A Study of gainful employment of learners receiving skills training in the informal sector using Machine Learning

by

Syed Ahsan Raizan
16201057
Sayed Tanjim Alam
16201076

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
January 2021

© 2021. BRAC University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Syed Ahsan Raizan
16201057

Sayed Tanjim Alam
16201076

Approval

The thesis/project titled "A Study of gainful employment of learners receiving skills training in the informal sector using Machine Learning" submitted by

1. Syed Ahsan Raizan (16201057)
2. Sayed tanjim Alam (16201076)

Of Fall, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January 15th, 2021.

Examining Committee:

Supervisor:
(Member)

Md. Khalilur Rahman, PhD
Associate Professor
Department of Computer Science and Engineering
BRAC University

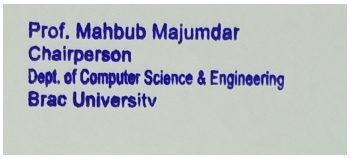
Co-Supervisor:
(Member)

Shifur Rahman Shakil
Manager
Skills Development Programme
BRAC

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)



Mahbubul Alam Majumdar, PhD
Professor and Chairperson
Department of Computer Science and Engineering
BRAC University

Ethics Statement

We used background data of learners involved in receiving training from BRAC Skills Development Programme (SDP), provided by BRAC. There was a data sharing agreement, based on which BRAC Skills Development Programme (SDP) shared this data. The data provided was void of any personal identifiers that could be associated with a learner, so this study is confident that all the learners will remain anonymous. This data was used only for research purposes.

Abstract

Analyzing data on the populace involved in BRAC's Skills Development Programme (SDP), receiving training for jobs and businesses in the informal sector of the economy, research aims to predict and/or classify whether or not learners had a gainful employment, based on background data inferred from past learners and to see how efficiently different machine learning algorithms can achieve this. As this work involves indirectly helping people who work in the informal sector, it is safe to assume that most of the learners will ask to be enrolled in trades that they see the majority pursuing, instead of making an informed decision. The observations from the research aims to find how effectively different machine learning algorithms discover correlation between a learner's background data and their chances of success in securing lucrative employment compared to their peers, and grouping learners into groups of successful and unsuccessful categories to determine how the performance of learners are in the job sector after receiving training. Some of the investigating criteria are their backgrounds, post training information and current salary, to name a few.

Keywords: Primary Data, Machine Learning, Supervised Learning, Unsupervised Learning, Informal Economy, Apprenticeship Program

Acknowledgement

First and foremost, we would like to thank Brac University for providing us with an opportunity to learn and continue our research, even during the COVID-19 pandemic.
Furthermore, we would like to express our deepest gratitude to our research supervisor Md. Khalilur Rahman PhD. for his valuable and constructive suggestions during the planning and development of this research work. His eye for details has vastly helped us improve some of the more delicate aspects of our work.
We would also like to immensely thank Mr. Shifur Rahman Shakil, our research co-supervisor, for his patient guidance, enthusiastic encouragement and useful critiques of this research work. His willingness to give his time so generously has been very much appreciated. And finally, we would like to thank Brac for facilitating access to the Skills Development Programme data, which made this research possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Despite being one of the world's quickest developing economies, Bangladesh has an unemployment rate of roughly 4.3% from 2017-2018 and only came down to around 4.2% in 2019 [1] [2]. Although the rates do not look too concerning statistically compared to other developing countries, reality speaks of a far more grim situation. According to the World Bank, Bangladesh had a population of 163 million in 2019 [3]. A third of the population of Bangladesh is 10-24 years old and 2.2 million young people join the workforce each year. Every 3 out of 4 business pioneers in Bangladesh have communicated concerns with respect to the accessibility of a skilled workforce, while roughly 10 million youngsters are accounted for to be underemployed or unemployed [4]. It is safe to say that taking the number of unaccounted underemployed or unemployed youngsters into consideration, the number would be significantly higher. It is extremely important to train and educate these youths in the informal sector as a lot of them tend to be the breadwinner of their family. Furthermore, the unemployed populace in the informal sector put indirect pressure on their families as their contribution to the family's income is null. As most of these families are already around the poverty line, this hampers them greatly. Moreover, they tend to lower the standard of living in general. BRAC's Skills Development Programme (SDP) aims to create jobs in the informal sector of Bangladesh's economy by training the enrolled learners in a specific discipline from a vast array of fields.

1.2 Problem Statement

According to the data in BRAC Skills Development Programme's (SDP) website [4], the job placement rate for SDP learners is astoundingly high (83.36%). As mentioned before, the SDP works with the informal sector and workers in that sector have a very low level of income. Skills Training for Advancing Resources (STAR) is an on-the-job apprenticeship model that equips underprivileged youth with the skills that employers need. It works on a scale which has educated more than 63,000 apprentices, minimizing early marriage and growing incomes dramatically [4]. Young learners study in 25 different demand-driven trades under master crafts persons. 95% remain in jobs or become entrepreneurs after graduation. Another project, Pro-poor Growth of Rural Enterprises through Sustainable Skills Development (Progress), works to catalyze the growth in the light engineering field of micro

and cottage businesses. To help 5,000 enterprises and tie them together with 10,000 young people, it leverages BRAC's experience with popular apprenticeship models. Bangladesh's light engineering potential is strengthened by establishing these partnerships, and young people acquire useful job skills [4]. The informal economy is growing at a rate of 2.4% per annum, through the initiatives taken by BRAC SDP [5]. With 50% of its learners being female, SDP managed to reduce early marriages by 62%. Moreover, 11% of its learners are people with disabilities. Despite so much work done by BRAC, there still exists opportunities to improve the training program, thereby paving the way for more female and people with disabilities learners to improve their lifestyles. Previously, there have been some effective attempts to increase the job placement rate by looking into businesses' mentality towards people with disabilities learners and working accordingly [6]. Due to these interventions, the aforementioned employment rate of 83.36% was achievable. According to a survey [5], some of the highest paying trades in the informal sector are: graphic design, welder, basic electronic technology, automobile, wood furniture design, wood furniture and beauty salon. The biggest issue here is the lack of any previous evidence for the people who have received these training courses.
In the formal sector, there are numerous studies on learners that can predict learner performance, predict behaviour, etc due to the huge amount of individual data available. Extensive research and money is spent on studies pertaining the formal sector. However, the same cannot be said for the informal sector. The scarcity of information in the latter sector makes it extremely difficult to apply predictions or classification approaches on the populace in the sector. Although there is information available when these people are enrolled in the Skills Development Programme (SDP) training courses from BRAC, it is still a trickle compared to the information available on anyone in the formal sector. Moreover, these learners rarely choose the trades they would like to receive training on, based on a well informed decision. They are more likely than not are influenced by their relatives and peers in choosing their training, which may or may not be the most suitable trade for them to train in.

1.3 Aim of Study

The aim of this research is to look into the background data of the learners in the Apprenticeship programme in order to identify successful individuals based on key and/or minor background features and use that information to help identify the trades best suited for future learners with similar backgrounds. Learner's background, geographical position and post training status were used to determine patterns in the data. These were analyzed to see if any new information could be generated about the learners and how it could weigh into determining the target variable. Based on the analysis, this research attempts to have a system be developed and put into place to identify attributes suitable for a trade to be successful in, or suggest the suitable trade based on certain attributes.

1.4 Research Methodology

This research involves exploratory data analysis, as well as explores the usage of machine learning, namely supervised and unsupervised learning, in the development sector. The use of machine learning in the development sector is rather new and has the potential to uncover new insights along with the help of the existing data mining approach. As stated before, the end goal of this research is to predict and /or classify whether or not learners had a gainful employment, based on background data and to see how efficiently different machine learning algorithms can achieve this. The provided data set did not have any clear output variable to classify/predict the outcome. Hence, an output variable was fabricated from the available variables with "Monthly Earning" carrying the defining weight in calculations for the output. Mean and standard deviation of the "Monthly Earning" variable was calculated and grouped using learners' training year and selected trade. These two new variables along with the previously mentioned "Monthly Earning" variable were used to calculate a new variable called "Earning Ratio". For supervised learning, regression algorithms used "Earning Ratio" as the target variable due to it being a continuous variable. The target variable had to be transformed to binary labels for classification algorithms. For unsupervised learning, we used clustering models such as K-means Clustering and Hierarchical Agglomerative Clustering.

1.5 Thesis Outline

This report describes the necessity, application and construction of a prediction model that would be advantageous to both SDP and its learners, as well as the development sector in general. This would help to make sure Brac's resources are properly utilized to provide its learners the edge needed to land jobs. The use of existing supervised and unsupervised machine learning models on SDP's dataset is described in this report.
Chapter 1, the introductory section, lays out the inspiration behind studies that motivated the authors to tackle this specific problem statement. The aim of this study and the summary are briefly discussed here.
Chapter 2, the related works section, discusses articles that presented related concerns. In addition, some scientific records that relate to the secondary data used are listed. The goal of the context analysis was to figure out the brief outcomes of previous studies.
Chapter 3, the data collection and feature selection section, outlines the quantity and quality of the data set and the preprocessing phase. A summary of the data set was also included. The feature selection portion demonstrated how to reduce the immense number of features to reduce the redundancy and the time complexity. The relevance and importance of indicators with respect to the target variable is focused in feature analysis.
Finally Chapter 4, Model Selection and Result Analysis, includes the proposed machine learning models and the comparative study between the corresponding models. Analysis between different supervised and unsupervised models are outlined with their respective model evaluations.

Chapter 2

Related Work

Analyzing and proper classification of data to determine trends and characteristics has great significance to all organizations in this information era. Even more so, choosing the right tool to process and observe the said data is of paramount importance. And with the exponential growth in data, the need to extract meaningful information to infer useful understanding arises.

Numerous research has been conducted on data analysis of successful teachers in different fields [7] [8] [9], however research on successful students is lacking compared to the former. Schibeci et al.1986 [10] had investigated the influence of students' background and perceptions on science attitude and achievement. They had used causal modeling procedures, in particular the LISREL method to analyze the data of 17-year-olds from the 1976-1977 National Assessment of Educational Progress (NAEP) survey. According to their paper, they examined the influence of five background variables(sex, race, home environment, amount of homework, and parents' education) on three dependent variables (student perception of science instruction, student attitudes, and student achievement).

Coleman et al.1966 [11] had previously analyzed the Equality of Educational Opportunity Survey (EEOS) data to find the differences between the characteristics of schools attended by minority and White students. Their research found the differences to be remarkably few and had summarized family inputs to have a greater weight than school inputs in predicting student achievements.

Konstantopoulos et al. 2011 in another research, approached the study by Coleman et al. 1966 and used modern statistical models to determine the predictive effectiveness of the school characteristics on student achievement net of family history effects [12]. They reanalyzed the EEOS 12th grade data using multilevel models and regression. Their research concluded that school had a significant effect on shaping student achievements, as opposed to the conclusion proposed by Coleman et al.

Kaur et al. 2015 [13] in their research, worked on developing data mining algorithms focused on classification and prediction in order to predict slow learners in the education sector. Their paper also discussed that students' academic success is not a result of only one determining factor, but also depends heavily on multiple factors such as personal, socioeconomic, psychological and other environmental variables.

Perhaps the closest work similar to this research was published in the International Journal of Advances in Scientific Research and Engineering (IJASRE) where a study of the Nigerian informal sector was performed using data mining approaches [14]. The study used data gathered by the National Salaries, Incomes and Wages Com-

mission (NSIWC) on the informal sector of the Nigerian economy between the year 2014 and 2016.
The lack of prior studies in the development/social sector similar to this research presented somewhat of a challenge over the course of the research. Few more limitations are discussed further into the report.

Chapter 3

Data Collection and Feature Selection

3.1 Data Collection

The informal economy, having no regulatory body present various challenges to those employed and to employers. The data used in this research has been collected from BRAC's Social Development Program's Apprenticeship Program for the years 2017 and 2018. To ensure the research maintains complete impartiality, no names or specific identifiers were present in the data set.The data is of applicants to the Apprenticeship program. Each learner admitted are school dropouts from impoverished families. The two data sets used are Preparatory Phase Data and Post Training Phase Data. The Preparatory Phase Data is data collected from applicants to the program, based on their background information. This dataset contained data from applicants from the years 2017 and 2018. The Preparatory Phase contains data on 15000 learner applicants. These background data are the basis of this research. The table below contains a list of features available from the data set, their data type and a brief description of each.

Feature Name	Data Type	Description
Branch	String	The branch the learner applied to
Cohort Lookup	String	The year the learner applied
Learner ID	String	ID assigned to applicant
Learner's Sex	String	Learner's sex
Learner's Religion	String	Learner's religious identity
Learner's Family Size	Integer	Learner's family size
Current Division	String	Current Division learner lives at
Current District	String	Current District learner lives at
Current Upazila	String	Current Upazila learner lives at
Current Area	String	Current Area learner lives at

Permanent Division	String	Division of learner's family home
Permanent District	String	District of learner's family home
Permanent Upazila	String	Upazila of learner's family home
Permanent Area	String	Area of learner's family home
SSC Dropout	Boolean	Did the learner drop out before his/her SSC
Learner trained before	Boolean	Did the learner take any such training before
Learner Family eager	Boolean	Is the learner's family eager about this program
Learner Family Female Headed	Boolean	Is the main earner of the learner's family a female
Earning Members	Integer	How many earning members does the learner's family have
Household Monthly Income	String with Range	Monthly income of the learner's family
Learner Meals	Integer	Number of meals the learner consumes daily
Learner Family Savings	String with Range	Family savings the learner or his family has
Learner Family Land	String	Land owned by the learner
Learner Education	String with Range	How much has the learner completed formal schooling
Learner Dropout	Boolean	At which level did the learner drop out
Learner Marital Status	String	Is the learner married
Learner Adivasi type	String	Is the learner an Adivasi
Learner Person with Disability	Boolean	Is the learner a person with disability
Learner Person with Disability body level	String	If disabled, to which extent does his/her physical capabilities allow him/her to function
Learner Person with Disability eyesight level	String	If disabled, to which extent does his/her eyesight allow him/her to function
Learner Person with Disability voice level	String	If disabled, to which extent does his/her vocal capabilities allow him/her to function
Learner Person with Disability hearing level	String	If disabled, to which extent does his/her hearing capabilities allow him/her to function

Learner Person with Disability mental level	String	If disabled, to which extent does his/her mental capabilities allow him/her to function
Preferred first trade	String	First preference of trade the learner wants to learn
Preferred second trade	String	Second preference of trade the learner wants to learn
Preferred third trade	String	Third preference of trade the learner wants to learn

Table 3.1: Feature Name List of Preparatory Phase Data

The Post Training Dataset is of learners who have successfully completed their apprenticeship program, and have gotten into various jobs. The data contained data on 50000 graduates from various years, but for the purposes of this research, having been constrained by the Preparatory Training Data of only 2017 and 2018. The Post Training Data of only 2017 and 2018 learners were used (By joining both data sets using Learner's ID). This data set shows their current earning and employment status, along with future goals. The table below contains a list of features available from the data set, their data type and a brief description of each.

Feature Name	Data Type	Description
Branch	String	The branch the learner applied to
Trade Chosen	String	The trade the learner chose and was trained in
Learner's ID	Integer	ID for the learner
Job Type	String	Currently Employment Status, either working under the Master Craftsperson, started their own business or has been employed in the formal sector
Monthly Earning	Float	The learner's current monthly earning
Future Scope	String	Future aspirations of the learners

Table 3.2: Feature Name List of Post Training Phase Data

3.2 Feature Selection

Initially, all rows containing empty fields were dropped, and columns with incomplete or incomprehensible data were also dropped. The data describing the persons with

disability's conditions were omitted as most of the data was either incomplete, or written in vague ways. This research took into account BRAC SDP's own process of categorising applicants into groups to section them based on disadvantages in the background data. These features, and the remaining features were encoded where needed. The table below outlines the features used, the encoding methodology used, as well as explanations as to why the type of encoding algorithm was used.

Feature Name	Encoding Type	Explanation
Learner ID	No Encoding needed	Integer type unique ID field
Enrollment Year	No encoding needed	This field is based on Cohort lookup. Only the year of application was extracted from that feature. This feature was used to calculate another feature: Earning Ratio, and later dropped
Family Size	No encoding needed	Integer Type Ordinal Data
Family Savings	Helmert Encoding	This feature was a String Type Ordinal Data with Range of values. The ranges were not of the same scale. Being so, Helmert Encoding was used
Family Headed	No encoding needed	Boolean Type Ordinal Data
Disability	No encoding needed	Boolean Type Ordinal Data
Education Level	Helmert Encoding	This feature was a String Type Ordinal Data with Range of values. The ranges were not of the same scale. Being so, Helmert Encoding was used
Earning Members	No encoding needed	Integer Type Ordinal Data
Learner's Sex	One-Hot Encoding	This feature was a String Type Nominal Data having 3 different values, so one hot encoding was used
Current Division	One-hot Encoding	This feature was a String Type Nominal Data having 3 different values, so one hot encoding was used
Chosen Trade	No encoding needed	This feature was used to calculate another feature: Earning Ratio, and later dropped
Monthly Earning	No encoding used	This feature was used as the target feature

Table 3.3: Feature Name List of Chosen Features to be used

The features selected were encoded. As not all models are capable of handling all

types of data, encoding was necessary. The two types of encoding used here were One-Hot Encoding and Helmert Encoding. One-Hot Encoding as described by [15] is used where there is Categorical Data, and the data does not represent any type of pattern with each other. Each data option is taken as a separate column, using binary representation is the original data was of that type. An example is shown below in Figure 3.2.1 using the Learner's Sex Field.

Learner's Sex
Male
Female
Transgender

Table 3.4: Feature: Learner's Sex

After One-Hot Encoding of the Learner's Sex field, 3 new fields are generated, as shown below:

Sex Male	Sex Male	Sex Transgender
1	0	0
0	1	0
0	0	1

Table 3.5: Feature: Learner' Sex after One-Hot Encoding

Even though this process results in an increase in dimensions, it was necessary as this is the only way to handle this type of data.
Helmert encoding as described by [16] is used when there are categorical variables, with ranges, and each range represents non-linear sections of information. This encoding process works by comparing levels of a variable with the mean of the subsequent levels of the variable. The Helmert code works using the following formula as given in [17]:

$$\mu_i = E(Y_i) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

In the above formula, i is the type of categories of data in the feature. The resulting estimated coefficients change meaning. The corresponding to the difference in the mean response between Option 1 and Option 2, under the Helmert coding, it would represent the difference between the mean response for Option 1 and and the "mean of the mean" response for the Options 2 through to Option n groups, where n is the number of options available as:

$$\mu_1 - \frac{\mu_2 + \mu_3 + ... + \mu_n}{n}$$

To see how this coding turns into these estimates setting up the Helmert matrix and augmenting it with the estimated mean response for each option,μ_i , then use Gauss-Jordan Elimination to put the matrix in row-reduced echelon form. This will

allow us to simply read-off the interpretations of each estimated parameter from the model. This is demonstrated in the following matrices in the figure below (Shown example for 4 options as done in Features Family Savings and Education Level):

$$\left[\begin{array}{cccc|c} 1 & \frac{3}{4} & 0 & 0 & \mu_1 \\ 1 & \frac{-1}{4} & \frac{2}{3} & 0 & \mu_2 \\ 1 & \frac{-1}{4} & \frac{-1}{3} & \frac{1}{2} & \mu_3 \\ 1 & \frac{-1}{4} & \frac{-1}{3} & \frac{-1}{2} & \mu_4 \\ & & & & \end{array}\right]$$

$$\left[\begin{array}{cccc|c} 1 & \frac{3}{4} & 0 & 0 & \mu_1 \\ 0 & 1 & \frac{-2}{3} & 0 & \mu_1 - \mu_2 \\ 0 & -1 & \frac{-1}{3} & \frac{1}{2} & \mu_3 - \mu_1 \\ 0 & -1 & \frac{-1}{3} & \frac{-1}{2} & \mu_4 - \mu_1 \\ & & & & \end{array}\right]$$

$$\left[\begin{array}{cccc|c} 1 & \frac{3}{4} & 0 & 0 & \mu_1 \\ 0 & 1 & \frac{-2}{3} & 0 & \mu_1 - \mu_2 \\ 0 & 0 & 1 & \frac{-1}{2} & \mu_2 - \mu_3 \\ 0 & 0 & -1 & \frac{-1}{2} & \mu_4 - \mu_2 \\ & & & & \end{array}\right]$$

$$\left[\begin{array}{cccc|c} 1 & \frac{3}{4} & 0 & 0 & \mu_1 \\ 0 & 1 & \frac{-2}{3} & 0 & \mu_1 - \mu_2 \\ 0 & 0 & 1 & \frac{-1}{2} & \mu_2 - \mu_3 \\ 0 & 0 & 0 & 1 & \mu_3 - \mu_4 \\ & & & & \end{array}\right]$$

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & 0 & \mu_1 - \frac{3}{4}\mu_1 - \mu_2 + \frac{2}{3}[\mu_2 - \mu_3 + \frac{1}{2}(\mu_3 - \mu_4)] \\ 0 & 1 & 0 & 0 & \mu_1 - \mu_2 + \frac{2}{3}[\mu_2 - \mu_3 + \frac{1}{2}(\mu_3 - \mu_4)] \\ 0 & 0 & 1 & 0 & \mu_2 - \mu_3 + \frac{1}{2}(\mu_3 - \mu_4) \\ 0 & 0 & 0 & 1 & \mu_3 - \mu_4 \\ & & & & \end{array}\right]$$

Reading off the pivot tables we get the following formula by derivation. First for β_0

$$\beta_0 = \mu_1 - \frac{3}{4}\mu_1 - \mu_2 + \frac{2}{3}[\mu_2 - \mu_3 + \frac{1}{2}(\mu_3 - \mu_4)] = \frac{1}{4}\mu_1 + \frac{1}{4}\mu_2 + \frac{1}{4}\mu_3 + \frac{1}{4}\mu_4 +$$

Next for β_1

$$\beta_1 = \mu_1 - \mu_2 + \frac{2}{3}[\mu_2 - \mu_3 + \frac{1}{2}(\mu_3 - \mu_4) = \mu_1 - \mu_2 + \frac{2}{3}\mu_2 - \frac{1}{3}(\mu_3 - \mu_4) = \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3}$$

Then for β_2

$$\beta_2 = \mu_2 - \mu_3 + \frac{1}{2}(\mu_3 - \mu_4) = \mu_2 - \frac{\mu_3 + \mu_4}{2}$$

Finally for β_3

$$\beta_3 = \mu_3 - \mu_4$$

As shown in the above figures by derivation, by using the Helmert contrasts, β values are calculated that represent the difference between the estimated mean at the current options and the mean of the subsequent options.

3.3 Target Variable Selection

This research, done using data from an apprenticeship based, looks to find learners who have achieved gainful employment after successfully completing their training. The only qualitative means of success found in the data is the Monthly Salary. Since this is also a very good estimate of success in the real world, this was chosen as the main way to determine the success of learners..
A variable called the Earning Ratio was found from the Monthly Earning . This variable is a calculated variable found using the learner's monthly income, trade chosen and year admitted into the program. This variable is used as it takes into account the discrepancies in earnings between various trades and year of admission, which essentially takes into account inflation to a certain degree.
First, the mean monthly earning of each learner is found, grouped by the Year, then the trade chosen. The figure below shows the monthly earning per trade for the year of 2017 of the learners.

		Monthly Earning
Enrollment Year	Trade Chosen	
2017	Aluminium Fabricators	3003.921569
	Basic Electrical Technology	3653.278689
	Beauty Salon - Female	2582.619048
	Graphics Design	2446.875000
	IT Support Technician	2305.312500
	Mobile Phone Servicing	2779.530917
	Motorcycle Service Mechanic	3122.054795
	Refrigeration & Air Conditioning	2662.171053
	Tailoring and Dressmaking - Female	2550.502661
	Tailoring and Dressmaking - Male	2674.765808
	Wood Furniture Design	3229.823151

Figure 3.1: Mean of monthly earning for each trade for the year 2017

Next, the Standard deviation of the monthly earning for each trade is calculated according to year. An example of this is shown in the figure below for the year 2017.

		Monthly Earning
Enrollment Year	Trade Chosen	
2017	Aluminium Fabricators	1132.070808
	Basic Electrical Technology	1607.687340
	Beauty Salon - Female	962.043306
	Graphics Design	694.241780
	IT Support Technician	602.003919
	Mobile Phone Servicing	1043.481493
	Motorcycle Service Mechanic	1695.733893
	Refrigeration & Air Conditioning	968.991353
	Tailoring and Dressmaking - Female	928.828543
	Tailoring and Dressmaking - Male	959.670305
	Wood Furniture Design	1218.142868

Figure 3.2: Standard Deviation of monthly earning for each trade for the year 2017

Using the mean and standard deviations of the monthly earnings, the Earning Ratio is calculated using the formula:

$$EarningRatio = \frac{MonthlyEarning \quad - \quad MeanEarning}{StandardDeviationofEarning}$$

The Earning Ratio shows how many standard deviations away from the mean value the learner's monthly salary is. After calculating the Earning ratio, some values

were found which seemed excessively high or low. For example, an Earning Ratio of 12 would suggest that the learner earns 12 Standard Deviations more than the Mean for his/her trade in his/her year. Values such as these were dropped, and the Means, Standard Deviations and Earning Ratios were recalculated. For Regression learning models, this Earning Ratio was used to find another target variable column named "Successful?". For learner's with an Earning Ratio greater or equal to 0.5, the "Successful?" column was '1'. This means that they were counted as successful. For those with an Earning Ratio less than 0.5, the "Successful?" column was '0'. This means that they were counted as unsuccessful. Using a binary target variable allowed many models to perform better. This division was determined after multiple trial and error using various levels of earning ratios. Using 0.5 as the turning point yielded in the best results. This is discussed in more details in Chapter 4.
The box-plot diagram below shows the earning ratio, as well the "Successful?" column selection, and the outliers which were removed in the 1st trial of the Earning Ratio calculation. As shown above, the Earning Ratio and Successful? columns are

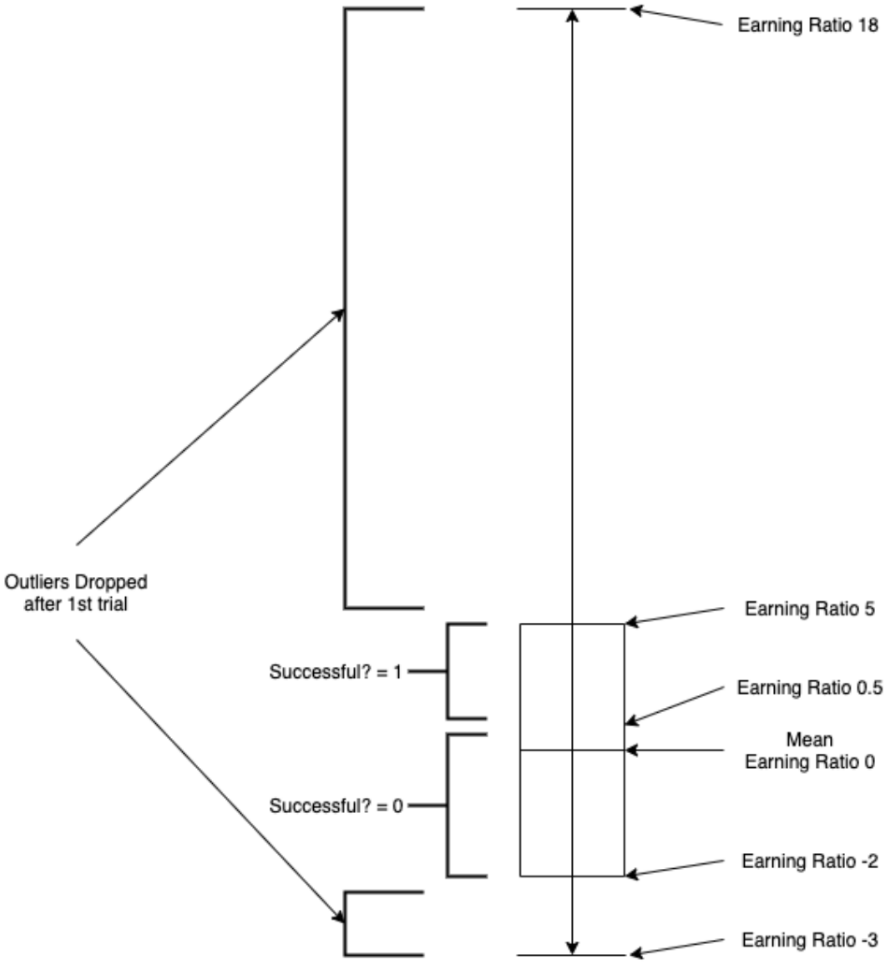


Figure 3.3: Box-plot of Earning Ratio and Successful?

closely linked. Thus both were not used together as joint Target variables in any models used.

3.4 Correlation Between Features

Following the encoding process, and generation of Earning Ratio and Successful? columns, the correlation between all the features were checked. This was done by using a correlation matrix, which was then plotted by using a heatmap. The heatmap is given in the figure below.

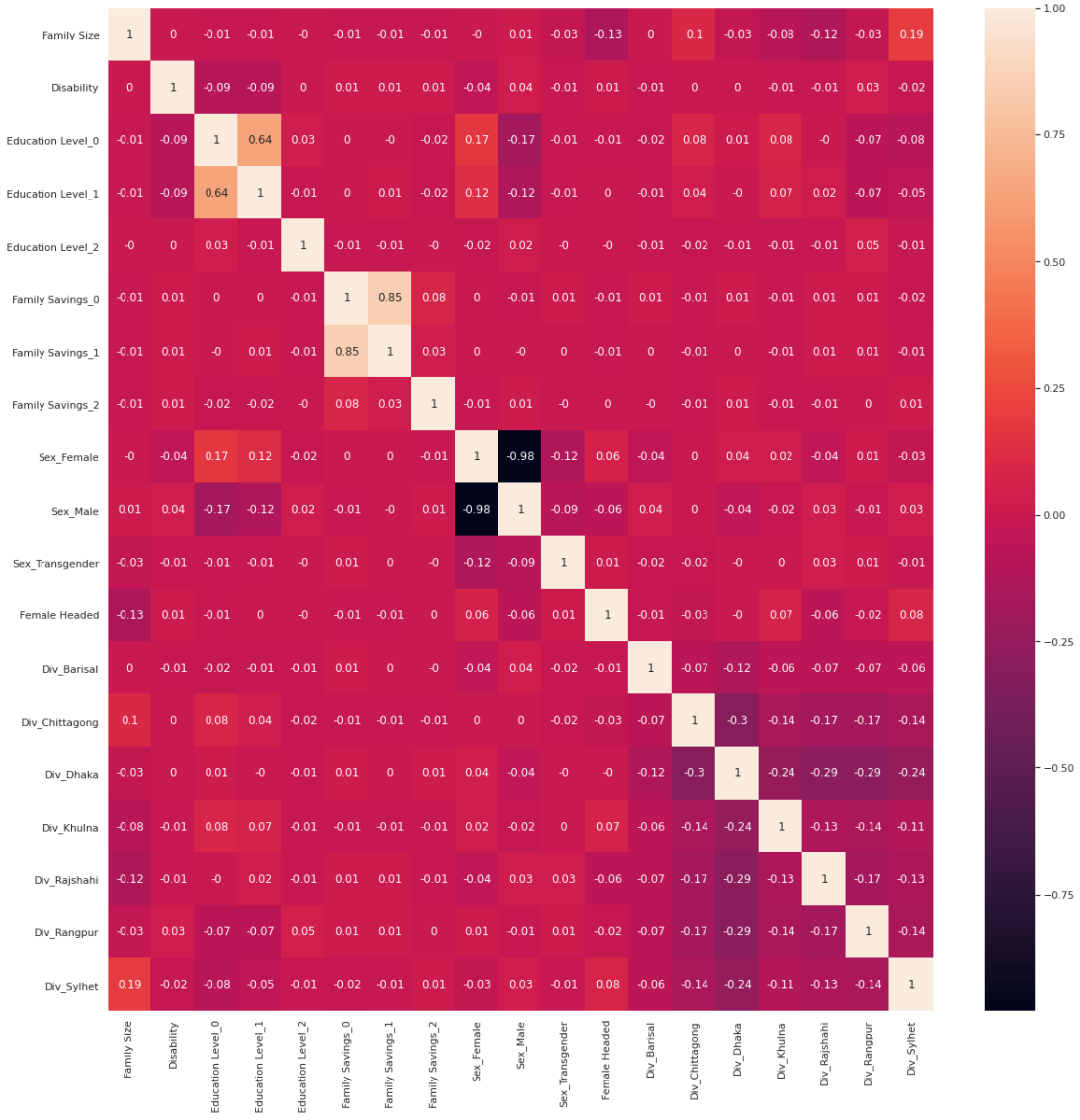


Figure 3.4: Heatmap of correlation Matrix

As seen from the above heatmap, all variables had very weak correlation to each other. This presented the biggest hurdle in using Machine Learning. Correlation is required as it can be used to predict one feature from another and might suggest causality. The only features with perfect correlation were the Monthly Earning and Earning Ratio, which is why both of these were not used as target variables together in Regression models.
But according to [18] correlation is not always causation. This is why this research looked into many different Machine Learning Models to try and test which models resulted in the best results, even when variables were not correlated.

The data was afterwards scaled using the MinMax Scalar using the process discussed by [19]. This step is done to ensure all data is normalised, to avoid bias on any feature. The MinMax scalar works by taking the minimum and maximum values of the selected feature, and scaling them to be between 0 and 1. This normalisation process is used in all models.

Chapter 4

Model Selection and Result Analysis

4.1 Supervised Learning

4.1.1 Regression Model Implementations

Regression models predict a continuous outcome variable based on the value of one or more predictor variables.
Four evaluation metrics were used in this study to assess the efficiency of regression models namely, Mean absolute error, Mean square error, Root mean squared error and R^2 score. Mean Absolute Error refers to the sum of absolute differences between the target and predicted variables. The lower the value, the better it is. Mean Squared Error is the sum of the square of the variance between the expected and actual target variables, divided by the number of data points. For this metric also, lower value translates to a better result for the model. Root Mean Squared Error is the standard deviation of the errors which occur when a prediction is made on a data set. Once again, lower value means a more robust model. The R^2 score is a statistical indicator that describes the fitting quality of a regression model. The ideal value for R^2 score is 1. The closer the value of R^2 score is to 1, the better is the model fitted.

Linear Regression

Linear regression model is represented as a linear equation integrating a particular set of input values [20][21]. The linear equation assigns one scale element, called a coefficient, to each input value or column, which is determined by beta (β). An additional coefficient is often applied, giving the line an extra degree of freedom which is also referred to as the intercept or the coefficient of bias. In a simple regression problem with a single x and a single y, the model would be:

$$y = \beta_0 + \beta_1 * x$$

Where β_0 and β_1 given by the formulas:

$$\beta_1 = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$\beta_0 = \frac{n\sum y - \beta_1(\sum x)}{n}$$

And, β_0 and β_1 are defined as:
β_0 = y-intercept of the line β_1 = Slope of the line
In higher dimensions, the line is referred to as a plane or a hyper-plane where there is more than one input (x). The representation, thus, is the equation structure and the unique values used for the coefficients.
It effectively excludes the influence of the input variable on the formula and hence, when a coefficient becomes zero, from the prediction that is made from the model. This becomes important when it comes to regularization methods that adjust the learning algorithm to minimize the complexity of regression models by putting pressure on the absolute size of the coefficients, driving sum to zero [20].
Linear regression was an initial selection, applied on the data to get an idea about how it behaves. Observing the data when fitted on a low complexity model helped in determining the machine learning models to be applied, moving further into the research.

```
Mean Absolute Error: 0.8089557169288493
Mean Squared Error: 0.9512197722324051
Root Mean Squared Error: 0.9753049637074576
R2_Score: 0.024428659652679374
```

Figure 4.1: Evaluation metrics used for Linear regression model

The above table shows the performance of the model when fitted on the learners' data. All the three error metrics have a significantly high value and are closer to 1, suggesting that the error rate in prediction for the model is very high. The R^2 score is also very low and close to zero, implying that the model fits the data rather poorly. This suggests that the linear regression model will not work with the data.

Ridge Regression

Ridge regression is a common method of regularized linear regression involving a penalty for L2. For certain input variables that do not add much to the prediction task, this has the effect of shrinking the coefficients [22]. An expansion of linear regression, ridge regression invokes the inclusion of penalties during preparation to the loss function, which supports simplified models with smaller coefficient values. One typical penalty is to penalize a model dependent on the sum of the values of the squared coefficient, called an L2 penalty. The size of all coefficients is reduced by an L2 penalty, but it prohibits any coefficients from being excluded from the model by causing their value to become zero [23]. Kuhn and Johnson [22] explained in their book that the consequence of this penalty is that the parameter values are only allowed to become high if the sum of squared errors (SSE) is decreased proportionately. In effect, as the lambda penalty becomes high, this approach shrinks the estimates to zero.

A hyper parameter called lambda (λ) is used, which controls the weighting of the loss function penalty. The penalty would be entirely weighted with a default value of 1.0 and a value of 0 excludes the penalty [23].
Ridge regression was applied because the sample size was less than a 100,000 and the number of features deemed important were more than a few [24].

```
Mean Absolute Error: 0.8090122916569377
Mean Squared Error: 0.951263201573317
Root Mean Squared Error: 0.9753272279462504
R2_Score: 0.024384118505028085
```

Figure 4.2: Evaluation metrics used for Ridge regression model

The above table shows the performance of the model when fitted on the learners' data. And once again, all the three error metrics have a significantly high value and are closer to 1, just like with linear regression. The R^2 score is also very low and close to zero, implying that the model fits the data rather poorly. This suggests that the ridge regression model with all its improvements, will still not work with the data.

Random Forest

Random forest is an ensemble of decision trees typically trained using the bagging method. It constructs and merges several decision trees to achieve a more detailed and stable prediction [25].
In Breiman's 2001 [26] approach, each tree in the set is generated by choosing each tree at random, at each node, and a small group of features to be divided on. The tree is grown to full height, without pruning, using the Classification and Regression Tree (CART) technique.
The model consists of a set of randomized base regression trees s[rn(x,θm,Dn),m 1], where θ_1,θ_2,... are independent identically distributed outputs of a randomizing variable θ. These random trees are combined to form the aggregated regression estimate:

$$rn(X, Dn) = E\theta[rn(X, \theta, Dn)]$$

Where Eθ denotes expectation with respect to the random parameter, conditionally on X and the data set Dn [26].
Random Forest was applied because the previous model, ridge regression, failed to generate conclusive accuracy over the data set. An ensemble model was the final one to determine whether or not regression models would work on the data set [24].
The table above shows the feature importance for each variable, outlining the weight that random forest assigns to each of these features whilst predicting the output. Here, Family Size has the most importance and the difference with the next most important feature is quite large.
The above table shows the performance of the model when fitted on the learners' data. This model too, shows all the three error metrics to have a significantly high value and as well as values closer to 1, just like with the other regression models.

```
Variable: Family Size                    Importance: 0.32
Variable: Div_Rajshahi                   Importance: 0.1
Variable: Education Level_0              Importance: 0.08
Variable: Female Headed                  Importance: 0.08
Variable: Disability                     Importance: 0.07
Variable: Education Level_1              Importance: 0.04
Variable: Family Savings_0               Importance: 0.04
Variable: Sex_Female                     Importance: 0.04
Variable: Sex_Male                       Importance: 0.04
Variable: Family Savings_1               Importance: 0.03
Variable: Div_Barisal                    Importance: 0.03
Variable: Div_Dhaka                      Importance: 0.03
Variable: Div_Chittagong                 Importance: 0.02
Variable: Div_Khulna                     Importance: 0.02
Variable: Div_Rangpur                    Importance: 0.02
Variable: Div_Sylhet                     Importance: 0.02
Variable: Sex_Transgender                Importance: 0.01
Variable: Education Level_2              Importance: 0.0
Variable: Family Savings_2               Importance: 0.0
```

Figure 4.3: Feature Importance table for Random Forest regression model

```
Mean Absolute Error: 0.8031728741676735
Mean Squared Error: 0.9662320914252955
Root Mean Squared Error: 0.982971053198056
R2_Score: 0.009032019691802229
```

Figure 4.4: Evaluation metrics used for Random Forest regression model

The R^2 score is also very low and close to zero and also worse than that of the previous models, implying that the model fits the data very poorly. This suggests that the random forest, and in general regression models, will not work with the data set.

4.1.2 Classification Model Implementations

Classification is the process of approximating the mapping function to discrete output variables from input variables. The primary objective is to determine which class/category the new data would fall under.
The metrics used to determine model effectiveness in this research were Accuracy score, Confusion Matrix, Precision score, Recall score and Null accuracy. Accuracy score, as the name suggests, states the accuracy of the classification algorithm. Higher score means more accuracy. The confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. The precision score is intuitively the ability of the classifier to not label a sample that is negative, as positive. The best value is closer to 100% and the worst value is closer to 0%. However, its importance is completely dependent on whether prediction precision is important or not for the data. Recall score is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. The best value is closer to 100% and the worst value is closer to 0%. This too, is dependent on whether prediction recall is important or not for the data. Null accuracy refers to the accuracy of a dumb model of the classification algorithm that always predicts the most frequent class. The lower it is compared to the accuracy score of the model, the better.
This research used 'Successful?' to be the output variable with binary labels.

Linear Support Vector Classifier

The purpose of a Linear Support Vector Classifier (SVC) is to fit the data to a hyper-plane that divides or categorizes it as best suited. From there, some features can be fed to the classifier after getting the hyper-plane to see what the predicted class is. It is implemented in terms of liblinear rather than libsvm, so it has more flexibility in choosing penalties and failure functions and can scale better to large numbers of samples.
This research started the classification approach by choosing Linear Support Vector Classifier as the number of samples were less than 100,000 [24]. Another reason was that it would converge faster to the output, as the sample size is quite large.
The above table shows the performance of the model when fitted on the learners' data. The test and training metrics are not that varied from each other suggesting that the model is able to replicate its classification accuracy from the training data set over to the test data set. The precision and recall scores are also somewhat high, further displaying the model's ability to classify the data to a certain degree of success.

K-Nearest Neighbor

K-Nearest Neighbor Classifier first finds the distances between a query and all the examples in the data. Then it selects the specified number examples (K) closest to

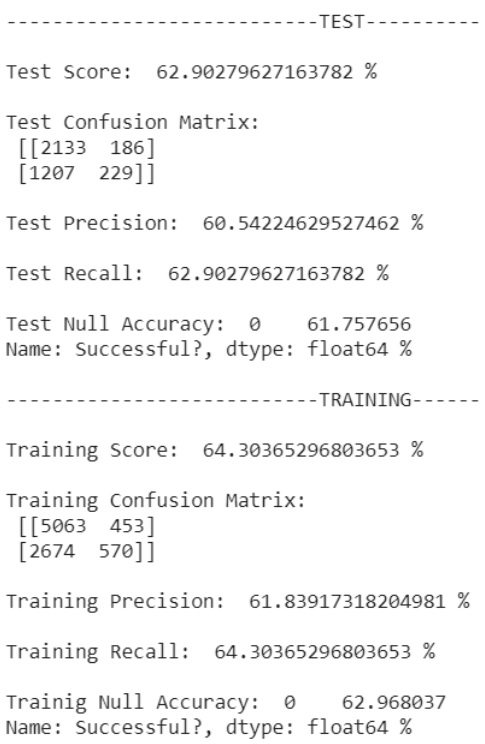


Figure 4.5: Evaluation metrics used for Linear Support Vector Classification

the query and finally votes for the most frequent label [27]. The algorithm decides a number k which is the nearest neighbor to that data point which is to be classified. If the value of k is 5, it would search for 5 neighbors closest to that data point.
The value of k is very important for this algorithm. A higher value of k instills more confidence about the accuracy of the prediction [28]. However, if the value is too high, decisions may become skewed. On the other hand, if the value of k is small, then the results would be more dependent on noise. There is a very high risk of over-fitting of the model in such situations.
The decision to choose K-Nearest Neighbor was based on the fact that the data was not textual in nature [24]. Since the previous algorithm, Linear Support Vector Classifier, was yielding average results, the search for a higher predictor accuracy was also the motivation behind choosing this algorithm.
The above table shows the performance of the model when fitted on the learners' data. Although the training scores are higher for this algorithm compared to Linear Support Vector Classifier, the test scores are actually lower. The precision and recall scores' differences for the training and test data is also seen to be higher for this model, suggesting the model performs poorly on unknown data set.

Gaussian Naive Bayes

Naive Bayes classifier is a probabilistic classifier which predicts the probability of the input being classified for all classes. The high independence between the features is an assumption for Naive Bayes. The classifier presumes that a single feature's value is independent of any other feature's value [29].
An inference often made when dealing with continuous data is that the continuous values associated with each class are spread according to a normal (or Gaussian)

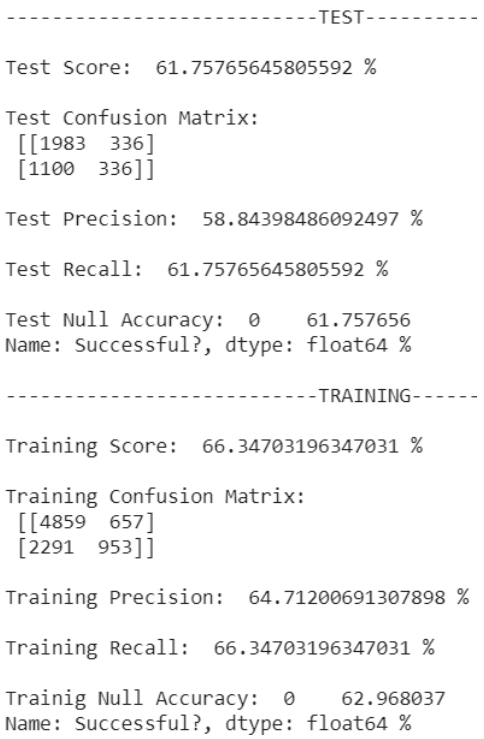


Figure 4.6: Evaluation metrics used for K-Nearest Neighbor

distribution. This means that the model implies that these values are sampled from the Gaussian distribution if predictors take continuous values instead of discrete ones [30].
Gaussian Naive Bayes was selected due to its naive approach where all the predictor variables are considered to have conditional independence between themselves. Another reason for its selection was its simplicity and speed.
Figure 4.7 shows the performance of the model when fitted on the learners' data. The performance is better than that of K-Nearest Neighbors but only slightly worse compared to Linear Support Vector Classifier. However, it has the best precision score with regards to the previous two classification algorithms.

Decision Tree Classifier

Decision Tree constructs a training model that, by studying basic decision rules derived from training data, can predict the class or value of the target variable [31]. The algorithm begins from the root of the tree to predict a class label for a record. The decision to choose Decision Tree was based on the fact that there is a high non-linearity between the dependent and independent variables.
Figure 4.8 shows the performance of the model when fitted on the learners' data. The training results show a very high performance boost in all aspects when compared to the previous three classification models. However, on the test data set, it barely outperforms K-Nearest Neighbor.

Support Vector Classifier, Kernel: Polynomial

Support Vector Machine (SVM) is based on the principle of finding a hyperplane that better divides the characteristics into various domains. The function of the

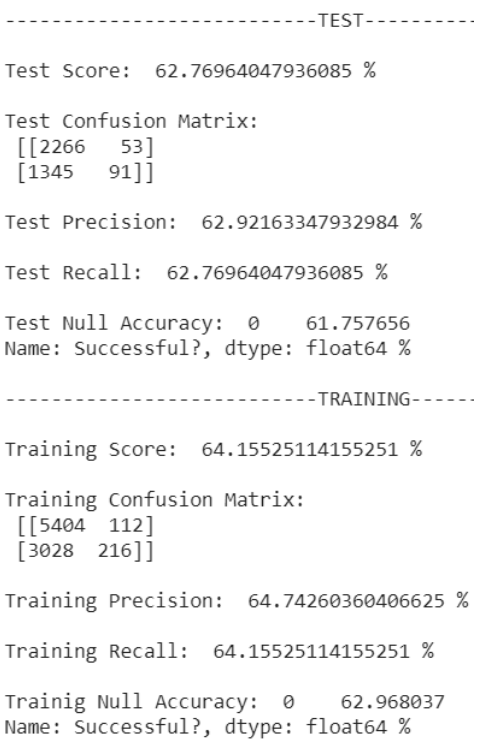


Figure 4.7: Evaluation metrics used for Gaussian Naive Bayes

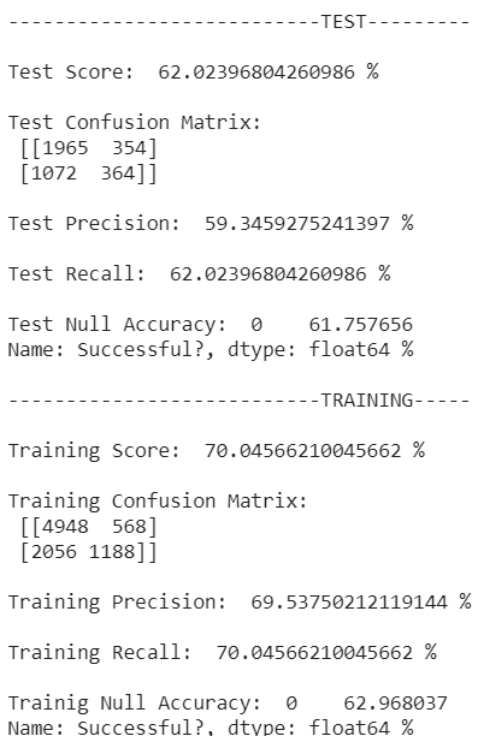


Figure 4.8: Evaluation metrics used for Decision Tree Classifier

kernel is to take data as input and transform it into required form. By applying the polynomial combination of all the current features, the polynomial kernel can be pictured as a transformer/processor to create new features [32].
The decision to choose Support Vector Classifier with polynomial kernel was based on the fact that the data set was non-linearly separable and the kernel trick would find a non-linear decision boundary.

```
---------------------------TEST----------

Test Score:  63.54194407456725 %

Test Confusion Matrix:
 [[2122  197]
 [1172  264]]

Test Precision:  61.684550487781486 %

Test Recall:  63.54194407456725 %

Test Null Accuracy:  0    61.757656
Name: Successful?, dtype: float64 %

---------------------------TRAINING------

Training Score:  66.4041095890411 %

Training Confusion Matrix:
 [[5147  369]
 [2574  670]]

Training Precision:  65.85606448425703 %

Training Recall:  66.4041095890411 %

Trainig Null Accuracy:  0    62.968037
Name: Successful?, dtype: float64 %
```

Figure 4.9: Evaluation metrics used for Support Vector Classifier

The above figure shows the performance of the model when fitted on the learners' data. The training results show a high degree of accuracy, precision and recall. On the test data set, it outperforms all the other classification algorithms used.

Multi-layer Perceptron Classifier

Multi-layer perceptrons are a mixture of several neurons linked in the form of a network, most widely referred to as artificial neural networks [30]. A multi-layer perceptron classifier has an input layer, one or more hidden layers, and an output layer [33].
Multi-layer perceptron classifier was chosen based on the its capability to learn non-linear models.
Figure 4.10 shows the performance of the model when fitted on the learners' data. The training results shows performance similar to that of the Decision Tree Classifier, only slightly worse. However, on the test data set, it performs better than all of the classification models, sans Support Vector Classifier with the polynomial kernel.
Although the six classification algorithms have almost similar performance when fitted to the data used by this research, Support Vector Classifier and Multi-layer Perceptron Classifier can be said to have the best two performances.

```
---------------------------TEST---------

Test Score:  63.19573901464713 %

Test Confusion Matrix:
 [[1983  336]
 [1046  390]]

Test Precision:  60.97438581800441 %

Test Recall:  63.19573901464713 %

Test Null Accuracy:  0    61.757656
Name: Successful?, dtype: float64 %

---------------------------TRAINING-----

Training Score:  68.37899543378995 %

Training Confusion Matrix:
 [[4915  601]
 [2169 1075]]

Training Precision:  67.4408993754311 %

Training Recall:  68.37899543378995 %

Trainig Null Accuracy:  0    62.968037
Name: Successful?, dtype: float64 %
```

Figure 4.10: Evaluation metrics used for Multi-layer Perceptron Classifier

4.2 Unsupervised Learning

4.2.1 K-Means Clustering Implementation

K-means Clustering as described by [34] et al and also by [35], is an Unsupervised iterative clustering algorithm which sections the data set into K non-overlapping and distinct groups, called clusters. It aims to make the clusters as far from each other as possible, while making the points inside as close as they can be to the centroid. The algorithm assigns data points so the sum of the squared distance between a data point and it's cluster's centroid (by using the arithmetic mean) is as low as possible. To make the data points more homogeneous within the clusters, the sum of squared distance must be the minimum value possible.
The algorithm works by first selecting the number of clusters, K. Then, the initial centroids are found by reorganising the dataset and selecting K points. This is an iterative process, which keeps being done till there is no change in the centroids. Next, the Sum of Squared distance is calculated between each data point and the centroid points. Based on the score, the points are assigned to each cluster. This process aims to solve the problem of Expectation Maximization. Mathematically this is done by the following process of E and M steps, where the E-Step is the assignment of centroids, and the M-Step is computing and calibrating the centroid. First, the object function:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_k i ||x^i - \mu_k||^2$$

Where $w_i k$=1 for data point x_i belonging to the cluster k; otherwise, $w_i k$=0. μ_k is

the centroid of the x_i cluster.
Next is a minimisation problem, divided into two parts. Firstly, the minimisation of $Jw.r.t.\ w_ik$ and treat μ_k fixed. Then, the $Jw.r.t.\ \mu_k$ is minimised and the treat w_ik fixed. Afterwards, $Jw.r.t.\ \mu_k$ is differentiated and the centroids are recomputed.
The next step, also called the E-Step is:

$$\frac{\partial J}{\partial w_ik} = \sum_{i=1}^{m}\sum_{k=1}^{K} w_ki||x^i - \mu_k||^2 \Rightarrow w_ik = \begin{cases} 1\ if\ k = argmin_j||x^i - \mu_j||^2, 0\ otherwise \end{cases}$$

The data point xi is assigned to the closest cluster in accordance to the lowest value of Sum of Squared distance between the cluster and the data point.
Next is the M-Step:

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{i=1}^{m} w_ik(x^i - \mu_k) = 0 \Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_ikx^i}{\sum_{i=1}^{m} w_ik}$$

This step is where re-computation of the centroids is done based on new assignments of data points.
To select the initial number of clusters K, the Silhouette Score and Sum of Squared Errors (SSE) [36] et al are used. The Silhouette Score is a measure of how homogeneous a data point is to other points within that cluster. The value ranges from 1 (the best score) to -1 (the worst score). A negative value signifies that the data points are wrongly clustered, whereas a 0 means that the clusters are overlapping. The SSE within clusters is calculated by summing up the squared distance between the data points and the centroid nearest to it. The error value must be as low as possible, which means that the SSE should be as small as possible. SSE being small means that the data points are close to the centroid of the cluster the point is assigned to. The combination of these two metrics are used to determine the initial value of K.
The figure shows two graphs, both charted for 50 iterations of the K-means algorithms with increasing number of clusters, finding the Silhouette Score and SSE for each.
From the first graph, ' Within Cluster SSE After K-Means Clustering ', it can be seen that as the number of clusters increases past 5, the sum of square of errors within clusters plateaus off. From the second graph, ' Silhouette Score After K-Means Clustering ', it can be seen that the Silhouette Score for most of the graph is bad. Since there is not much of a difference in SSE after 5 clusters and that the Silhouette Score is relatively good at 2 clusters, and 2 clusters provide very good SSE, the number of clusters taken is 2.
For the K-means clustering, the features taken are Monthly Earning and Earning Ratio. Using any combination of the other features presents SSE scores lower than 500 for any number of clusters, while the Silhouette Score is on average 0.1.
The table above shows the results of the first trial of K-means clustering, showing the average Monthly Earning, Earning Ratio and number of points in each cluster. This shows two very distinct groups. Cluster 0 with a very low average Monthly Earning and Earning Ratio and cluster 1 with a very high average Monthly Earning and Earning Ratio. It can be thus inferred that, cluster 0 is the less successful group, and cluster 1 is the successful group.

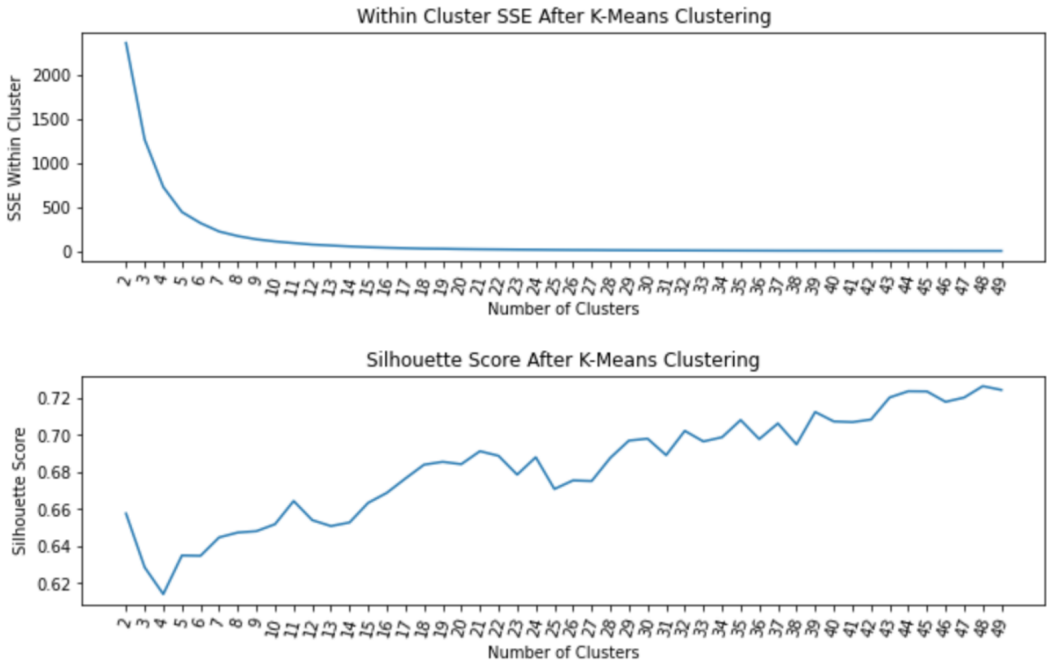


Figure 4.11: Silhouette Score and SSE for varying number of clusters

cluster	Monthly Earning	Earning Ratio	count
0	2555.949143	-0.523756	8927
1	4659.691834	1.307486	3576

Figure 4.12: Results of the first trial of K-means

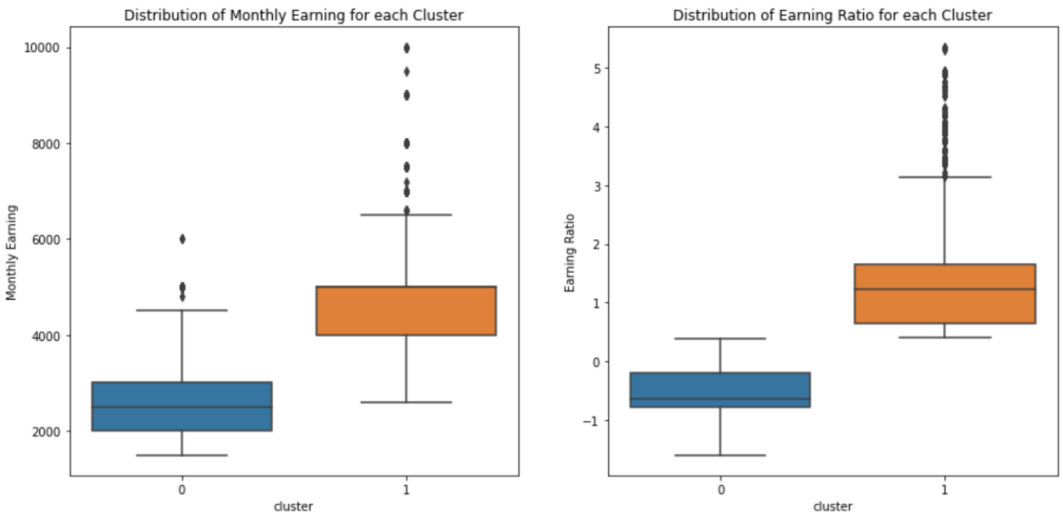


Figure 4.13: Box-Plot of results of the first trial of K-means

The above figure also shows a box-plot of the first trial. Some outliers can be seen in cluster 1. These outliers are data points which had too high Monthly Earning compared to the others in their trade.
Cluster 0 has a large population compared to cluster 1. Hence, K-means is run on that group again to try and divide this group to check if it provides a more balanced grouping of the points in cluster 0.

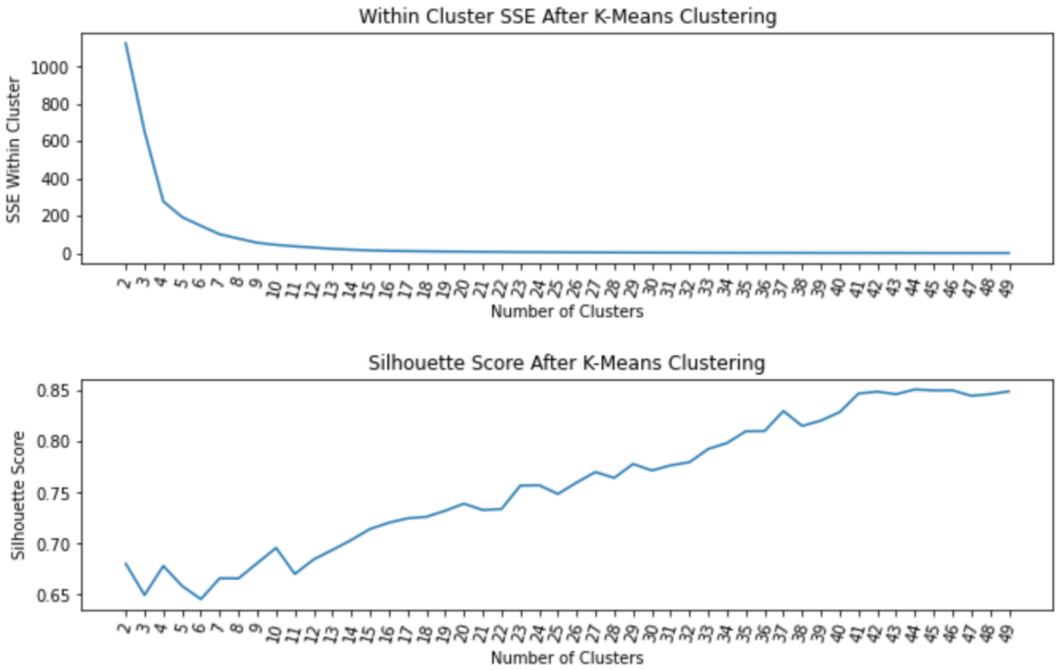


Figure 4.14: Silhouette Score and SSE for varying number of clusters from cluster 0

From the first graph, ' Within Cluster SSE After K-Means Clustering ', it can be seen that as the number of clusters increases past 4, the sum of squares of errors within clusters plateaus off. From the second graph, ' Silhouette Score After K-Means Clustering ', it can be seen that the Silhouette Score for most of the graph is bad. Since there is not much of a difference in SSE after 4 clusters and that the Silhouette Score is relatively good at 2 clusters, and 2 clusters provide very good SSE, the number of clusters taken is 2.

cluster	Monthly Earning	Earning Ratio	count
0	3055.596309	-0.074795	3468
1	2238.532698	-0.808972	5459

Figure 4.15: Results of the second trial of K-means

The above table shows the results of the second trial of K-means on cluster 0 from the first trial. This new trial divides that cluster into two new clusters. From this trial, the cluster 0 has a higher Monthly Earning and Earning Ratio. The Earning Ratio can be rounded off to 0.
The above figure also shows a box-plot of the second trial. Some outliers can be seen in both clusters. These outliers are data points which had too high or too low

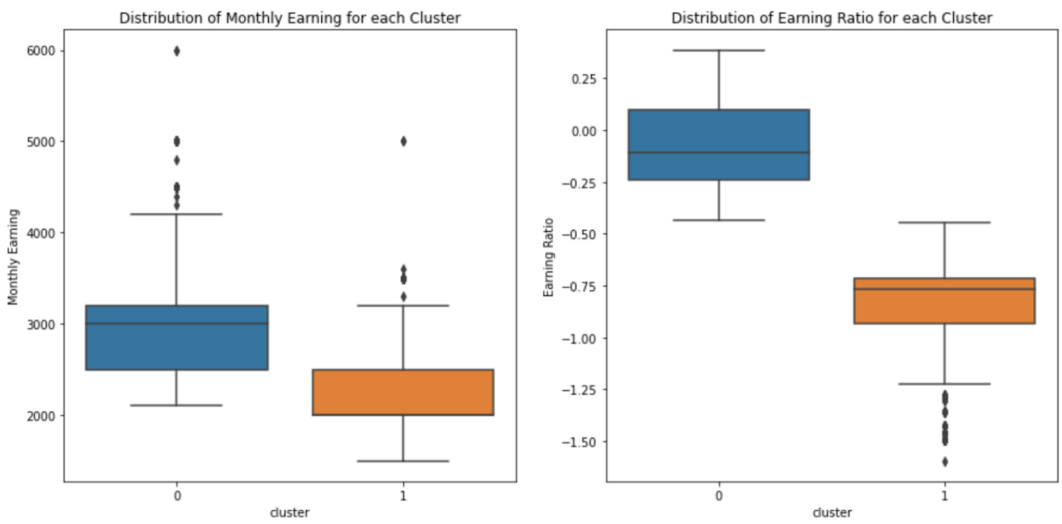


Figure 4.16: Box-Plot of results of the second trial of K-means

Monthly Earning compared to the others in their trade.
The outcome of the implementation of the K-means algorithm on this data set presents 3 groups. The 'Successful' group, which is cluster 1 from the first trial, with the highest average Monthly Earning and Earning Ratio, the "Average" group from the second trial's cluster 0, and finally the 'Unsuccessful' group from the second trial's cluster 0.

4.2.2 Agglomerative Hierarchical Clustering Implementation

The Agglomerative Hierarchical Clustering by Zhou et al [37] and also by Patlolla [38] is a bottom-up clustering approach with a complexity of $O(n^3)$ where clusters have sub-clusters, which in turn have their own sub-clusters. The algorithm works by first calculating the proximity of individual points, each of which is taken as individual clusters. Next, similar clusters are merged together. Proximity is calculated again and similar clusters are merged. This iterative process is continued till there is only a single cluster. The advantage of this process is that there is no initial selection of clusters, so that the data clusters based on the proximity of data points. The proximity is calculated by using the Squared Euclidean Distance [39]. The euclidean distance is the absolute distance between two points on a line. This formula works in multiple dimensions. This is given by:

$$d^2(p,q) = (p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2$$

For deciding which clusters to combine, a dissimilarity measure is needed. This is done by using any of a few available metrics. For this research, we have chosen the Ward Metric [40]. With this method, groups are made such that the pooled within-group's sum of squares is minimized. Which is to say, for each step needed, two clusters are joined together resulting in the minimal increase in the pooled

within-group sum of squares. This can be formulated by:

$$d_w(A,B) = \frac{nm}{n+m}\left[\left(\frac{1}{n}\sum_{a\epsilon A}\right) - \left(\frac{1}{n}\sum_{b\epsilon B}\right)\right]^2$$

where d: D×D→R+ is a distance metric on the underlying space of the data, such as:

$$d_2(A,B) = ||a+b||_2 \quad \text{or} \quad d_\infty(\alpha,\beta) = max|\alpha_i - \beta_i| assuming D = R_k$$

A Dendrogram as described by Podani [41], which is a tree diagram which shows the clusters being split into their sub clusters. The vertical axis of the dendrogram represents the distance or dissimilarity between clusters. The horizontal axis represents the objects and clusters. This is used to demonstrate the outcome of the algorithm. As the number of clusters cannot be decided, the level is selected. For this research, using the features Monthly Earning and Earning Ratio, the level was kept to 1. Higher levels resulted in very high dissimilarity. The dendrogram is given below.

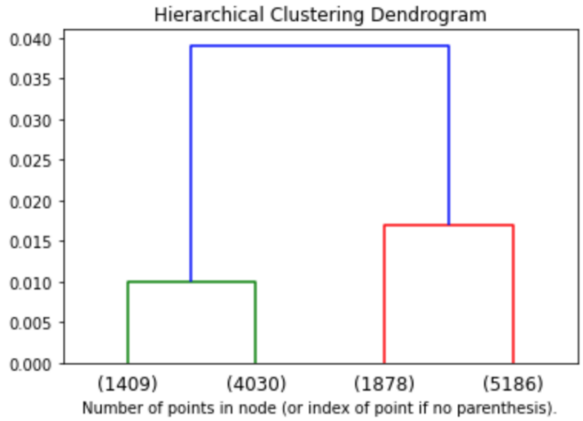


Figure 4.17: Hierarchical Clustering Dendrogram

cluster	Monthly Earning	Earning Ratio
0	2228.948336	-0.809965
1	3854.504827	0.512783
2	5145.265176	1.847443
3	2893.640854	-0.208603

Figure 4.18: Result of Agglomerative Hierarchical Clustering

The above figures of the dendrogram and the table show the results of the clustering. 4 clusters are the result of using 1 level. From the resultant data, it can be interpreted that cluster 2 is the "Most Successful" group of learners. Cluster 0 is the "Least Successful" group. But, it can be seen from the dendrogram that clusters 0 and 1 are in closer proximity, but their average Monthly Earning and Earning Ratio are not similar at all. Hence, we can deduce that the results are not accurate.

Chapter 5

Conclusion and Future Work

The main purpose of this study is to find out whether background data of applicants can be used to predict their future success. Being able to extrapolate such data would allow apprenticeship based programs to look into which groups of people are generally disadvantaged when it comes to being able to perform better in their careers, while also being able to fine tune their programs to support these people. Being a developing nation, Bangladesh has a big need for a skilled labour force and the knowledge to tackle such cases can be a very important boon for the socioeconomic conditions of the learners, their families and also the nation. As found in the research, there seems to be a lack of any such quantitative feature which can be used to extrapolate such information. All features used had very low correlation to the success of the learners. Metrics used to judge all types of models have shown this repetitively.

This research would serve as a big milestone for both BRAC Skills Development Program (SDP) as well as the informal sector, as previous work done on the informal sector on this scale is sparse. This research faced quite a few hurdles due to a lack of time, and also problems brought about by the COVID-19 pandemic. It was not possible to gather more primary data from the field and attempt more of an in-depth research. Being able to gather psychological information about these learners could enable future researchers to gauge the learners' mindset, which can offer invaluable information pertaining to their success in the future. Questionnaires taking information about the learners' experience after the program could also provide information about their experience, which can also serve as a feature. For future work, being able to gather more data across many years, and use more modern models, including deep learning approaches, would ensure data robustness. Furthermore, cognitive ability and aptitude test data would also significantly improve the accuracy of the research. Currently, BRAC Skills Development Programme (SDP) does not use or collect these features. But it is hoped that the sharing of this research with BRAC would encourage them to collect and mine these data, to extract valuable findings.

Bibliography

[1] H. Plecher. (Nov. 2020). "Unemployment rate in bangladesh 2020," [Online]. Available: https://www.statista.com/statistics/808225/unemployment-rate-in-bangladesh/ (visited on 2021).

[2] T. Economies. (2019). "Bangladesh unemployment rate1991-2019 data — 2020-2021 forecast — historical," [Online]. Available: https://tradingeconomics.com/bangladesh/unemployment-rate (visited on 2021).

[3] T. W. Bank. (2020). "Population, total - bangladesh," [Online]. Available: https://data.worldbank.org/indicator/SP.POP.TOTL?locations=BD (visited on 2021).

[4] B. SDP, *Brac skills development programme*. [Online]. Available: http://www.brac.net/program/skills-development/#howwedoit.

[5] SDP. (Feb. 2019). "Data talk 2," [Online]. Available: https://docs.google.com/document/d/13JnUt3xGIDxCTRfCUjdb4p_9SALxfIGDDwQPRYLcr24/edit (visited on 2021).

[6] S. R. SHAKIL and S. SYED. (Jul. 2019). "How brac used machine learning to reduce the drop-out rate of learners with disabilities to zero," [Online]. Available: http://blog.brac.net/how-brac-used-machine-learning-to-reduce-the-drop-out-rate-of-learners-with-disabilities-to-zero/ (visited on 2021).

[7] C. J. Mills, "Characteristics of effective teachers of gifted students: Teacher background and personality styles of students," *Gifted Child Quarterly*, vol. 47, no. 4, pp. 272–281, 2003.

[8] E. Hanushek, "Teacher characteristics and gains in student achievement: Estimation using micro data," *The American Economic Review*, vol. 61, no. 2, pp. 280–288, 1971.

[9] P. L. Crawford and H. Bradshaw, "Perception of characteristics of effective university teachers: A scaling analysis," *Educational and Psychological Measurement*, vol. 28, no. 4, pp. 1079–1085, 1968.

[10] R. A. Schibeci and J. P. Riley, "Influence of students' background and perceptions on science attitudes and achievement," *Journal of Research in Science teaching*, vol. 23, no. 3, pp. 177–187, 1986.

[11] J. S. Coleman, J. Campbell, E. Campbell, C. Hobson, J. McPartland, A. Mood, F. Weinfeld, and R. York, "Equality of educational opportunity (washington, dc: Us department of health, education, and welfare, us government printing office)," 1966.

[12] S. Konstantopoulos and G. D. Borman, "Family background and school effects on student achievement: A multilevel analysis of the coleman data," *Teachers College Record*, vol. 113, no. 1, pp. 97–132, 2011.

[13] P. Kaur, M. Singh, and G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector," *Procedia Computer Science*, vol. 57, pp. 500–508, 2015.

[14] *I. J. of Advances in Scientific Research and Engineering-IJASRE (ISSN: 2454 - 8006)*, vol. 5, no. 10, pp. 237–250, Oct. 2019. DOI: 10.31695/IJASRE.2019.33565. [Online]. Available: https://ijasre.net/index.php/ijasre/article/view/652.

[15] Dansbecker, "Using categorical data with one hot encoding," 2018. [Online]. Available: https://www.kaggle.com/dansbecker/using-categorical-data-with-one-hot-encoding.

[16] A. Kumar, "What is helmert encoding ?," 2020. [Online]. Available: https://www.kaggle.com/getting-started/185167.

[17] S. (https://stats.stackexchange.com/users/7962/statsstudent), *How to calculate helmert coding*, Cross Validated, URL:https://stats.stackexchange.com/q/411837 (version: 2020-10-15). eprint: https://stats.stackexchange.com/q/411837. [Online]. Available: https://stats.stackexchange.com/q/411837.

[18] W. Badr, [Online]. Available: https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4.

[19] R. Cuninghame-Green, "Minimax algebra and applications," in, ser. Advances in Imaging and Electron Physics, P. W. Hawkes, Ed., vol. 90, Elsevier, 1994, pp. 1–121. DOI: https://doi.org/10.1016/S1076-5670(08)70083-1. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1076567008700831.

[20] J. Brownlee. (Mar. 2016). "Linear regression for machine learning," [Online]. Available: https://machinelearningmastery.com/linear-regression-for-machine-learning/ (visited on 2021).

[21] S. Naithani. (Mar. 2020). "What is linear regression algorithm? introduction implementation," [Online]. Available: https://in.springboard.com/blog/what-is-linear-regression/ (visited on 2021).

[22] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[23] Qshick. (Jan. 2019). "Ridge regression for better usage," [Online]. Available: https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db (visited on 2021).

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] N. Donges. (Jun. 2019). "A complete guide to the random forest algorithm," [Online]. Available: https://builtin.com/data-science/random-forest-algorithm (visited on 2021).

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] N. Suguna and K. Thanushkodi, "An improved k-nearest neighbor classification using genetic algorithm," *International Journal of Computer Science Issues*, vol. 7, no. 2, pp. 18–21, 2010.

[28] Tutorialspoint. (2020). "Knn algorithm - finding nearest neighbors," [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm (visited on 2021).

[29] P. Majumder. (2020). "Gaussian naive bayes," [Online]. Available: https://iq.opengenus.org/gaussian-naive-bayes/ (visited on 2021).

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] N. Singh. (2020). "Decision tree algorithm, explained," [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html.

[32] L. Chen, *Support vector machine-simply explained*, Jan. 2019. [Online]. Available: https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496.

[33] R. Keim, *How many hidden layers and hidden nodes does a neural network need? - technical articles*, Jan. 2020. [Online]. Available: https://www.allaboutcircuits.com/technical-articles/how-many-hidden-layers-and-hidden-nodes-does-a-neural-network-need/.

[34] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003, Biometrics, ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(02)00060-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320302000602.

[35] I. Dabbura, "K-means clustering - algorithm, applications, evaluation methods, and drawbacks," 2018. [Online]. Available: https://imaddabbura.github.io/post/kmeans-clustering/.

[36] "The clustering validity with silhouette and sum of squared errors," *3rd International Conference on Industrial Application Engineering 2015*, 2015.

[37] S. Zhou, Z. Xu, and F. Liu, "Method for determining the optimal number of clusters based on agglomerative hierarchical clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 12, pp. 3007–3017, 2017. DOI: 10.1109/TNNLS.2016.2608001.

[38] C. R. Patlolla, *Understanding the concept of hierarchical clustering technique*, 2019. [Online]. Available: https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec.

[39] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: Essential theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12–30, 2015. DOI: 10.1109/MSP.2015.2398954.

[40] user3658307 (https://math.stackexchange.com/users/346641/user3658307), *What's the algorithm for agglomerative hierarchical clustering?* Mathematics Stack Exchange, URL:https://math.stackexchange.com/q/2847132 (version: 2018-07-10). eprint: https://math.stackexchange.com/q/2847132. [Online]. Available: https://math.stackexchange.com/q/2847132.

[41] S. Podani János and Dénes, "On dendrogram-based measures of functional diversity," *Oikos*, vol. 115, no. 1, pp. 179–185, DOI: https://doi.org/10.1111/j.2006.0030-1299.15048.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2006.0030-1299.15048.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2006.0030-1299.15048.x.