assess accuracy of model developed. It correctly categorized 28 testing data with an accuracy of 96.42 percent (Mulyati & Setiani., 2018)

Naïve Bayes has been used to predict student eligibility in vocation schools. Interviews, competency, final exam reports and psychology were the variables used. Each variable was first transformed into ranging values and then possibility values were calculated for each variable. From a sample of 270 students and validation test of 199 students, recall was 99.3% and precision of 96.1%. The algorithm had an accuracy of 74.87%, meaning that not all students were accepted in school (Melian & Nursikuwagus, 2018).

2.5 Career Prediction Parameters

Many factors are considered when predicting a student's career path in computing, including the student's knowledge of numerous areas, specialties', programming and analytical talents, certificates, and interest in particular courses. All of these criteria play an important part in predicting the right job path (Reddy, 2021).

Career choices are influenced by an individual's surroundings, abilities, skills, and academic accomplishment. If you make the wrong decision, you may experience failure and disappointment as a result. Other drivers of profession choice have been demonstrated to be aptitude, personal circumstances, and academic accomplishment. (Kazi & Akhlaq, 2017).

A bachelor's degree in computing, any experience in and professional experience, are the most sought-after qualifications. Other non-technical abilities are teamwork, feel of responsibility and communication skills are also important for career prediction (Gruger & Schneider, 2019). Individual personalities, interests, and aptitudes, as well as other explainable and measurable attributes, give more weight in vocational selection (Nie, et al., 2018).

According to (Malik & Al-Emran, 2018), the choice of career is influenced by social factors, skills required to be successful in the career and characteristics of a computer professional. The research identified social factors as parents, teachers and counselors, friends, job opportunities, job image, personal interest and abilities, financial rewards, flexible hours and sponsorships. Personal interest and abilities were identified as the most influential in choosing computer science career. Skills required for a computer science career professional was identified as math, problematic solving, graphics, computing abilities, reasoning, ability to communicate, creativity, typing speed and business know how. Basic computer skills, communication and problem solving are most important required skills for a computer professional. Other skills are math, business knowledge creativity, graphics and logic. The research addressed on the characteristics of computer professionals that include clever, persistent, systematic, well-informed, hardworking, recall, teamwork, ordered, go-getter and ready to learn. Most importantly, characteristics for computer professional were team player, analytical, smart, good memory and hardworking.

In summary, a computer professional parameters include professional skills and abilities, aptitude, communication skills, analytical skills, team player, personal interest and professional experience.

2.6 Dataset

A dataset is an organized collection of data. A dataset can be used to hold any kind of data, including database tables and collections of arrays. A tabular dataset is a database table where each row corresponds to a dataset's fields and each column to a certain variable. A tabular dataset's most often accepted file type is "Comma Separated File(CSV)". The success of machine learning classification depends on the availability of datasets (Althnian, et al., 2021).

According to (Althnian, et al., 2021), the performance of classifier doesn’t depend on the size of the dataset but the extent of distribution of original dataset representation. Additionally, just because a machine learning model is resilient for a small dataset does not always mean that it outperforms other models.

According to (Alwosheel, Sander, & Chorus, 2018), classification dataset size for machine learning should be at least fifty to one thousand the number classes to be classified according to a rule of thumb. Additionally, it specifies that the dataset should be a minimum of ten to a hundred times the available characteristics or attributes. Data 4 ver1 has nineteen attributes and 2064 instances.

2.7 Feature selection Techniques

When utilizing machine learning to create a predictive model, feature selection refers to the process of choosing and manipulating variables/features in a dataset. Feature selection removes redundant attributes or features that are non-informative from the model. There are different techniques used for feature selection in machine learning by reducing input variables and selecting those that are believed to be more useful predicting the objective variable to a model (Sharma & Dey, 2012).

Overview of feature selection techniques

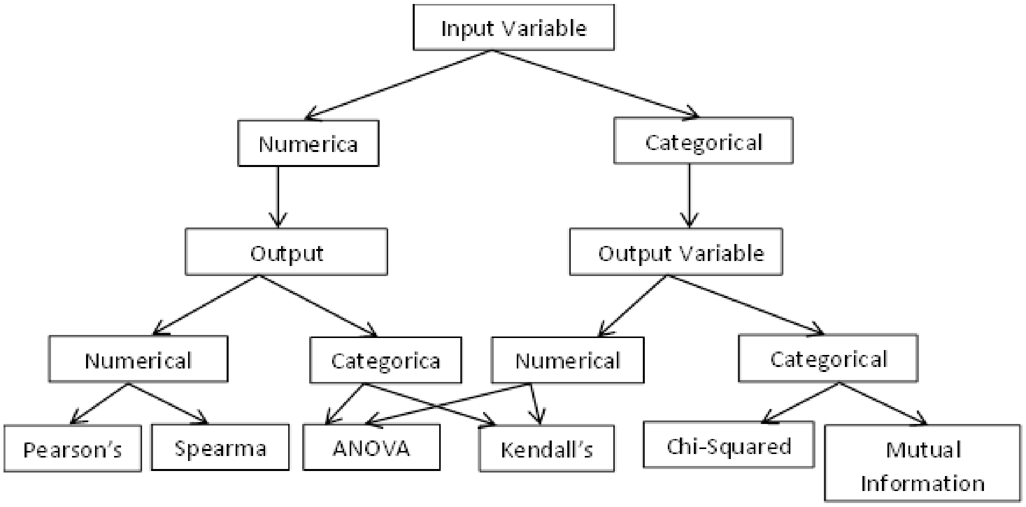


Figure 2 Criteria for Feature selection techniques for Machine Learning (Brownlee, 2019)

According to (Venkatesh & Anuradha, 2019), the following feature selection techniques has been described:

Pearson's Correlation is a measure for figuring out how two continuous variables, X and Y, relate to one another linearly. Its value is between -1 and +1. fs2 is the Pearson's correlation coefficient.

ANOVA (Analysis of Variance): It functions similarly to Linear Discriminant Analysis (LDA), except it only requires one continuous dependent feature and one or more categorical independent variables. To ascertain whether or not the means of several groups are equal, it performs statistical test..

Chi-Square: This is a statistical test that examines the potential for correlation between sets of categorical variables according to the frequency distribution of the variables.

Mutual information is the application of information gain (usually used in the building of decision trees) to feature selection from the discipline of information theory. Mutual information between two variables measures how much uncertainty is reduced for an attribute when the output of another attribute is known.

Reasons for feature selection according to Venkatesh & Anuradha, 2019.

1. It allows quick learning of the machine learning system.
2. Reduction of model complexity as well as making it simple to understand.
3. Choice of right subset increases the model accuracy.
4. Over fitting is reduced.

2.8 Machine Learning Algorithms

2.8.1 Prediction using Naïve Bayes(NB)

NB classifier is a branch of mathematics that deals with theory of probability. It finds the highest opportunities for likely classification by getting frequency of training data of each classification. NB classifier being a statistical classifier is used to predict class membership probability. Using a modest quantity of training data to calculate the parameter estimations, it has been demonstrated to have a high accuracy (Gerhana, et. al., 2019).

Naïve Bayes algorithms have been known for their accuracy, robust and efficiency in prediction of students' delayed graduations (Lagman, 2019) and in academic performance (Razaque, 2017).

Thomas Bayes introduced the Naïve Bayes algorithm. It provides knowledge, data and practical learning which is combined to provide prediction. The formula for Naïve Bayes is:

$$P\left(\frac{H}{E}\right) = \frac{P(E|H) * P(H)}{P(E)}$$

Where,

P(H) - prior probability,

P(E) - evidence probability,

P(E|H) -evidence probability when the hypothesis is true

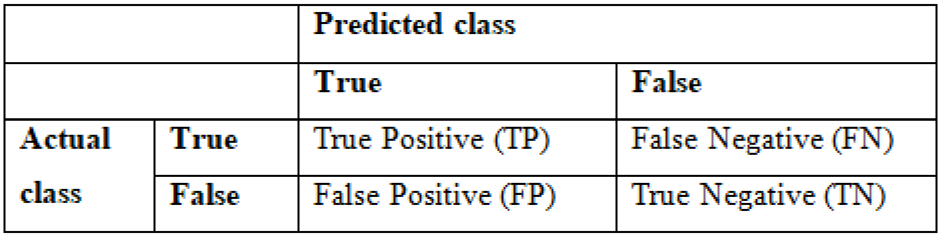
P(H|E) - probability of the hypothesis when evidence is true (Melian & Nursikuwagus, 2018).

According to (Gerhana, et. al., 2019), Naïve Bayes is implemented in four steps:

1. To calculate the number of classes or attributes to be used is the first step.
2. The number of cases in each attribute are calculated based on testing data.
3. All variable results are then multiplied
4. Finally, the results of each class are compared

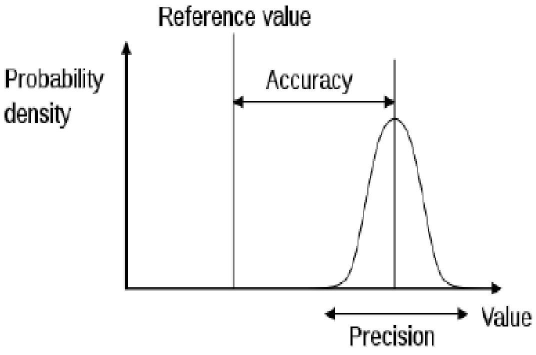
To test the performance of NB, confusion matrix is used with the area under the curve interpreted. Confusion matrix is a machine learning tool that contains two or more categories.

Example of confusion matrix



		Predicted class	
		True	**False**
Actual class	**True**	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

Diagram below shows Precision and accuracy. Precision is how near result is to the real result and accuracy is the shift of the result from the actual value (Melian & Nursikuwagus, 2018).



Advantages of Naïve Bayes

- ✓ Works quickly and saves time
- ✓ Multi-class prediction problems are suitably solved by Naïve Bayes
- ✓ Naïve performs better with assumption of independence of features and less time is required for training
- ✓ Naive Bayes works well for categorical inputs as compared to numerical variables.

Disadvantages of Naïve Bayes

- ✓ The Naive Bayes algorithm’s applicability is limited in real-world situations due to the assumption of independence.
- ✓ Naïve Bayes algorithm has the problem of “Zero-Frequency” where it assigns probability of zero to categorical variables which has no category in training dataset.

Naïve Bayes Categories

According to (Singh, Kumar, Gaur, & Tyagi, 2019), Naïve Bayes are categorized as:

GaussianNB: It assumes attributes must be distributed normally when classifying. **MultinomialNB**: Discrete counts are conducted using this method of counting. Assume the problem is with text categorization.

BernoulliNB: If attributes paths are binary, binomial model is practical. The "bag of words" paradigm is used to categorize text, with binary standing for " the documents’ word is existing" as 1s and " no word in the document" as 0s.

ComplementNB- The Complement Naive Bayes classifier was created to correct the normal Multinomial Naive Bayes classifier's "severe assumptions." It's especially well-suited to unbalanced data sets.

Applications of Naïve Bayes classifier

Instantaneous Prediction: NB is a rapid and enthusiastic classifier for learning. It may be used to produce forecasts instantaneously.

Prediction of many classes: The capability of many classes prediction in this technique is also renowned. It can predict the probabilities of several target variable classes.

Text classification, sentiment analysis and spam filtering: When it comes to text categorization, Naive Bayes classifiers outperform other methods (due to superior outputs in many classes conditions and the individuality criteria). Consequently, it is frequently used in emotion analysis, filtering of spams, identification of sentiments from customers as negative or positive in social media.

System Recommender: NB Classifier and Filtering Collaborator come together to develop a System Recommender that uses data mining and machine learning methods to separate unknown data as well as forecast the output will be resourceful to the user.

2.8.2 Random Forest

Decision tree algorithm is used to build a Random Forest (RF) which is a supervised machine learning technique. Random forest maintains accuracy in a large proportion of data by handling missing values. It's main advantage is that it performs classification and regression tasks. RF has the capability of avoiding overfitting when managing large amount of datasets (VidyaShreeram & Muthukumaravel, 2021).

The steps for Random Forest classifier according to VidyaShreeram & Muthukumaravel, 2021 are:

1. A number "k" is chosen from the total number of "m" features.
2. The value "d" is computed using split point from "k" features
3. Using the data best split, the nodes are divided into sub nodes
4. Repeating of step 1-3 until "I" nodes reached
5. Steps 1-4 is repeated to construct the forest.

A Random forest has many decision trees. Random forest algorithm generate the "Forest" through bootstrap aggregating. Bootstrap aggregating improves the accuracy of the algorithm. Random forest prediction outcomes are based on prediction of decision trees by taking mean of various trees output. The higher the number of trees the higher the precision.

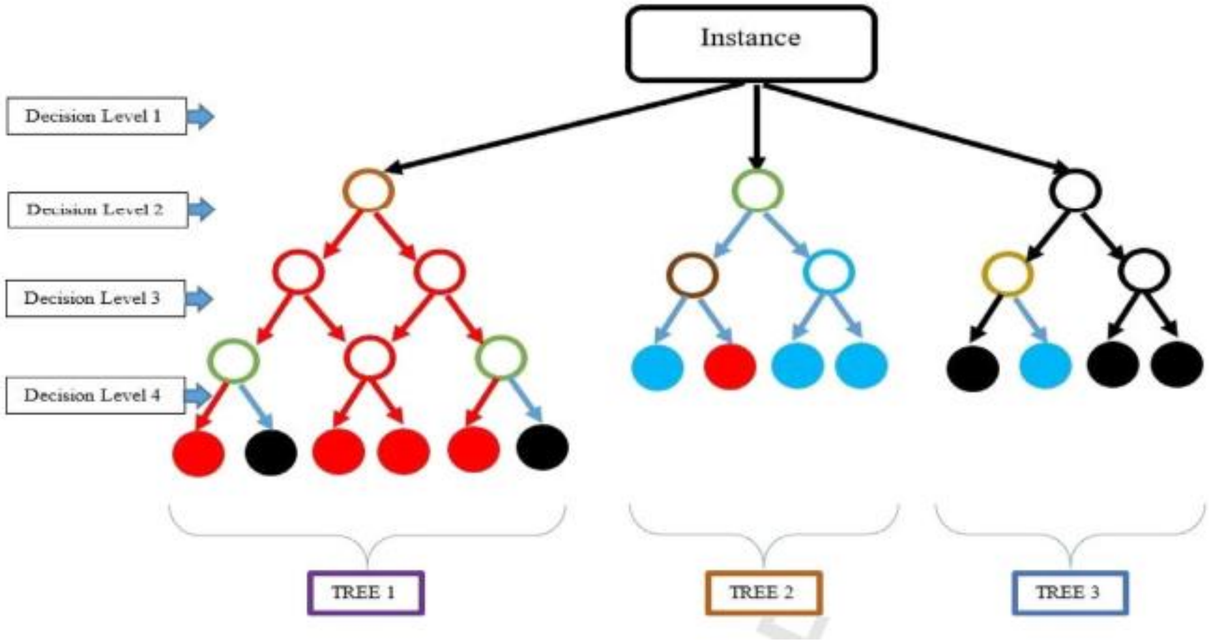


Figure 3 Random Forest Classifier by Subudhi, Dash, & Sabut, 2019

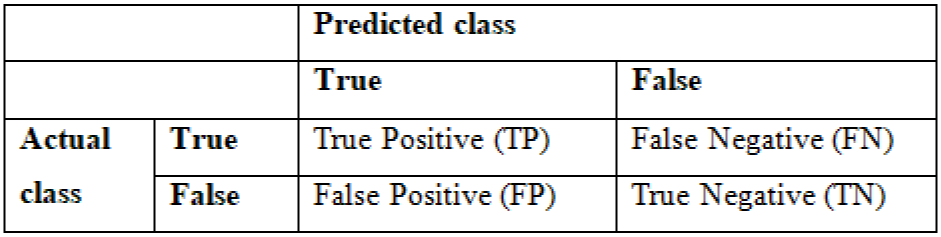
Benefits of Random Forest

- As compared to decision tree, RF is considered more accurate.
- Missing data is handled practically because it has a way of doing.
- Without adjustment of hyper-parameter, a fair estimate can be generated.
- It is able to handle the challenge of over-fitting the occur in decision trees.
- It randomly selects features in a subset where there is node splitting in RF.

2.9 Evaluation

This step involves verifying that the results are valid and correct. The confusion matrix was employed to verify the model's predictions. Confusion matrix is a representation of the results of a binary testing. In a classification problem, confusion matrix represent the summary of prediction results (Gerhana, Fallah, Zulfikar, Maylawati, & Ramdhani, 2019). Inaccurate and accurate predictions are summed up and each class is separated after counting. The advantage of a confusion matrix is that it is simple to determine whether the system is confusing the two groups.

Confusion matrix example



		Predicted class	
		True	False
Actual class	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

Accuracy is degree of closeness to true value.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is how often an a process will provide the same value.

$$Precision = \frac{TP}{TP + FP}$$

Recall is equal to total count of true positive which is divided by the sum of the number of false negatives and true positives.

$$Recall = \frac{TP}{TP + FN}$$

F1 score is a measurement that is calculated from recall and precision.

$$F1\ SCORE = 2 * \frac{precision * recall}{precision + recall}$$

As a rule of thumb, the measure for selecting the most appropriate classifier is the accuracy of classification. Accuracy is defined as the proportion of cases that are properly identified during testing to the total number of examples tested (Jeyarani, Nagaveni, & Ram, 2013).

2.10 Research gap

Personality and student grades have been considered by researchers but background information and environmental factors, which are also key contributors to the specialty of a computing student have not been put into consideration. Therefore, they lack high accuracy prediction of college students' specialty, hence limiting the ability to predict and classify computing students' specialty accurately.

Random forest and Naïve Bayes algorithms have been implemented in different prediction models with relatively high accuracy. However, the implementation of these algorithms in career prediction of computing specialty has not been done. The comparison of RF and NB techniques in career prediction to identify the most appropriate one has not been done.

With the current models used in computing specialty prediction, there are no higher levels of accuracy and applicability in prediction for computing students. The integration of Cumulative Grade Point Average (CGPA),environmental factors, background information and previous experience in computing specialty prediction model with appropriate classification techniques has not been achieved.

3.0 CHAPTER THREE: METHODOLOGY

3.1 Research design

The researchers objective was to create a predictive model for computing careers and it was achieved using the CRISP-DM model. Due to advantages and requirements of the academic research community, this approach was chosen above others like KDD, SEMMA, and KDP in order to provide a more generic, research-oriented explanation of the procedures (Abd Alazeez & Thaher, 2021). The design will involve manipulations of variables and predictions on the basis of previous observations or history. The study will be concerned with the identification and integration of the appropriate parameters to create an optimal model which is accurate and reliable in predicting computing careers.

3.2 The Proposed Model's Overall Architecture

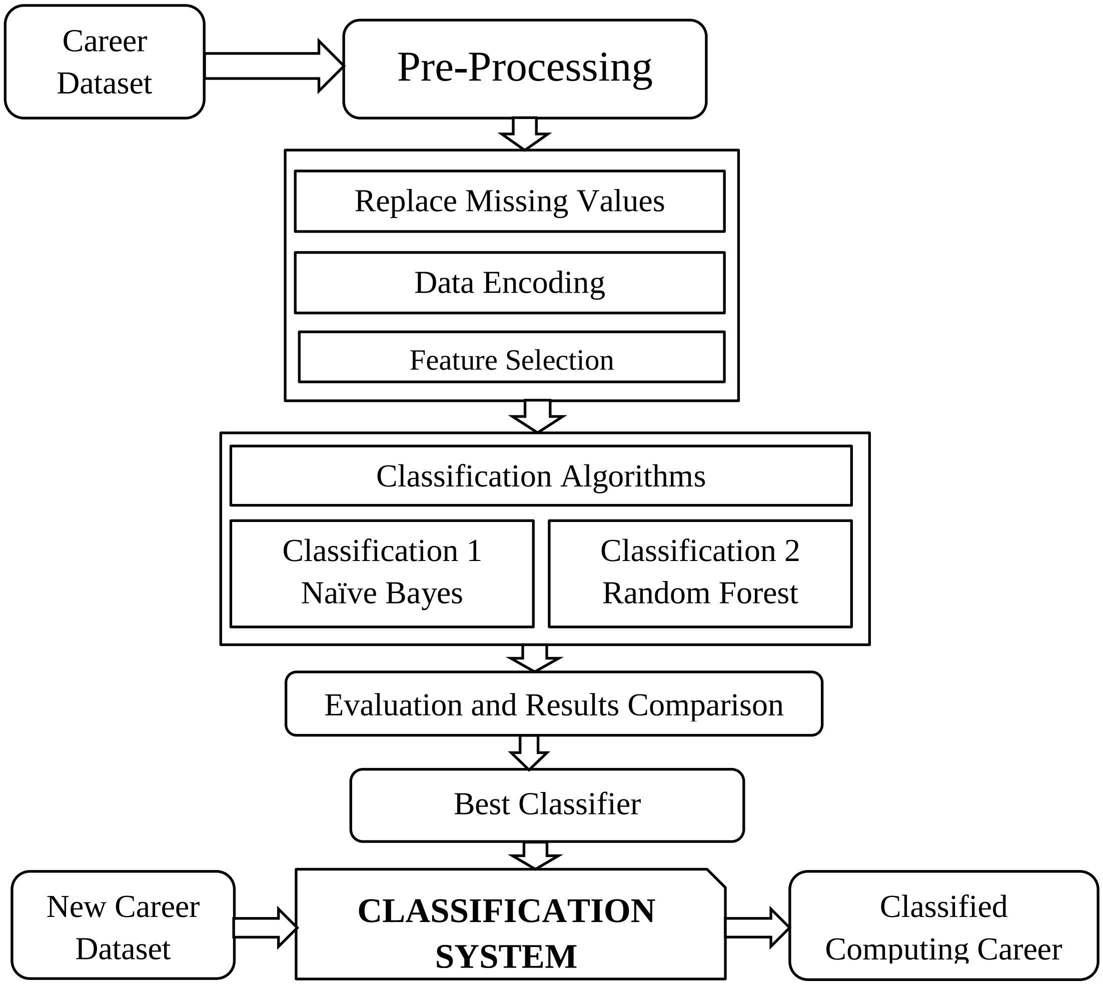

Career Dataset
Pre-Processing
Replace Missing Values
Data Encoding
Feature Selection
Classification Algorithms
Classification 1 Naïve Bayes
Classification 2 Random Forest
Evaluation and Results Comparison
Best Classifier
New Career Dataset
CLASSIFICATION SYSTEM
Classified Computing Career

Figure 4 Proposed Computing career classification system overall Architecture

Career Dataset

It's critical to get trustworthy data so that to uncover the right trends in machine learning model. Accuracy of the model developed depends on the quality of the dataset used. For this study, the dataset was from Kaggle repository which is reliable and has datasets that has been used in machine learning.

Pre-Processing

The data preprocessing component is in charge of cleaning the career dataset. The dataset with missing values were dropped to minimize false prediction. The dataset contained categorical and numerical features. Therefore the dataset was encoded for modeling. It was also visualized with an aim to understand   the relationship between various attributes after encoding. Encoding created duplicate attributes which were dropped because they added no value to the dataset. Feature selection was done using Chi-square and relevant attributes were selected for modeling.

Classification Algorithms

The component known as "model building and comparison" is in charge of compiling data on careers using appropriate classification procedures like RF and NB and contrasting the two approaches. The task of the classification system component is to map the pattern in the rules developed using the new career data in order to forecast possible careers in computing.

Evaluation

The model must be evaluated once it has been trained. This is accomplished by putting the model through its paces on data that has never been seen before. The testing set was divided from dataset before it was utilized in a ratio of 70:30. If model is tested on the same data that was used for training, the receive result will not be accurate since the model is already familiar with it and recognizes the same patterns. The results of both Random forest and Naïve Bayes were compared.

Classification system

The model was used to make predictions using new inputs or unseen data.

3.3 CRISP-DM Overview

In order to structure data mining initiatives, a technique called CRISP-DM was settled in 1996. According to (Steffen, Hajo, Dorothea, & Steffen, 2019), CRISP-DM is a six step methodology which has cycle iterations depending on needs of developers. The procedures are as follows: understanding the business, comprehending the data, preparing the data, modeling, evaluating, and deploying.

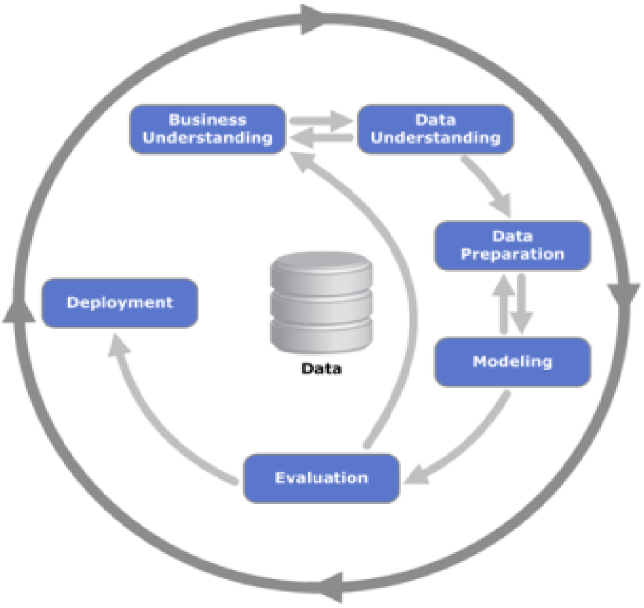


Figure 5 CRISP-DM Steps guide

3.3.1 Business Understanding

The objective of business understanding is to give context to the goals and to the data so that the data engineer is able to particularly get a view of data with respect to business model. This step comprises documentation reading and listing ways that help the development of system and make questions on context relevancy.

Several literatures were evaluated to assess methodologies and machine learning technologies in prediction and classification in this field in order to gain insight into what was needed. To better comprehend the applicability of machine learning in the prediction of computing careers using the data 4 ver1 dataset, a number of books and research articles on prediction models from the internet were analyzed.

3.3.2 Data Understanding

Understanding what can be gained from the data is the purpose of this stage. The completeness and dispersion of the data are examined to determine the quality of the data. It is important because it defines if the final result is viable and trustworthy. The research study used samples from Kaggle repository where data 4 ver1 was identified which was in Comma separated values (CSV) format. The dataset used contained 20 attributes and 2064 instances. This original dataset contained 18 categorical, one continuous input attributes and one categorical class labeled Role which is the speciality of the known computing career. The dataset contains ten computing career roles (Marketing, Network Engineer, Developer, Computer Analyst, Data Analysis, ML engineer, Content Writer, Data Engineer, Management and Security) which can be predicted. The dataset utilized a range of questions that students must respond to and CGPA in order to measure their previous knowledge.

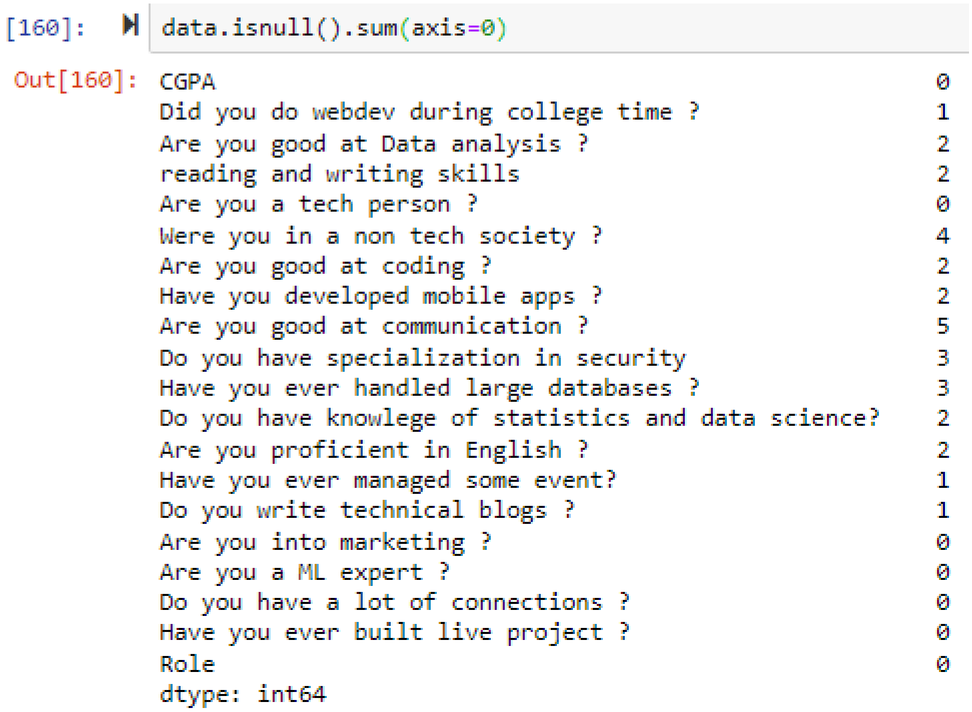
Table 1 Dataset parameters distributions

	Parameters	**Instances**	**Samples**
1.	CGPA	0-9.9	2064
2.	Did you do web development during college time	Yes	968
		No	1092
3.	Are you good at Data analysis ?	Yes	448
		No	1608
4.	Reading and writing skills	Excellent	724
		Medium	648
		Poor	684
5.	Are you a tech person ?	Yes	1016
		No	1048
6.	Were you in a non-technical society ?	Yes	764
		No	1284
7.	Are you good at coding ?	Yes	1328
		No	728
8.	Have you developed mobile apps ?	Yes	620
		No	1436
9.	Are you good at communication ?	Yes	1152
		No	892
10.	Do you have specialization in security	Yes	116
		No	1936
11.	Have you ever handled large databases?	Yes	540
		No	1512
12.	Do you have knowledge of statistics and data science?	Yes	500
		No	1556

13.	Are you proficient in English ?	Yes	1704
		No	351
14.	Have you ever managed some event?	Yes	596
		No	1464
15.	Do you write technical blogs ?	Yes	516
		No	1544
16.	Are you into marketing ?	Yes	248
		No	1816
17.	Are you a ML expert ?	Yes	380
		No	1684
18.	Do you have a lot of connections ?	Yes	684
		No	1380
19.	Have you ever built live project ?	Yes	632
		No	1432

3.3.3 Data preparation

The data utilized for analysis was chosen at this stage by evaluating its importance to the purpose of the data mining process, and its value, amount, type as well as technological limitations. The data was changed into the proper format to enable analysis of the selected algorithms. The data was saved as a CSV file. To preprocess the data, Juypter notebook was used and Sklearn libraries were imported. Data was Checked for missing values using Isnull library from Sklearn. The data had missing values.

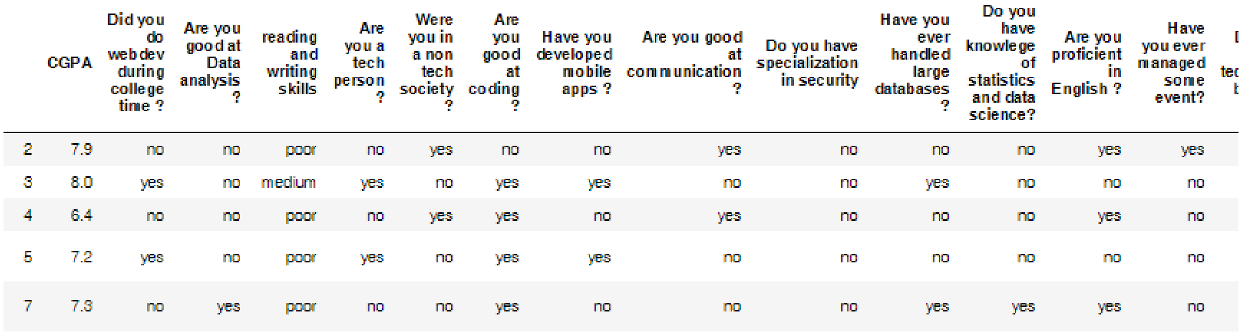


```
[160]:   data.isnull().sum(axis=0)

Out[160]: CGPA                                                   0
          Did you do webdev during college time ?                1
          Are you good at Data analysis ?                        2
          reading and writing skills                             2
          Are you a tech person ?                                0
          Were you in a non tech society ?                       4
          Are you good at coding ?                               2
          Have you developed mobile apps ?                       2
          Are you good at communication ?                        5
          Do you have specialization in security                 3
          Have you ever handled large databases ?                3
          Do you have knowlege of statistics and data science?   2
          Are you proficient in English ?                        2
          Have you ever managed some event?                      1
          Do you write technical blogs ?                         1
          Are you into marketing ?                               0
          Are you a ML expert ?                                  0
          Do you have a lot of connections ?                     0
          Have you ever built live project ?                     0
          Role                                                   0
          dtype: int64
```

Figure 6 Preview of missing values on the dataset

To clean the data for null instances, rows that had no values in any of the field were dropped.

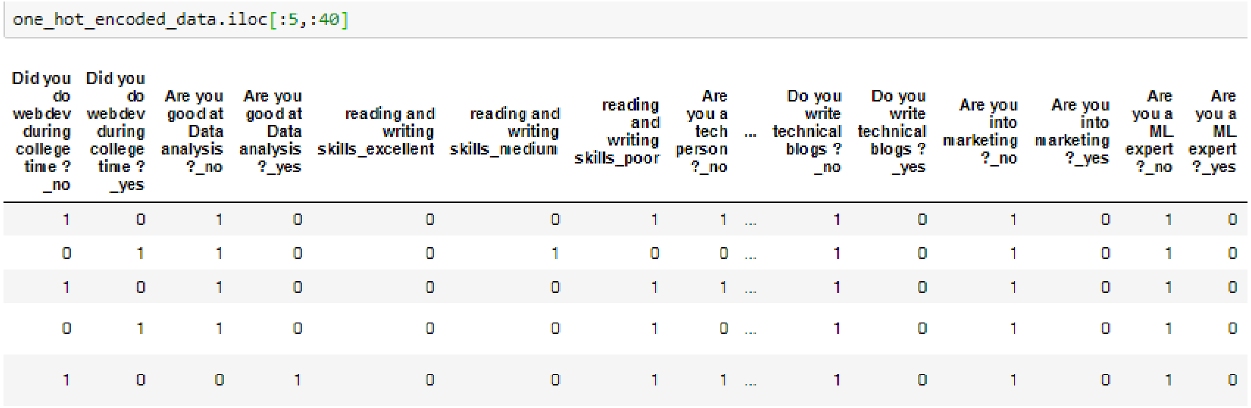


	CGPA	Did you do webdev during college time ?	Are you good at Data analysis ?	reading and writing skills	Are you a tech person ?	Were you in a non tech society ?	Are you good at coding ?	Have you developed mobile apps ?	Are you good at communication ?	Do you have specialization in security	Have you ever handled large databases ?	Do you have knowlege of statistics and data science?	Are you proficient in English ?	Have you ever managed some event?
2	7.9	no	no	poor	no	yes	no	no	yes	no	no	no	yes	yes
3	8.0	yes	no	medium	yes	no	yes	yes	no	no	yes	no	no	no
4	6.4	no	no	poor	no	yes	yes	no	yes	no	no	no	yes	no
5	7.2	yes	no	poor	yes	no	yes	yes	no	no	no	no	no	no
7	7.3	no	yes	poor	no	no	yes	no	no	no	yes	yes	yes	no

Figure 7 Preview of data after dropping rows with missing values on the dataset

Encoding

The data set contained continuous and categorical data. The categorical data contained nominal data. Nominal data has no any order that has to be maintained. Categorical data being nominal, one hot encoding was selected for encoding. In one hot encoding, it creates a new variable for each level of a category feature. For each category, a binary variable with a value of 0 or 1 is used as a representation. The absence of that category is indicated by 0 and its presence by 1, respectively. Dummy variables are the names of freshly produced binary characteristics. The number of dummy variables used is determined by the category of variable's levels.



```
one_hot_encoded_data.iloc[:5,:40]
```

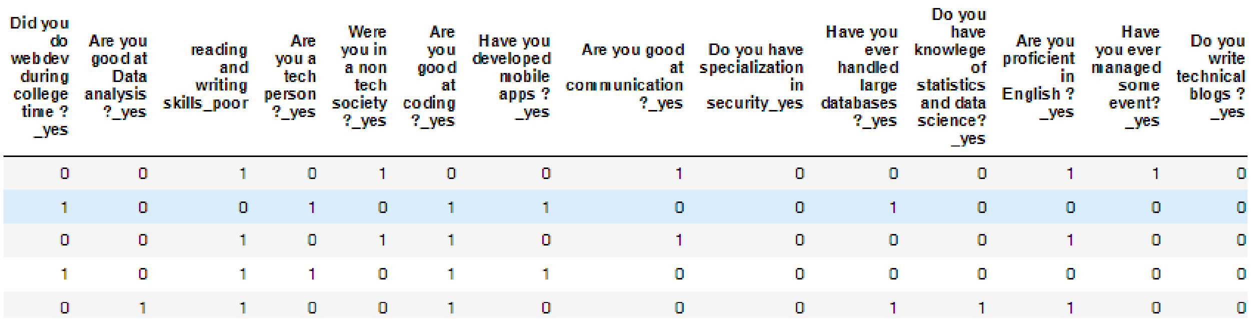
Did you do webdev during college time ?_no	Did you do webdev during college time ?_yes	Are you good at Data analysis ?_no	Are you good at Data analysis ?_yes	reading and writing skills_excellent	reading and writing skills_medium	reading and writing skills_poor	Are you a tech person ?_no	...	Do you write technical blogs ?_no	Do you write technical blogs ?_yes	Are you into marketing ?_no	Are you into marketing ?_yes	Are you a ML expert ?_no	Are you a ML expert ?_yes
1	0	1	0	0	0	1	1	...	1	0	1	0	1	0
0	1	1	0	0	1	0	0	...	1	0	1	0	1	0
1	0	1	0	0	0	1	1	...	1	0	1	0	1	0
0	1	1	0	0	0	1	0	...	1	0	1	0	1	0
1	0	0	1	0	0	1	1	...	1	0	1	0	1	0

Figure 8 Sample Data after One hot Encoding career dataset

One-hot encoding gives an extra column for each level. Therefore, dropping one column for a level does not affect the information.

Table 2 Performance Metric After one-hot encoding

Evaluation Metric	**Random Forest**	**Naïve Bayes Models**			
		GaussianNB	BernoulliNB	MultinomialNB	ComplementNB
Number of Attributes	39	39	39	39	39
Accuracy	80.952381	74.829932	84.353741	78.231293	72.108844
Precision	78.137765	66.623230	77.832076	72.802358	70.438755
Recall	74.284632	73.853111	81.776961	76.171956	67.995726
F1 score	75.007436	67.504038	79.248164	74.149138	62.358394



Did you do webdev during college time ? _yes	Are you good at Data analysis ?_yes	reading and writing skills_poor	Are you a tech person ?_yes	Were you in a non tech society ?_yes	Are you good at coding ?_yes	Have you developed mobile apps ? _yes	Are you good at communication ?_yes	Do you have specialization in security_yes	Have you ever handled large databases ?_yes	Do you have knowlege of statistics and data science? _yes	Are you proficient in English ? _yes	Have you ever managed some event? _yes	Do you write technical blogs ? _yes
0	0	1	0	1	0	0	1	0	0	0	1	1	0
1	0	0	1	0	1	1	0	0	1	0	0	0	0
0	0	1	0	1	1	0	1	0	0	0	1	0	0
1	0	1	1	0	1	1	0	0	0	0	0	0	0
0	1	1	0	0	1	0	0	0	1	1	1	0	0

Figure 9 One-hot encoded data after dropping a column in each level

Feature selection technique to use and the reason

Features selection was carried out to identify relevant attributes for modeling. Chi-square was used to identify the best features for modeling computing careers. In statistics, the chi-square investigation is used to scrutinize the independence of both occurrences. It is expected for count O and count E to be seen when two data variables are presented. Chi-square calculates the difference between the observed count O and the predicted count E.

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

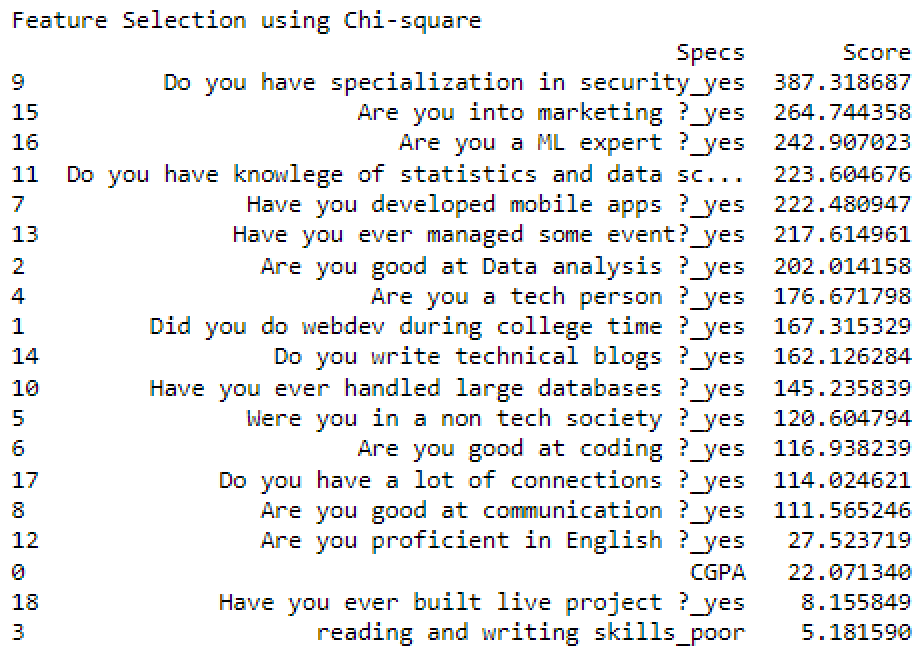
where

c= degrees of freedom

E=expected value(s)

O=observed value(s)

The chi-square value will be lower if two features are autonomous and the experiential counts are near to the foreseen count. When two characteristics are independent, the observed count is very close to the predicted count, hence the chi-square value will be lower. A higher chi-square score indicates that the independence hypothesis is incorrect. As a result, the feature can be chosen for model training if the chi-square value is larger and indicates a greater dependency on the response.



```
Feature Selection using Chi-square
                                                Specs       Score
9         Do you have specialization in security_yes  387.318687
15                      Are you into marketing ?_yes  264.744358
16                         Are you a ML expert ?_yes  242.907023
11 Do you have knowlege of statistics and data sc...  223.604676
7               Have you developed mobile apps ?_yes  222.480947
13             Have you ever managed some event?_yes  217.614961
2                Are you good at Data analysis ?_yes  202.014158
4                      Are you a tech person ?_yes  176.671798
1          Did you do webdev during college time ?_yes  167.315329
14                Do you write technical blogs ?_yes  162.126284
10         Have you ever handled large databases ?_yes  145.235839
5               Were you in a non tech society ?_yes  120.604794
6                       Are you good at coding ?_yes  116.938239
17            Do you have a lot of connections ?_yes  114.024621
8                Are you good at communication ?_yes  111.565246
12               Are you proficient in English ?_yes   27.523719
0                                               CGPA   22.071340
18            Have you ever built live project ?_yes    8.155849
3                    reading and writing skills_poor    5.181590
```

Figure 10 Output of Feature selection using Chi-square

3.3.4 Modeling

The models are categorized as supervised and unsupervised. In supervised, both inputs and expected outputs are accessed by machine making it possible to do classification. In unsupervised the expected output is not available. The dataset in this case has inputs and expected outputs which inform the decision of supervised learning. The step involved evaluation, selection and application of appropriate modeling features. To build a classification model from the preprocessed data 4 ver1, Sklearn library in python was used. Naïve Bayes and Random Forest models were applied to predict computing career path from data 4 ver1 dataset. Data prediction models were created and tested using Random Forest and Naïve Bayes models.

3.3.5 Evaluation

The labeled dataset was separated into training and testing in a ratio of 70:30, where 70% was used to train the model and 30% was used for evaluation. To identify the number of attributes appropriate for the prototype, evaluation was done using 5, 10, 15 and 19 features. The metric to be examined for RF and NB models are recall, accuracy, F1 measure as well as precision.

3.2.6 Deployment

After development of the prototype, different categories of users were given the prototype to use for prediction. The categories involved were: Career counselors, Human Resource, IT Professionals and Graduate Computing Students within Nairobi County in Kenya. Random sampling was used to select users in each category.

The population to be sampled was infinite and therefore sample size determination provided by Godden(2004) was used since the target population of computer professionals is greater than 10,000. The following formula was used to determine the sample size:

$$n = \frac{Z^2 \times P\,(1-P)}{E^2}$$

Where: n = Infinite population sample Size

Z = Confidence level at 95% (Standard value 1.96)

P = Population proportion assumed to be 0.1 (10%)

E= Margin of Error at 5% (0.05)

$$n = \frac{1.96 \times 0.1(1-0.1)}{0.05^2} = \frac{1.96 \times 0.1(0.9)}{0.0025}$$

Therefore, sample size for computer professional was seventy. This was then divided into seventeen career counselors, fourteen human Resource, thirteen IT professionals and twenty six graduate computing students.

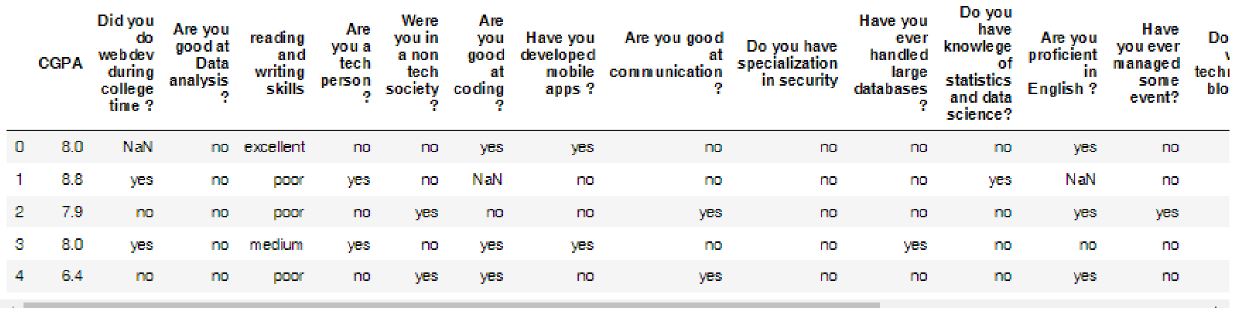
4.0 CHAPTER FOUR: RESULTS AND DISCUSSION

4.1 Introduction

Several categorization experiments were put up, concentrating on various supervised machine learning approaches comprising RF and NB models. The objective was to predict the best careers for computing students using different features. This was accomplished by using data 4 ver1 to evaluate several Nave Bayes and Random Forest models using various evaluation measures.

4.2 Experiment setup

Dataset of the format CSV (Comma Separated values) was imported into Jupyter Notebook.

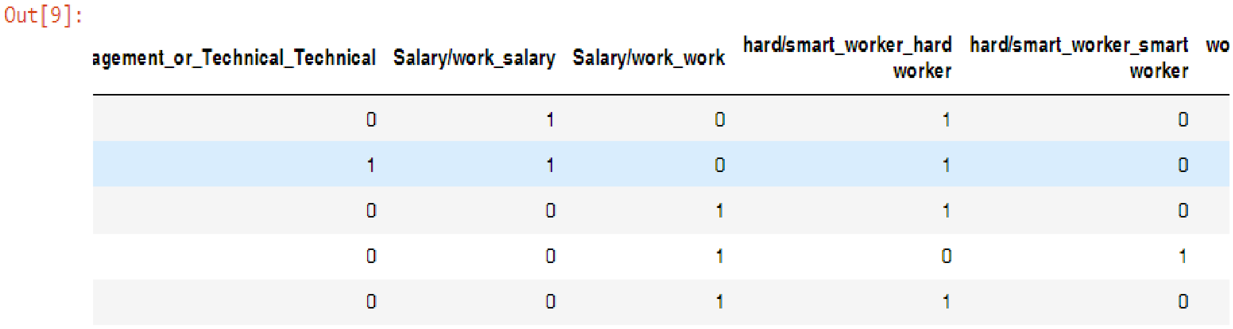


	CGPA	Did you do webdev during college time ?	Are you good at Data analysis ?	reading and writing skills	Are you a tech person ?	Were you in a non tech society ?	Are you good at coding ?	Have you developed mobile apps ?	Are you good at communication ?	Do you have specialization in security	Have you ever handled large databases ?	Do you have knowlege of statistics and data science?	Are you proficient in English ?	Have you ever managed some event?	Do v techı blo
0	8.0	NaN	no	excellent	no	no	yes	yes	no	no	no	no	yes	no	
1	8.8	yes	no	poor	yes	no	NaN	no	no	no	no	yes	NaN	no	
2	7.9	no	no	poor	no	yes	no	no	yes	no	no	no	yes	yes	
3	8.0	yes	no	medium	yes	no	yes	yes	no	no	yes	no	no	no	
4	6.4	no	no	poor	no	yes	yes	no	yes	no	no	no	yes	no	

Figure 11 Raw data Loaded on Juypter Notebook

Encode data

The data was encoded using one hot encoding and dummy variables were created for categorical variables.



Out[9]:

agement_or_Technical_Technical	Salary/work_salary	Salary/work_work	hard/smart_worker_hard worker	hard/smart_worker_smart worker	wo
0	1	0	1	0	
1	1	0	1	0	
0	0	1	1	0	
0	0	1	0	1	
0	0	1	1	0	

Figure 12 Sample Data after One hot Encoding career dataset

4.3 Model Building

In the hunt for the ideal Model, the model construction supported in this study was classification. Seventy percent of the dataset was used for training and thirty percent as testing to prevent overfitting and increase accuracy and, consequently, applicability in the performance dataset.

4.4 Modeling Techniques and Tools Used

The supervised learning (classification) approach was used in the research to create a prediction model in machine learning. Jupyter Notebook was utilized as the software tool, which is an open-source and free program for knowledge analysis that can be downloaded on the internet. Using sklearn, Juypter Notebook implements many machine learning techniques.

4.5 Performance Evaluation for Predictive Model

4.5.1 Predictive model and basic classification results using Jupyter Notebook

On data 4 ver1 dataset, experiments were carried out using the Jupyter notebook to examine several Nave Bayes and Random Forest classification methods. The experiments tested various classification model outcomes based on percentage accuracy of correctly categorized occurrences. For each model, dataset and the environmental variables were the same. Several metrics were compared, they include precision, recall, accuracy and F1 score. Precision is how often an a process will provide the same value. Recall is equal to total count of true positive which is divided by the sum of the number of false negatives and true positives.

Random forest and Naive Bayes were the models chosen for the experiments because of their widespread use and value in resolving classification issues. A total of 70% of the dataset was used for training and 30% was used for testing in order to train the classifier on data 4 ver1 dataset. RF and NB models were employed to develop the prediction model for the computing careers.

4.5.2 Training dataset

NB and RF algorithms were implemented and assessed via Sklearn which is a library in python on data 4 ver1. NB and RF models were created using training data with known output values. Next, the data was run through the model to create the desired output whenever a new data point with an unknown output value was used.

Table 3 Performance metrics of Prediction Algorithms

NO OF ATTRIBUTES	ACCURACY	PRECISION	RECALL	F1 SCORE
	GAUSSIANNB			
5	23.770	16.609	26.607	19.348
10	55.738	64.441	59.608	56.499
15	78.689	76.765	70.289	69.290
19	71.311	69.123	74.384	69.151
	RANDOM FOREST			
5	65.574	41.127	50.647	43.374
10	84.525	79.990	85.335	82.282
15	86.623	78.116	85.361	84.544
19	89.885	79.190	83.190	79.972
	BERNOULLINB			
5	56.557	36.843	45.656	37.293
10	81.967	73.368	74.266	73.528
15	83.607	74.881	84.786	79.086
19	87.705	78.714	81.535	79.562
	COMPLEMENTNB			
5	49.180	30.205	36.718	32.971
10	66.393	55.176	52.758	49.337
15	74.590	55.870	65.119	59.484
19	77.869	61.398	76.884	66.965
	MULTINOMIALNB			
5	46.721	28.950	38.372	32.085
10	74.590	64.404	67.870	64.632
15	77.049	70.313	72.222	68.076
19	77.869	76.504	76.741	74.568

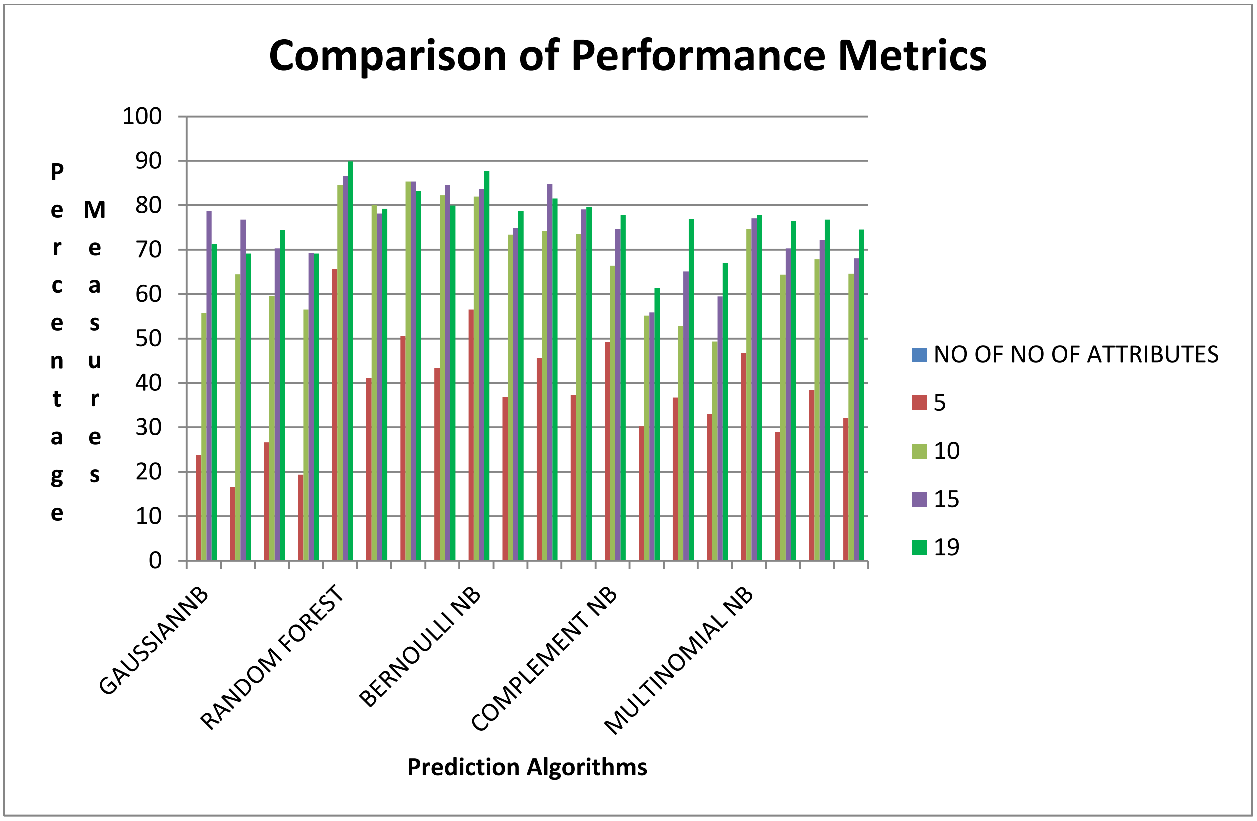


Figure 13 The performance metrics of Prediction Algorithms

Figure 13, the performance evaluation results for the various Naive Bayes models and Random Forest were displayed as a column graph by the researcher. The highest accuracy was found in RF model with 89.885% and 86.863% accuracy using 19 and 15 columns respectively. It was therefore viewed as the most likely scenario. The Random Forest model outperformed the other Naive Bayes algorithms in terms of performance.

4.5.3 Results of interpretation of the training data set

Comparing the Random Forest and Naive Bayes models and choosing the one that performs the best was one of the goals of this study. As a result, all of the aforementioned models were used in each experiment that was conducted for this study. The exact identical datasets were utilized throughout all tests. According to the results of these tests, Random Forest outperformed BernoulliNB, GaussianNB, MultinomialNB, and ComplementNB Naïve Bayes models. RF performance was the most appropriate model in terms of accuracy based on all the benchmarks used to evaluate the models used in this study. Based on this data, emphasis was placed on developing a predictive system using Random forest algorithm in this particular domain.

The efficiency analysis of each data classification approach is presented in Table 3 Performance Metrics of Prediction Algorithms above. It shows instances of properly classified data with 5 columns, 10 columns, 15 columns, and 19 columns for each algorithm. The Random Forest model had an accuracy of 89.885%, precision of 79.190%, recall of 83.190% and F1 score of 79.972% while using 19 columns for appropriate prediction. The GausssianNB model had an accuracy of 71.311%, precision of 69.123%, recall of 74.384% and F1 score of 69.151% while using 19 columns. The BernoulliNB model had an accuracy of 87.705%, precision of 78.714%, recall of 81.535% and F1 score of 79.562% while using 19 columns. The ComplementNB model had an accuracy of 77.869%, precision of 61.398%, recall of 76.884% and F1 score of 66.965% while using 19 attributes. The MultinomialNB model had an accuracy of 77.869%, precision of 76.504%, recall of 76.761% and F1 score of 74.568% while using 19 columns.

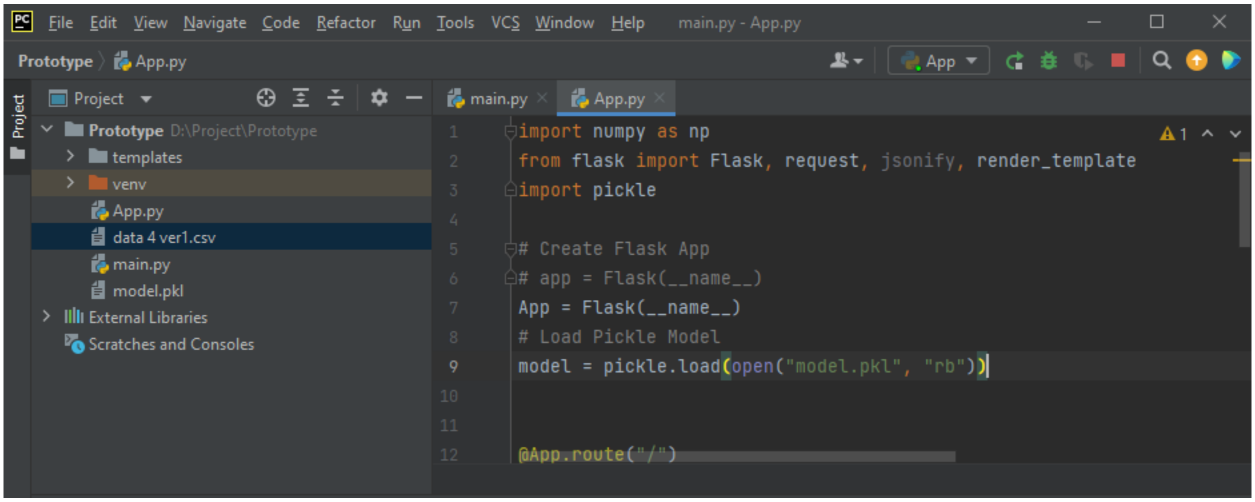
Prediction using One-Hot encoded data require many attributes to be provided as inputs to increase accuracy. In table 3, all algorithms performance was low for five attributes and it increased with an increase of attributes. There was the gradual increase in all performance metrics until it reaches the point of stagnation. The highest accuracy was obtained when all nineteen attributes were used.

Random forest performed better with One-Hot encoded data than Naïve Bayes. Both KNN and Stochastic Gradient Descent were used in career prediction and gave an accuracy of 73.74% and 81.035% respectively using aptitude test (Reddy, 2021).

4.6 Development and Implementation of the Proposed Prototype

The prototype based on the Random forest algorithm was developed using Pycharm community edition and HyperText Markup Language (HTML). Pycharm community edition has been used to develop database using Pickle. Python often uses the pickle library to serialize and deserialize object structures. The original object hierarchy can be recreated by unpickling the stored byte stream. The stored element generates a case of the unique piece when unpickling a byte stream and then fills it with the appropriate data.

HTML was used to prepare the GUI. The user uses GUI for input of data. For filtering and classification, the prototype loads the relevant dataset. The user's front-end choice is taken into account as input. On the server machine, Pycharm Community Edition must first be installed before the project can be loaded.

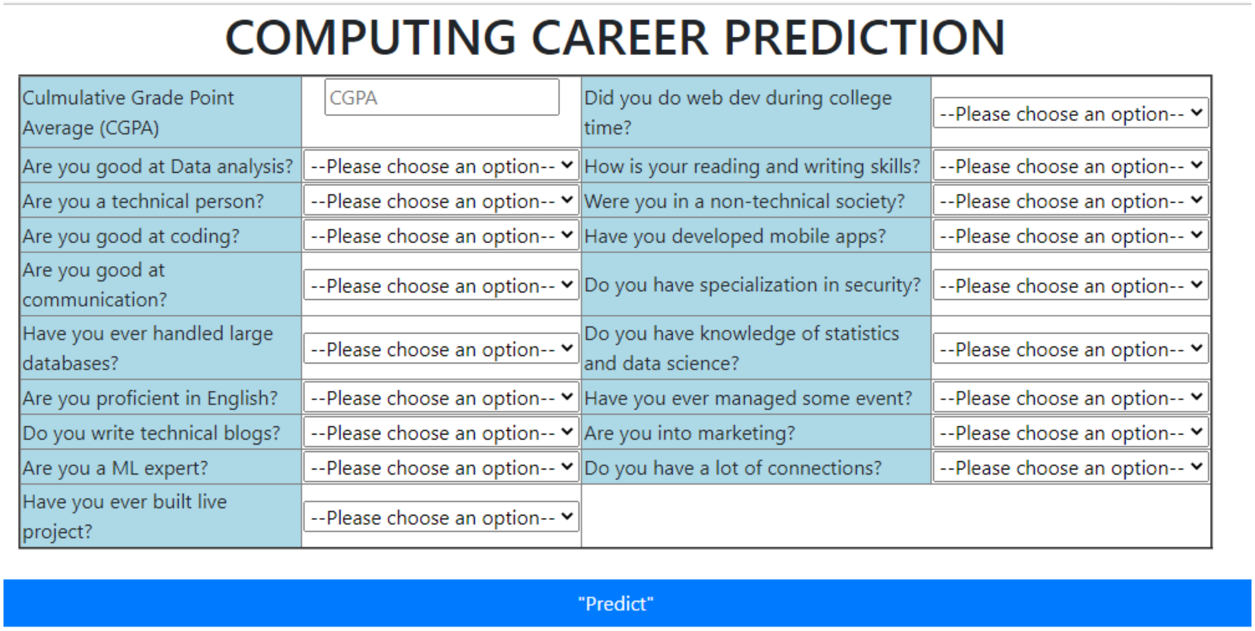


```
import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle

# Create Flask App
# app = Flask(__name__)
App = Flask(__name__)
# Load Pickle Model
model = pickle.load(open("model.pkl", "rb"))


@App.route("/")
```

Figure 14 The Pycharm of the proposed Computing career classification system

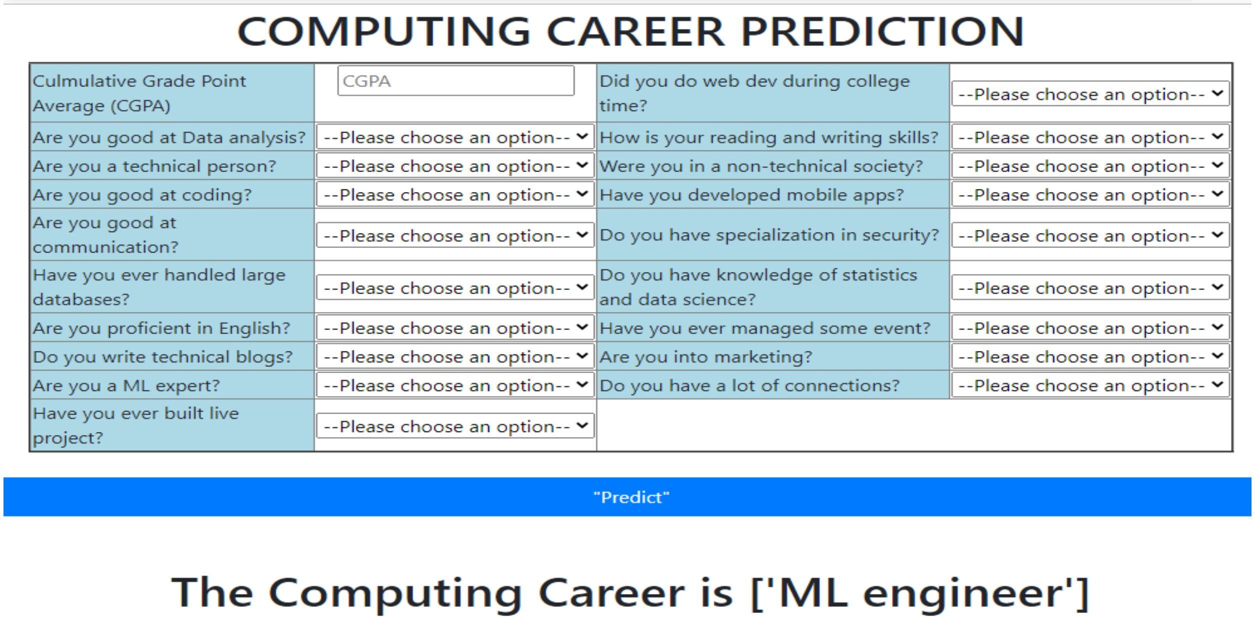


COMPUTING CAREER PREDICTION

Culmulative Grade Point Average (CGPA)	CGPA	Did you do web dev during college time?	--Please choose an option--
Are you good at Data analysis?	--Please choose an option--	How is your reading and writing skills?	--Please choose an option--
Are you a technical person?	--Please choose an option--	Were you in a non-technical society?	--Please choose an option--
Are you good at coding?	--Please choose an option--	Have you developed mobile apps?	--Please choose an option--
Are you good at communication?	--Please choose an option--	Do you have specialization in security?	--Please choose an option--
Have you ever handled large databases?	--Please choose an option--	Do you have knowledge of statistics and data science?	--Please choose an option--
Are you proficient in English?	--Please choose an option--	Have you ever managed some event?	--Please choose an option--
Do you write technical blogs?	--Please choose an option--	Are you into marketing?	--Please choose an option--
Are you a ML expert?	--Please choose an option--	Do you have a lot of connections?	--Please choose an option--
Have you ever built live project?	--Please choose an option--		

Figure 15 GUI for the proposed prototype

Figure 15 above, is a form of user interface for users to interact with the proposed prototype where they are required to enter relevant data. Once the data is entered, the user will click on the Predict button for the data to be loaded into the model.



COMPUTING CAREER PREDICTION

Culmulative Grade Point Average (CGPA)	CGPA	Did you do web dev during college time?	--Please choose an option--
Are you good at Data analysis?	--Please choose an option--	How is your reading and writing skills?	--Please choose an option--
Are you a technical person?	--Please choose an option--	Were you in a non-technical society?	--Please choose an option--
Are you good at coding?	--Please choose an option--	Have you developed mobile apps?	--Please choose an option--
Are you good at communication?	--Please choose an option--	Do you have specialization in security?	--Please choose an option--
Have you ever handled large databases?	--Please choose an option--	Do you have knowledge of statistics and data science?	--Please choose an option--
Are you proficient in English?	--Please choose an option--	Have you ever managed some event?	--Please choose an option--
Do you write technical blogs?	--Please choose an option--	Are you into marketing?	--Please choose an option--
Are you a ML expert?	--Please choose an option--	Do you have a lot of connections?	--Please choose an option--
Have you ever built live project?	--Please choose an option--		

The Computing Career is ['ML engineer']

Figure 16 proposed prototype prediction output.

Figure 16 is an output sample of the proposed prototype on prediction of a Machine Learning engineer. The prediction occurs when the user enters all relevant data on the GUI and clicks on the prediction button.

4.7 Prototype Testing

The following were the results of different users after using the prototype.

Table 4 Prototype Testing Results

SNO	Users	No of tests	Accepted Prediction	% Accepted Prediction
1.	Career counselors	17	15	88.24%
2.	Human Resource	15	11	73.33%
3.	IT Professionals	13	12	92.3%
4.	Graduate Computing Students	26	21	80.8%

In all tests done by the career counselors, 88.24% of the predictions conformed with the expectation of the counselee. For all predictions by the human resource, 73.33% of the prediction were acceptable. Out of thirteen IT professionals, 92.3% of the prediction were acceptable and 81.25% of the twenty six sampled graduate computing students was as their career expectations.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

The research’s results are communicated in this chapter together with the conclusions and recommendations, along with ideas for future research subjects. The first section summarizes the research findings, as well as the accomplishments made possible by performing this study. The recommendations and conclusion are outlined in the second portion of this chapter. The goal is to show that the offered conclusion and recommendations follow logically from the results analyzed.

5.2 Research findings summary

The objectives of the study have been realized. One of the objectives was to determine the fundamental career prediction parameters. In literature review, the parameters identified were overall average grades for the student and background information. Random forest outperformed Naïve Bayes and was used to predict careers successfully and was considered as the best machine learning algorithm for creating the career prediction model. A prototype has been developed and used to evaluate the computing career.

5.3 Conclusion

Students considering higher education need to understand their capabilities and choose courses based on their passion. The results of the career prediction model developed will be employed by computing college students to identify their speciality and recruiters or job-providers.

The fundamental career prediction parameters for college students were identified from previous research. The fundamental parameters are categorized as social factors, environmental factors, skills required to be successful in the career and characteristics of a computer professional. They include academic accomplishment, professional skills and abilities, aptitude, communication skills, reading and writing skills, analytical skills, team player, personal interest, professional experience and environment surrounding.

The dataset contained input and expected classes or output, therefore a supervised classification algorithms were identified. For the construction of the computing career system, the supervised algorithms, RF and NB were considered because of their excellent prediction accuracy.

NB and RF machine learning classifiers have been used in the research to predict the careers of graduate students using the same dataset with different number of features. The figure of attributes were five, ten, fifteen and nineteen. Python programming language was used to implement the classifiers. With an accuracy of 89.885 percent and an F1 score of 79.972 percent using 19 features, the Random Forest classifier surpassed the Naive Bayes classifier. Computing career prototype developed using Random Forest was very effective in prediction based on the students background, skills and college grade achieved.

Prediction using One-Hot encoded data require many attributes to be provided as inputs to increase accuracy. In table 3, all algorithms performance was low for five attributes and it increased gradually with an increase of attributes with the highest performance metrics obtained with nineteen attributes.

5.4 Challenges

One of the challenge was to get the dataset with academic performance for specific units for college student and background information. The dataset had cumulative grade point average (CGPA) which did not give specific performance of the units done by the student. This could have improved the accuracy of the model.

Another limitation was the number of career roles in the dataset. It contained ten career roles in computing where as there are more career areas.

5.5 Limitations

This model will not be able to address all specialties in computing as technology is dynamic and changing every day.

The model will not be able to cater to other students in colleges who are undertaking courses that are not related to computing.

5.6 Recommendations

By assessing and scrutinizing students' analytical, technical, memory-based, logical, general awareness, psychometric, skill-based examinations and interest, a more complicated web application may be developed in which inputs are not directly provided but rather student parameters are established.

Future research should base on datasets with more than nineteen attributes in order to obtain the optimal performance metrics with increasing number of attributes.

REFERENCES

Abd Alazeez, Y., & Thaher, A. (2021). Data Mining Between Classical and Modern Applications: A Review. *AL-Rafidain Journal of Computer Sciences and Mathematics*, 171-191.

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., et al. (2021). Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Sciences*.

Alwosheel, A., Sander, v. C., & Chorus, G. C. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of choice modelling*, 167-182.

Bandura, A., Caprara, G. V., Barbaranelli, C., Pastorelli, C., & Regalia, C. (2001). Sociocognitive self-regulatory mechanisms governing transgressive behavior. *Journal of Personality and Social Psychology*, 125–135.

Brownlee, J. (2019). How to Choose a Feature Selection Method For Machine Learning. *Machine Learning Mastery*.

Eiroa-Orosa, F. J. (2020). Understanding Psychosocial Wellbeing in the Context of Complex and Multidimensional Problems. *International Journal of Environmental Research and Plublic Health*, 5937.

Gerhana, Y. A., Fallah, I., Zulfikar, W. B., Maylawati, D. S., & Ramdhani, M. A. (2019). Comparison of naive Bayes classifier and C4.5 algorithms in predicting student study period. . *Journal of Physics: Conference Series*.

Godden, B. (2004). Sample Size Formulas: http://williamgodden.com/samplesizeformula.pdf.

Gruger, J., & Schneider, J. G. (2019). Automated Analysis of Job Requirements for Computer Scientists in Online Job Advertisements. *WEBIST*, 226-233.

Jeyarani, R., Nagaveni, N., & Ram, R. V. (2013). Self adaptive particle swarm optimization for efficient virtual machine provisioning in cloud. In *In Organizational Efficiency through Intelligent Information Technologies* (pp. 88-107). IGI Global.

Kabari, L. G., & Agaba, F. (2019). An Intelligent Career Advisor Expert System. . *International Journal of Advanced Research and Publications*, 91-94.

Kazi Afaq, A., Sharif, N., & Ahmad, N. (2017). Factors influencing students' career choices: empirical evidence from business students. *Journal of Southeast Asian Research*, 1-15.

Kazi, A. S., & Akhlaq, A. (2017). Factors Affecting Students' Career Choice. *Journal of Research & Reflections in Education (JRRE)*.

Lagman, A. C. (2019). Embedding naïve Bayes algorithm data model in predicting student graduation. *In Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering*, (pp. 51-56).

Macharia, M. S. (2019). Dr Reengineering mass career acquisition through technical vocational education training counseling in Kenya. *International Journal of Research in Business and Social Science*, 212-218.

Malik, S. I., & Al-Emran, M. (2018). Social Factors Influence on Career Choices for Female Computer Science Students. *International Journal of Emerging Technologies in Learning*, 56-70.

Melian, L., & Nursikuwagus, A. (2018). Prediction Student Eligibility in Vocation School with Naïve-Byes Decision Algorithm. *IOP Conference Series: Materials Science and Engineering.* Bandung, Indonesia: IOP Publishing Ltd.

Michiharu, Y., Yunqi, L., Thanh, T., Yongfeng, Z., & Dongwon, L. (2022). Looking Further into the Future: Career Pathway Prediction. *In Proceedings of the First International Workshop on Computational Jobs Marketplace.*

Mualuko, N. J. (2007). The issue of poverty in the provision of quality education in Kenyan secondary schools. *Educational Research and review 2, no. 7*, 157-164.

Mulyati, S., & Setiani., N. (2018). IDENTIFYING STUDENTS'ACADEMIC ACHIEVEMENT AND PERSONALITY TYPES WITH NAIVE BAYES CLASSIFICATION. *Sebatik*, 64-68.

Nie, M., Yang, L., Sun, J., Su, H., Xia, H., Lian, D., et al. (2018). Advanced forecasting of career choices for college students based on campus big data. *Frontiers of Computer Science*, 494-503.

NIE, M., YANG, L., SUN, J., SU, H., XIA, H., LIAN, D., et al. (2018). Advanced forecasting of career choices for college students based on campus big data. *Front. Comput. Sci. 12*, 494–503.

Nunsina, T., & Situmorang, Z. (2020). Analysis Optimization K-Nearest Neighbor Algorithm with Certainty Factor in Determining Student Career. *In 2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)* (pp. pp. 306-310.). IEEE.

Ojenge, W., & Muchemi, L. (2008). Career Guidance Using Expert System Approach. *Information Systems*, 123-131.

Rangnekar, R. H., Suratwala, K. P., Krishna, S., & Dhage, S. (2018). Career prediction model using data mining and linear classification. *In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.

Razaque, F. N. (2017). Using naïve bayes algorithm to students' bachelor academic performances analysis. *In 2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1-5). IEEE.

Reddy, V. M. (2021). Career Prediction System. *International Journal of Scientific Research in Science and Technology*, 54-58.

Roy, K. S., Roopkanth, V. U., Bhavana, V., & Priyanka, J. (2018). Student Career Prediction Using Advanced Machine Learning Techniques. *International Journal of Engineering & Technology*, 26-29.

Sampurno, T., Hediyantama, B., & Widiyati., M. A. (2019). Predicting Candidates For Fit And Proper Test Using K-Nearest Neighbor. *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)* (pp. 413-416). Yogyakarta, Indonesia: IEEE.

Sharma, A., & Dey, S. (2012). A Comparative study of Feature Selection and Machine Learning Techniques for Sentiment Analysis. *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, (pp. 1-7).

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between multinomial and Bernoulli naïve Bayes for text classification. *In 2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 593-596). IEEE.

Steffen, H., Hajo, W., Dorothea, S., & Steffen, I. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM Model. *Science Direct*, 403- 408.

Subahi, A. F. (2018). Data Collection for Career Path Prediction Based on Analysing Body of Knowledge of Computer Science Degrees . *Journal of Software*, 533-546.

Subudhi, A., Dash, M., & Sabut, S. (2019). Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier. *Biocybernetics and Biomedical Engineering*, 277-289.

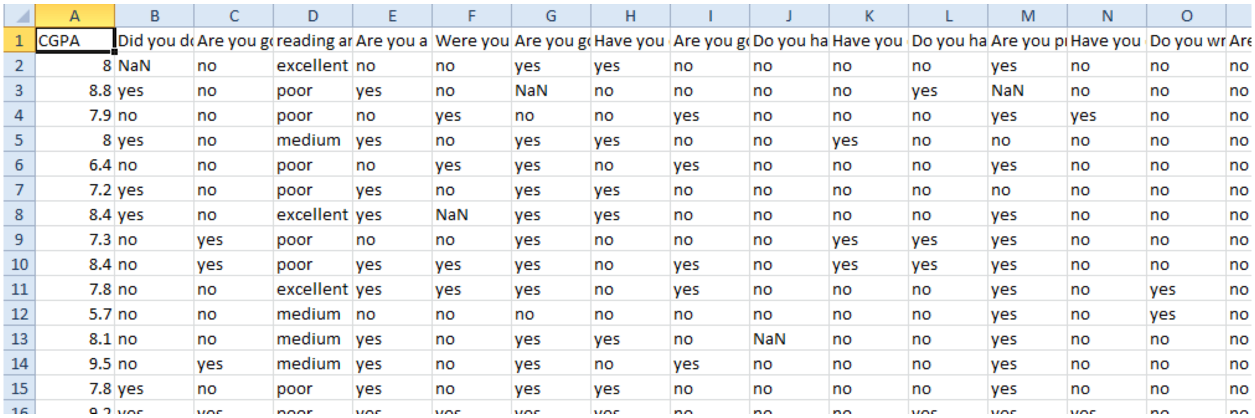
Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, 3-26.

VidyaShreeram, N., & Muthukumaravel, A. (2021). Student Career Prediction Using Machine Learning Approaches. *Proceedings of the First International Conference on Computing, Communication and Control System.* Chennai, India: EAI.

Wong, B., & Kemp, P. (2017). Technical boys and creative girls: the career aspirations of digitally skilled youths. *Cambridge Journal of Education*, 306.

APPENDICES

Appendix A: Sample data 4 ver1 dataset for model building



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	CGPA	Did you d	Are you g	reading ar	Are you a	Were you	Are you g	Have you	Are you g	Do you ha	Have you	Do you ha	Are you p	Have you	Do you wr	Are
2	8	NaN	no	excellent	no	no	yes	yes	no	no	no	no	yes	no	no	no
3	8.8	yes	no	poor	yes	no	NaN	no	no	no	no	yes	NaN	no	no	no
4	7.9	no	no	poor	no	yes	no	no	yes	no	no	no	yes	yes	no	no
5	8	yes	no	medium	yes	no	yes	yes	no	no	yes	no	no	no	no	no
6	6.4	no	no	poor	no	yes	yes	no	yes	no	no	no	yes	no	no	no
7	7.2	yes	no	poor	yes	no	yes	yes	no	no	no	no	no	no	no	no
8	8.4	yes	no	excellent	yes	NaN	yes	yes	no	no	no	no	yes	no	no	no
9	7.3	no	yes	poor	no	no	yes	no	no	no	yes	yes	yes	no	no	no
10	8.4	no	yes	poor	yes	yes	yes	no	yes	no	yes	yes	yes	no	no	no
11	7.8	no	no	excellent	yes	yes	yes	no	yes	no	no	no	yes	no	yes	no
12	5.7	no	no	medium	no	no	no	no	no	no	no	no	yes	no	yes	no
13	8.1	no	no	medium	yes	no	yes	yes	no	NaN	no	no	yes	no	no	no
14	9.5	no	yes	medium	yes	no	yes	no	yes	no	no	no	yes	no	no	no
15	7.8	yes	no	poor	yes	no	yes	yes	no	no	no	no	yes	no	no	no
16	9.2	yes	yes	poor	yes	yes	yes	yes	no	no	no	yes	yes	yes	no	no