FUTURE JOB PREDICTION IN TRIVANDRUM USING MACHINE LEARNING TECHNIQUES

Dissertation submitted in part fulfilment of the requirements for the degree of MSc in Information Systems with Computing at Dublin Business School

Akku George Saju
Student Id: 10376281

MSc Information Systems with Computing August, 2018
DECLARATION

I, Akku George Saju, declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this work and this work is fully compliant with the Dublin Business School’s academic honesty policy.

Signed: Akku George Saju

Date: 20/08/2018
ACKNOWLEDGEMENTS

First and foremost, I would like to thank God Almighty for giving me the knowledge and opportunity to do this dissertation. I would like to thank Abhishek Kaushik, my supervisor who provided me valuable motivation, feedback and direction throughout the dissertation process. He has helped me to think deeply about the research problems and provided me freedom to shape this thesis.

I am also very much thankful to all the lecturers and academic staffs of Dublin Business School, who provided me support and confidence to move on in this journey of thesis.

I take this opportunity to sincerely thank the companies located in Trivandrum, Kerala who provided me required data for my thesis process.

And most importantly I would like to thank my parents for believing in me and supporting me in all the difficult moments. Also, I am very much thankful to my brother and sister who provided me incredible strength and encouragement from beginning to the end of this journey.

I would also like to thank all my friends who stood beside me during the hard times of this study.
This dissertation explores the impact of machine learning on job aspirants through an interactive web application. Job seeking is a very difficult task in India and determining a career path is a much more difficult challenge. There is a lack of planning in the life of job aspirants. Job aspirants don’t get enough updated information regarding the job opportunities in different sectors. There is a need of a proper medium through which job aspirants could get information on the number of job vacancies, number of intakes, salary etc. in each job industry for the future years. Using such predictions, job aspirants can plan on a specific career path and the specific qualification required for the job can be attained in the following years. Nowadays, websites act as a powerful medium to reach to the job aspirants. The major contribution of this dissertation will be a web application using angular platform, through which job aspirants get updated knowledge regarding future job opportunities in the form of statistics, charts and graphs. Therefore, this dissertation endeavours to minimize the challenges faced by job aspirants through machine learning techniques. The study was conducted by adopting a mixed method approach: both statistical and predictive in nature. Decision tree, linear regression, decision forest and neural network regression are some of the machine learning algorithms considered in this study. The key findings using these algorithms are also displayed in this study.
# TABLE OF CONTENTS

## CHAPTER 1: INTRODUCTION

- 1.1 BACKGROUND ........................................................................... 9
- 1.2 AIM AND OBJECTIVE ................................................................. 10
- 1.3 ORGANISATION OF DISSERTATION ........................................ 10
- 1.4 DISSERTATION SCOPE AND LIMITATIONS .............................. 11
- 1.5 DISSERTATION CONTRIBUTIONS ............................................. 11

## CHAPTER 2: LITERATURE REVIEW

- 2.1 INTRODUCTION ....................................................................... 13
- 2.2 MACHINE LEARNING ............................................................... 13
- 2.3 DECISION TREE ALGORITHM .................................................. 13
- 2.4 LINEAR REGRESSION ALGORITHM ......................................... 15
- 2.6 WEB APPLICATION .................................................................. 17
- 2.7 CONCLUSION ........................................................................... 18

## CHAPTER 3: RESEARCH METHODOLOGY AND METHODS

- 3.1 INTRODUCTION ....................................................................... 20
- 3.2 RESEARCH PHILOSOPHY .......................................................... 20
- 3.3 RESEARCH APPROACH ............................................................. 21
- 3.4 RESEARCH STRATEGY ............................................................... 21
- 3.5 STRENGTHS AND WEAKNESSES .......................................... 24
- 3.6 SAMPLE .................................................................................... 26
- 3.7 ETHICAL PROCEDURES ............................................................ 27
- 3.8 CONCLUSION ........................................................................... 27

## CHAPTER 4: ARTEFACT DESIGN AND DEVELOPMENT

- 4.1 RAPIDMINER ........................................................................... 29
- 4.2 TABLEAU ................................................................................... 30
CHAPTER 5: FINDINGS ................................................................................................................. 41
5.1 INTRODUCTION ..................................................................................................................... 41
5.2 DATA FLOW DIAGRAM OF TEST RESULTS ........................................................................ 41
5.3 LINEAR REGRESSION TEST RESULTS .................................................................................. 44
5.4 DECISION FOREST TEST RESULTS ..................................................................................... 50
5.5 NEURAL NETWORK REGRESSION TEST RESULTS ............................................................... 53
5.6 DECISION TREE TEST RESULTS ......................................................................................... 58
5.7 COMPARISON TEST RESULTS ............................................................................................. 63

CHAPTER 6: DISCUSSION .............................................................................................................. 66
6.1 INTRODUCTION ..................................................................................................................... 66
6.2 ADDRESSING RESEARCH QUESTIONS .................................................................................. 66
6.3 DEMONSTRATING DISSERTATION LIMITATIONS ................................................................. 67

CHAPTER 7: CONCLUSIONS ........................................................................................................ 69

BIBLIOGRAPHY ............................................................................................................................ 71

APPENDICES .................................................................................................................................. 77
# LIST OF FIGURES AND TABLES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Research &quot;Onion&quot; (Mohamad, 2015)</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>Methodology Pattern (Abhishek Kaushik, 2016)</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>Steps in Visualization (Abhishek Kaushik, 2016)</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>Use case diagram of the process</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>RapidMiner model using decision tree algorithm</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>RapidMiner model using SVM operator</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>Tableau workbook</td>
<td>31</td>
</tr>
<tr>
<td>8</td>
<td>Azure studio model using decision forest regression operator</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>Azure studio web service experiment for linear regression algorithm</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>Azure API management services window</td>
<td>33</td>
</tr>
<tr>
<td>11</td>
<td>Gateway data flow diagram</td>
<td>33</td>
</tr>
<tr>
<td>12</td>
<td>Homepage of the application opened in Visual studio code</td>
<td>34</td>
</tr>
<tr>
<td>13</td>
<td>Increasing demand for angular technology (Chauhan, 2015)</td>
<td>34</td>
</tr>
<tr>
<td>14</td>
<td>Login data flow diagram</td>
<td>35</td>
</tr>
<tr>
<td>15</td>
<td>Login Page of the web application</td>
<td>36</td>
</tr>
<tr>
<td>16</td>
<td>Home Page of the web application</td>
<td>36</td>
</tr>
<tr>
<td>17</td>
<td>Second Page of the Web application</td>
<td>37</td>
</tr>
<tr>
<td>18</td>
<td>Predicted results based on the number of vacancies using decision tree algorithm</td>
<td>37</td>
</tr>
<tr>
<td>19</td>
<td>Third page of the web application</td>
<td>38</td>
</tr>
<tr>
<td>20</td>
<td>Web application in android mobile</td>
<td>39</td>
</tr>
<tr>
<td>21</td>
<td>Predicted data visualization on the prediction input form</td>
<td>40</td>
</tr>
<tr>
<td>22</td>
<td>Google map API</td>
<td>40</td>
</tr>
<tr>
<td>23</td>
<td>Azure data flow diagram containing tests</td>
<td>41</td>
</tr>
<tr>
<td>24</td>
<td>Azure data flow diagram for algorithms</td>
<td>42</td>
</tr>
<tr>
<td>25</td>
<td>Split Test technique</td>
<td>43</td>
</tr>
<tr>
<td>26</td>
<td>Cross validation Test technique</td>
<td>43</td>
</tr>
<tr>
<td>27</td>
<td>Chi-square test results</td>
<td>43</td>
</tr>
<tr>
<td>28</td>
<td>Linear regression on Number of Intakes</td>
<td>44</td>
</tr>
<tr>
<td>29</td>
<td>Linear regression on Number of vacancies</td>
<td>44</td>
</tr>
<tr>
<td>30</td>
<td>Scored label from split technique Vs Scored label (2) from cross validation technique</td>
<td>44</td>
</tr>
<tr>
<td>31</td>
<td>Error Histogram of linear regression on Number of vacancies using Cross validation technique (Testing score)</td>
<td>45</td>
</tr>
<tr>
<td>32</td>
<td>Error Histogram of linear regression on number of vacancies using split technique (Training score)</td>
<td>45</td>
</tr>
<tr>
<td>33</td>
<td>Linear regression on Salary tableau in application giving sector wise statistical analysis for a user</td>
<td>46</td>
</tr>
<tr>
<td>34</td>
<td>Application Prediction based on Salary in Linear regression</td>
<td>46</td>
</tr>
<tr>
<td>35</td>
<td>Scored Label visualization on linear regression using Cross validation technique</td>
<td>47</td>
</tr>
<tr>
<td>36</td>
<td>Scored Label visualization on linear regression using split technique</td>
<td>47</td>
</tr>
<tr>
<td>37</td>
<td>Scatter plot on Tableau of linear regression on vacancies</td>
<td>48</td>
</tr>
<tr>
<td>38</td>
<td>Scatter plot 3D on RapidMiner for linear regression on intakes</td>
<td>48</td>
</tr>
<tr>
<td>39</td>
<td>Linear regression on salary Tableau</td>
<td>49</td>
</tr>
<tr>
<td>40</td>
<td>Linear regression on vacancy on Azure</td>
<td>49</td>
</tr>
<tr>
<td>41</td>
<td>Decision Forest on Salary</td>
<td>50</td>
</tr>
<tr>
<td>42</td>
<td>Decision forest on salary in Azure</td>
<td>50</td>
</tr>
<tr>
<td>43</td>
<td>Application prediction</td>
<td>50</td>
</tr>
<tr>
<td>44</td>
<td>Scored label visualization on Decision Forest using cross validation technique</td>
<td>51</td>
</tr>
<tr>
<td>45</td>
<td>Scored Label visualization on Decision forest using split technique</td>
<td>51</td>
</tr>
</tbody>
</table>
Figure 46 Error Statistics of Decision Forest on Number of vacancies using split technique (Training score)
Figure 47 Error Statistics of decision forest on Number of vacancies using Cross validation technique (Testing score)
Figure 48 application gives tableau result on number of intakes giving sector wise statistical analysis for a user
Figure 49 Scatter plot 3D in RapidMiner
Figure 50 Application prediction
Figure 51 Neural network regression model
Figure 52 On year in Azure
Figure 53 On vacancies in Tableau
Figure 54 Neural Networks on Years of experience required
Figure 55 Error Histogram of neural network on years of experience using split technique (Training score)
Figure 56 Error Histogram of neural network on years of experience using Cross validation technique (Testing score)
Figure 57 Scored Label visualization on neural network using Cross validation technique
Figure 58 Scored label visualization on neural network using split technique
Figure 59 Tableau showing predicted results on salary giving sector wise statistical analysis for a user
Figure 60 Application Prediction
Figure 61 On years in Tableau
Figure 62 On number of intakes in Azure
Figure 63 For Number of intakes
Figure 64 Error Histogram of decision tree on Number of vacancies using Cross validation technique (Testing score)
Figure 65 Error Histogram of decision tree on Number of vacancies using split technique (Training score)
Figure 66 Scatter plot on Tableau
Figure 67 Decision tree model in RapidMiner
Figure 68 Scored label visualization on decision tree using split technique(left) and Cross validation technique(right)
Figure 69 Scatter plot 3D on RapidMiner
Figure 70 Map of Trivandrum
Figure 71 prediction graphs showing less variation

Table 1 Test results summary
Table 2 Multilayer perceptron for other labels
Table 3 MLP for different parameters
Table 4 accuracy measures for sector wise prediction in azure
CHAPTER 1: INTRODUCTION

1.1 BACKGROUND
This dissertation proposes to predict job prospects in Trivandrum city located in Kerala, India using machine learning techniques. A job is an activity which people do often regularly and, in most cases, remunerated. India, the second largest country in terms of population, has different culture and is very diverse. Indian economy is the third largest by purchasing power parity and tenth largest around the world by nominal GDP (Johnson, 2017). But today people who are actively searching for a job are unable to get one, since country has been facing the challenges of unemployment for a long time. In India unemployment has become a chronic issue. To absorb the growing population, country is lacking ability to create adequate job opportunities (Johnson, 2017). Kerala, the southern state in India known as ‘God’s own country’ has many graduates entering the job search pool each year. Many Keralites work for leading multi-national companies around the world. Despite these facts, a report on ‘The Hindu’ highlights that Kerala tops in unemployment rates (THE HINDU, 2016). Among the big states in India, Kerala has the highest unemployment rate at 7.4 per cent, which is three times the national level. The future of youth is adversely affected by the unemployment crisis. In the 15-29 age group, the unemployment rate is 21.7 percent in rural areas while it is 18 percent in urban areas (THE HINDU, 2016). The unemployment rates of male and female in Kerala in both rural and urban areas were found to be normally three to four times higher than the all India average (Times Of India, 2018). The major causes of high unemployment rates are the following (Rajeev, 2018):

1. Rapid population growth: Higher rates of population growth is a crucial factor that contributes to higher levels of unemployment.
2. Economic growth: Population explosion together with a dip in the economic growth rates resulted in increased rates of unemployment
3. Agricultural growth: Nowadays, agriculture is considered as an unprofitable sector, the agricultural lands are getting converted into residential plots throughout the country.
4. Industrial growth: The state lacks a healthy industrial growth.
5. Flight of capital: Lower credit deposit ratio of nationalized banks in Kerala is also a factor.
6. Utilization of remittances: Massive flow of remittances were not utilised properly to solve the problem of unemployment.

Finding a job is a very difficult task in India and finding a career path is a much more difficult challenge. There is a lack of planning in the life of job aspirants. Job aspirants don’t get enough updated information
regarding the opportunities in various job industries. There is a need of a proper medium through which people could estimate number of job vacancies, number of intakes, salary etc. for the future years. Nowadays as the number of internet users are on rise in the country, people could get updated knowledge in the form of statistics, charts and graphs. Therefore, this dissertation will seek to resolve most of the above challenges using predictive analysis.

1.2 AIM AND OBJECTIVE
The primary aim of this dissertation is to create a medium for job aspirants to predict job prospect features like job vacancies, intakes, salary available in different sectors like management, administration, finance, IT for the future years. Thereby, users also could grasp what educational qualifications are required to meet a specific range of salary.

Machine Learning is the science of getting computers to learn and behave just like how humans do, moreover enhance their learning in an autonomous fashion over time, by giving them required information in the form of real world interactions and observations (Faggella, 2017). The main objectives of this research are the following:

1. To analyse and predict job prospect features in Trivandrum, the capital city of Kerala, using machine learning techniques.
2. To develop a web application that shows job features for future with visualization effects.

Lack of communication and analytical skills, demand for highly skilled labours, unsatisfied incomes or salaries, inadequate technical skills for the prospective job role etc. are some of the individual drawbacks that leads to unemployment. Most of these drawbacks can be resolved by introducing a proper medium like web applications, which helps to recognize the future job vacancies available in a sector. After unravelling the job opportunities about to come in future, users could decide what qualifications are vital to achieve jobs and high salaries in future. The primary research questions that this research work would focus is addressed as follows:

1. How accurately the machine learning techniques of analysing and predicting job prospect features of a company in Trivandrum city can be implemented and linked to a web application?
2. Examine the prediction accuracy based on different sectors on the end to end user demands?

1.3 ORGANISATION OF DISSERTATION
Chapter one discusses the background of the dissertation topic, defining the problem and its causes. This section also details about the research question, aim and objectives, scope and limitations of the research. Chapter two deals with literature review, incorporating the significance for the inclusion of each topic. Chapter three contains research methodology and methods featuring the strengths and weaknesses of the
dissertation. Chapter four discusses artefact design and development, showing how the artefact came about and contributes to the dissertation. Chapter five illustrates the findings in a way that general conclusions are not disclosed. Multiple sections are included with respective aims and objectives. Chapter six examines the discussion part, where the work is reviewed. Also, this section will carry out interpretation of the results, answering the research questions and discussing the inferences of the findings. Chapter seven draws general conclusions by summarising the findings in a reflective and coherent manner including recommendations for the future work. Chapters eight and nine list bibliography and appendices respectively.

1.4 DISSERTATION SCOPE AND LIMITATIONS
The scope of this dissertation is to predict job prospect features in different sectors like administration, management, IT, teaching, servicing and finance in the core city of Kerala, Trivandrum. In this dissertation, a web application is created using angular technology. Angular technology brings consistency in the codes, enhance productivity, maintains modularity and catches errors earlier. The web application will predict job features like number of vacancies, number of intakes, salary, experience for the future years that the user request. Machine learning models are designed using trending tools like RapidMiner and Microsoft azure studio. Azure machine learning studio is very user-friendly and provides a set of required tools without any limitations. In this research work, azure gateways are enabled to bring the API’s created in the azure portal into application. As a result, application will deliver options for the user to input required features for prediction and hence becomes interactive with user. Options are provided by azure for publishing the data models. Google map API is used in the application which helps to bring location into the web on giving latitudes and longitudes of Trivandrum. Tableau and charts.js are the visualization packages used in this dissertation to view the results. A simple and an interactive approach is attained in the visualization section. It will help user to increase the understanding level of predicted statistics. I hope that this research will help to provide necessary job information to job aspirants and minimize the challenges faced by users today.

One major limitation is that the application will predict job features only for one city. Data is collected only from Trivandrum city and therefore prediction is done only for the Trivandrum city. Another limitation faced by this research work is the quantity of data obtained. Due to the data privacy reasons, companies in Trivandrum provided only less amount of data.

1.5 DISSERTATION CONTRIBUTIONS
The major contribution of the dissertation will be to create a web application with angular technology as the backbone which predicts required job prospect features for the future years that the user inputs. At
present, there are only very few applications using angular platform for prediction purposes. In angular technology, there is a need to install different packages each time to use foreign components. As well as, some additional effort is required to convert the html codes into a format accepted by the angular. That is why the web application will be a core component of this dissertation. This research work employs dominant visualization tools like Tableau and charts.js, which will provide a better understanding to the user who need future information regarding job attributes. This dissertation will critically assess the impact of an interactive and simple medium to reach to the people. This thesis will be a starting point for the present and new researchers, who are keen to discover more on angular technology and machine learning domain. The web application will help users to recognize the number of vacancies, number of intakes, salaries available in different sectors like administration, management, IT, servicing etc. for the future years. A simple and user-friendly interface provided by the web application aid user to analyse and predict the job features on their own. Some of the current challenges like lack of proper information on the job prospect features and career planning can be reduced using the application. This research work will seek to signify the importance of machine learning domain, angular platforms and visualization tools. This dissertation will answer the question: How accurately the machine learning techniques can be implemented for analysing, predicting future jobs in Trivandrum and linked to a web application with sector wise classification?
CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION
It is necessary to review important literatures that underpins the research topic to justify the directions of the study and better locate my research among other vivid studies. This chapter will brief significant literatures in the machine learning domain. The aim of this chapter is to describe some of the main concepts like machine learning, neural network regression, Linear regression, support vector machine, decision tree, etc. Literature review plays a crucial role to bring clarity in the topic of study and helps to broaden knowledge in the research area. An awareness of what research has been formerly completed in machine learning domain and how it was done is acquired through this chapter. The following section provides a review of key literatures highlighting the relevance of each topic in research.

2.2 MACHINE LEARNING
Machine Learning is the technology in which systems can learn directly from data, experience and examples. This technology allows a machine to do specific tasks, complex processes wisely via learning from data and examples (The Royal Society, 2017). Systems learn from previous data and previous experience to improve the future performance. The term machine learning was coined out by Arthur Samuel. Machine learning is a field of computer science in which the machine is trained, learned automatically rather than being explicitly programmed (Cheolsoo Park, 2018). Nowadays, machine learning techniques have enhanced the ability of a system to achieve a higher performance than humans. For instance, a recent progress in image recognition have made the systems more precise compared to earlier times. There has been a great advancement in the field of machine learning over the recent years. Systems which struggled previously to attain accurate results can now outperform humans at specific tasks (The Royal Society, 2017). This dissertation is based on predictive analytics where future job features are predicted. The ability of machine learning technology to perform decisions or predictions by processing huge amounts of data makes machine learning a key tool in this thesis.

2.3 DECISION TREE ALGORITHM
In decision tree algorithm, instances are classified by sorting them based on feature values. A decision tree has many nodes and branches where node in a decision tree symbolizes a feature of the instance to be classified. A branch in a decision tree depicts a value that the node can assume. Decision tree algorithm allows automation of different tasks currently done by humans (F.Y.Osisanwo, 2017). This algorithm is given the name decision tree as it begins with a root, which then leads to several branches until when a prediction or decision is made for a problem. Based on certain constraints a tree like structure will be constructed involving solutions to a problem (Kajaree Das, 2017). One main reason to select decision tree algorithm in this dissertation is due to the efficiency of this algorithm. To capture non-linearities and other
effects efficiently, there is a need of highly interpretable models which can be built using decision tree algorithm. In a study performed by Mahjoobi and Etemad-Shahidi, prediction is done by binary decision trees in which the input characteristics will be split from root node on to the leaves (bottom nodes). There will be exactly two outgoing edges for each internal node of the decision tree structure: left child and right child. Some characteristic tests of an instance will be done by every node in the decision tree and every branch descending from node leads to feasible outputs. Terminal nodes are formed when no more splitting of nodes take place, which is the final stage of decision tree construction process (Lei Ren, 2016). A similar pattern of decision tree construction takes place in this dissertation as the algorithm parameters of decision tree remains the same. Outputs obtained from the designed models are presented in the findings section of this dissertation.

In this research, decision tree algorithm is used for classifying the data and to predict specific job features. In decision tree, every node except the root node has an incoming link. Every node is representing a predictive nature and terminal nodes characterize predicted values of the predictor variable. On tracking path from root of the tree to a leaf, it is possible to generate a decision or prediction about the condition of the predictor variable after obtaining a new sample. The algorithm must make a confirmation regarding the splitting variables, split points and shape of the tree. Selection of random variables out of all parameters of a dataset is the first stage in developing the decision tree. $\sqrt{N}$ is the equation used to choose number of random variables at each step, where N is equal to the total number of variables of a dataset (Fahim Ahmed, 2017).

All the data collected is split into training and testing sets, followed by giving the training dataset as input to a decision algorithm. In this research there should not be a factor that limits the relationship between label attribute and other key attributes. If the label attribute is number of vacancies, then there are many other key attributes like salary, year of experience, number of intakes etc. that relates with the label attribute. Number of vacancies is the total number of vacancies in the company and number of intakes is the number of vacancies filled by recruitment procedure. One main feature of decision tree algorithm is that partitioning of training dataset does not limit relationship between label attribute and input variables (Michele Fratello, 2017). Ease of use, easy to understand, no hidden assumptions, simple way of presentation, uncomplicated guidance for a user to arrive at a conclusion are some of the benefits of using decision tree in this research work (Paula A.Harrison, 2018). Decision forest algorithm is also used in this dissertation as a predictive model. In decision forest, a group of decision trees are viewed as a whole. Capabilities of decision tree algorithm are inherited in decision forest also; a good thing is that many of the shortcomings of decision tree are resolved in decision forest (Md Nasim Adnan, 2017).
2.4 LINEAR REGRESSION ALGORITHM

The main objective of this algorithm is to examine the relationship between variables. Type I regression is the most common form of regression technique in which an attempt is done to figure out the relationship between independent and dependent variables. In Type II regression there are no independent variables, that means all variables are capable of influencing each other (Inna Boldina, 2015). Here in this research work relationship between job attributes is determined using Type I regression. Both curvilinear and straight-line relationships can be modelled using linear regression algorithm.

In simple linear regression model there are two variables, one is the input variable and the other is response variable. Response variable Y for a specified input value X is given by

\[ Y = \alpha + \beta X + e \]

where \( \alpha \) and \( \beta \) are parameters. ‘e’ is a variable called the random error. A simple linear regression is the relationship between the input variable X and the response variable Y (M.Ross, 2017). Linear regression is represented by a general mathematical equation

\[ Y = a + b X \]

where ‘a’ is the y-intercept and ‘b’ is the slope of the line. There are three consecutive steps in performing a linear regression. First step is to properly specify the regression model and second step is determination of regression parameters. Third step is validation of the model (Inna Boldina, 2015). Scatter plot is a graph on the cartesian plane in which every point is having certain coordinates in the form (Xt, Yt). The collection of points will be approximately a straight line if there exists a linear relationship between Xt and Yt, otherwise these points will take other shapes (Massimo Guidolin, 2018). Scatter plot obtained in this dissertation is presented in the findings section. A study conducted by Tomorn Sunthornnapha in September 2017 compared multilayer perceptron method (MLP) and linear regression method (LR) for creating a reliable energy baseline model. The nature of plant variables in that study were having high nonlinearity and increased complexity. Linear regression method was more sensitive to the outliers. MLP method was introduced to reduce the effect of nonlinearity by supervised learning. But LR method produced better results of the uncertainty parameters for short-term period (Sunthornnapha, 2017). This feature of linear regression method is very necessary in predictive analytics. The assumptions which must be considered while using simple linear regression models are the following:

i. Relationship between independent and dependent variables is linear.

ii. A normal probability distribution must be maintained by the residuals, following a normal distribution in the difference between actual value and estimated value.
iii. Residuals or errors change over time without repeating negative or positive value containing patterns.

iv. Variance in the neighbourhood of regression line is the same for all independent variable values (Idriss Elsiddig, 2015).

In a study conducted by Grzegorz Dudek, ARIMA was taken into consideration for the comparison of designed models with other models. Autoregressive Integrated Moving Average (ARIMA) is one of the conventional short-term load forecasting methods which is commonly used in many prediction tools like RapidMiner. ARIMA model’s linear nature is a disadvantage. Order selection process is normally assumed as subjective and is difficult to apply (Dudek, 2015). This is one of the main obstacles in using these models. In this dissertation, simple linear regression models are used to find the relationship between variables. On realising the demerits of using ARIMA, it was decided to use Microsoft azure studio for prediction purposes and comparison of the models. Figuring out independent variable that builds influence on dependent variable, proper estimation of dependent variables according to the alteration in independent variables are some of the common benefits of regression analysis that play a significant role in predicting the job features of Trivandrum city in this thesis (Jeon, 2015).

2.5 NEURAL NETWORK REGRESSION ALGORITHM

Multilayer perceptron (MLP) neural network is the most widely used network in several applications. MLP is an artificial neural network that has high ability to learn (Urszula Markowska-Kaczmar, 2015). Artificial neural networks (ANN) also known as neural networks play a great role in solving complex non-linear problems. This algorithm is closely related to biological neuron system. ANN, a computational model comprises huge collection of bridged artificial neurons. To represent relations between inputs and outputs, it is possible to train neurons and their connections with data. One key feature of ANN is its advantages in predicting output or result in the absence of real information knowledge, regarding the modelled system. Computer science, finance, engineering and many other fields depend on ANN to solve vivid crucial problems. The main reason for deploying ANN in real-time applications is its computational efficiency. This algorithm has the genuine potential to handle very high nonlinear complex problems and provide good accuracy rates (Bin Wu, 2018). Computational efficiency and ability to handle high nonlinear complex problems are the two factors that helped to get satisfactory results in this dissertation.

When it comes to architecture of neural networks, input layer is the leftmost layer in the network, input neurons are the neurons within the layer, output neurons are present in the output or rightmost layer. The neurons in the middle layer are neither inputs nor outputs. This middle layer is called hidden layer (Nielsen, 2015). Dr Robert Hecht-Nielsen, inventor of one of the first neurocomputers defined a neural network as “a computing system made up of a number of simple, highly interconnected processing
elements, which process information by their dynamic state response to external inputs” (Fumo, 2017). Information first reaches the Input layer, where no sort of computation takes place then it is passed to the next layer(hidden layer). Intermediate processing takes place in hidden layer, in which computations are performed and transfer the information to either output layer or next hidden layer if present. In output layer, an activation function is used for mapping to the required output format (Fumo, 2017).

In a study managed by Malgorzata Pawul and Malgorzata Sliwka, artificial neural network was used to predict particulate matter PM 10 concentration levels. Particulate matter PM 10 is considered as the main cause for occurrence of sulfuric smog. Input contained meteorological data like average temperature, average wind speed etc. and the goal was to get predicted concentrations of PM 10. They split the data into three subsets containing 75 percent of training data, 15 percent of validation set and 15 percent of testing data. Selection of dataset is very crucial in building neural network model. Universal approximation property, robustness, massive parallelism and learning capabilities are some of the merits of using neural network in this thesis work (Dudek, 2015). A common pattern followed by neural network model is that it compares input and output data, so that input data that affects the output data must be used (Malgorzata Pawul, 2016). In this study, job prospect features like number of vacancies, number of intakes, years of experience, salary etc. are given the role as a label in corresponding stages. Even if the experimental functions have high non-linear negative issues, there is presence of adaptive capabilities in ANN to map these functions. On increasing the number of processing nodes, adaptive capabilities of ANN also increase. One standard method of ANN to develop a model is network training and another one is network testing, to verify the performance (Fontanella, 2018). Efficiency of algorithms, computing power, quality of data are the factors that affects ANN model performance (Weipeng Cao, 2017).

2.6 WEB APPLICATION

Both developed and developing countries are now facing unemployment, one of the serious social issues. One of the main reasons for this crisis is lack of information regarding the job opportunities in future. People are not able to get updates and data based on future employment opportunities. In the past, it was difficult for the job seekers to get enough information about the job opportunities via traditional medias. Some of the traditional methods users relied on are the following:

i. Private and public recruitment agencies
ii. Job fairs
iii. Newspaper advertisements
iv. Radio and television advertisements
v. Present employee contacts
These traditional job finding methods were very slow, expensive, challenging and stressful. Today, internet has changed the living standards of people (Yasin, 2014). There has been a massive rise in the number of internet users in India soon after the declaration of central governments ambitious digital India program. A report on ‘India Today’ states that government has introduced 965 megahertz spectrums to different telecom service providers in India as a part of pushing the digitisation program (Bhardwaj, 2017). Millions and millions of people have started to use web applications as part of their day to day routine. Job seekers prefer online methods in their job hunt as the most convenient and time saving approach (Yasin, 2014).

It has been observed that, at the end of December 2017 India had 481 million internet users, rocketing by 11.34 % compared to 2016. A recent report by internet and mobile association of India (IAMAI) revealed that number of internet users will reach about 500 million by June 2018 (Ayyar, 2018). The most literate state in India, Kerala, has declared internet access as a basic right for all the people just like water, food and education (Varma, 2017). Using surveying and interviewing as the primary methods, Parry and Tyson organised a study over a period of six years on the recruitment processes of corporations. Human resource directors and managers, managing directors and recruitment specialists, finance directors from UK were the respondents of this study. Respondents were asked for reason behind the acceptance or rejection of online methods, obtaining about 25,524 responses in the survey. The most common reasons obtained for the usage of websites were ease of use(64%), cost effectiveness(75%), time saving(52%).

Parry and Tyson summarised that online approaches increased the overall efficiency and self-satisfaction among users. Similarly, a study conducted by Sylvia and Mol revealed that online methods were more user friendly. Most of the candidates were satisfied with the features via online approach (P.Bringula, 2013). These facts answer the question why this research is focused on building a web application. Web application is the best and convenient medium for users to grasp knowledge regarding future job opportunities. Web application discussed in this research work is developed using programming languages like html and CSS. An advanced feature of this web application is that its responsiveness, ability to reach out and perform in all devices, adaptable in iOS, android and windows phone. This website does not need to be downloaded and installed from app stores like android play store or apple play store. A simple URL is enough to grab this website into any device (yeeply, 2017).

2.7 CONCLUSION

The literature review provides a significant insight into the domain of machine Learning. The studies conducted by people are taking wide variety of approaches There are many ongoing researches on these machine learning techniques. Research studies seen in this section highlights the potential advantages and popularity of machine learning. Literature review enabled to grasp more knowledge regarding the different
algorithms. Most of the studies mentioned in this chapter have repeatedly discovered the key merits in relying on machine learning technology.
CHAPTER 3: RESEARCH METHODOLOGY AND METHODS

3.1 INTRODUCTION
Research methodology is the analysis of theory behind a research work. A research methodology is needed to achieve dissertation aim and objectives (Abdulquadri Ade Bilaua, 2017). A research method is an organised and systematic investigation of a problem with an intention to solve it (Mohamad, 2015). In this dissertation, a technical framework is designed to provide future information to the users. The purpose of this study is to get a clear understanding of job vacancies, intakes and salaries associated with different sectors like administration, management, finance, IT etc. for users in the future years. As a result, some common problems like career planning issues, lack of information about future job prospects etc. gets minimized.

3.2 RESEARCH PHILOSOPHY
Research is a process through which people try to boost their ability and knowledge by learning new concepts, theories, ideas, models, graphs etc. in an organised way. Research methodology is very important in a study. Initial development of the research process can be visualized in the form of an onion as shown in figure 1 (Mohamad, 2015). Overlapping layers in the onion structure is very helpful in understanding design of the research methodology (Hassani, 2017). Research philosophy, the first layer

Figure 1 The Research "Onion" (Mohamad, 2015)
on research onion signifies an idea regarding the analysis and interpretation of data. This layer influences strategy of the researcher with respect to the research method he or she adopts. Moreover, it results in knowledge development on a specific field (Mohamad, 2015). A positivistic approach is taken in this dissertation on the belief that experiments, and observations impart true knowledge. In this study, quantitative methods are used to measure certain values and an attempt is done to determine the relationship between attributes or variables (Mark Saunders, 2009).

3.3 RESEARCH APPROACH
Second layer of the research onion is research approach, where two types of approaches are normally considered: inductive approach and deductive approach (Mohamad, 2015). In inductive approach, a theory is developed after the researcher has observed the results and alternative explanations can be derived from the process. There is no rigid framework in inductive approach. Whereas deductive approach has a rigid framework and the test must follow this framework to proceed to the results. A theory already exists, and the researcher must prove this theory in deductive approach (Rahi, 2017). In this thesis we are following both deductive and inductive approaches. A rigid framework is designed using tools like angular, Microsoft Azure studio, RapidMiner, Tableau and charts.js. Firstly, model must be designed via RapidMiner, azure studio and the model is fit into this framework built on angular platform constituting visualization tools like Tableau, charts.js. Existing tools are used to build a framework into which new data is loaded. Inductive approach is now considered for the data analysis part where a theory based on the accuracy, efficiency, root mean squared error etc. is developed after observing the results.

3.4 RESEARCH STRATEGY
Experimental stage of this research work begins with data investigation. Data of several departments are collected from different sources. This data is assembled, integrated and cleaned up. There are many challenges in integrating these data together as different departments keep data in vivid formats (Ian H. Witten, 2017). Several data types are present in the collected dataset, namely polynomial, real, integer, date, time etc. These data types play an important role while passing them into each operator in tools like RapidMiner and azure studio. A list of 600 rows of data was obtained from one company and a list of 809 rows of data was obtained from another company. Both these data are combined and mixed. Therefore, a total of 1409 rows of data is taken into consideration in this study. Number of vacancies, number of intakes, salary, year of experience, opening date of the vacancy, deadline, year, quarter, department, job title, eligibility and job description are the attributes used in the dataset. Out of these attributes number of vacancies, intakes, salary and year of experience required are the label attributes selected for the prediction purposes. After cleaning the data, all data is converted into a common format. A few rows of data are picked up from the whole dataset for verification. Some basic statistical concepts are also used in this study to get a better understanding of the results. Mean or average can easily be obtained on dividing the
sum of all values by the number of values. Median is the number present in the middle of a list of regularly ordered numbers. When the list contains odd number of values, then it is easy to find the median (satyam, 2017). Average of the squared differences from mean is called variance. Variance is given by the following equation:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

The square root of variance gives standard deviation. There are two factors which determine the relationship status between the attributes: covariance and correlation. Covariance is the measure of degree in which two variables vary in tandem to their means, whereas correlation is the measure of movement of two variables with respect to each other. If correlation coefficient is 1, then it is a perfect positive correlation and if it is -1, then it is a perfect negative correlation (bhattacharjee, 2017). Chi-square goodness of fit test and chi-square test for independence are generally the two types of chi-square tests. Whether a sample data matches a population is determined by the chi-square goodness of fit test while the latter test checks if there is a relation between two variables. Formula for the chi-square test is given by

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

where $O_i$ represents the observed value and $E_i$ represents the expected value. There exists a high correlation if test results have low chi-square value. It means that if observed and expected values are equal, then the chi-square value will be zero which is very rare to occur. If correlation is more, then the attributes used are not good. Chi-square tests are mostly used on numbers, gives the difference between observed counts and expected counts (Stephanie, 2018).

Machine learning algorithms are broadly classified into two: Supervised machine learning algorithms and unsupervised machine learning algorithms. If the objective is to identify structural characteristics of the data and the dataset is not provided with quality labelling, then it presents an unsupervised learning. There are extreme variations in unsupervised learning algorithms which make it difficult for a unified treatment. In supervised learning, proper labels are provided for data and the objective is to study a correct mapping from data to their labels (Krishnamurthy, 2015). Some of the supervised machine learning algorithms are Linear regression algorithm, neural network algorithm, support vector machine algorithm, decision tree etc. (Osisanwo F.Y., 2017). During the proposal stage of the dissertation it was planned to use only three algorithms namely linear regression algorithm, naïve Bayes algorithm and k nearest neighbours(KNN). The above proposal was based on the understanding that any algorithm can be used for the prediction purposes, but later needed reconsideration. Each algorithm has got certain specifications and parameters which are very important in prediction. After further research, it was discovered that other algorithms like
neural network, decision tree, support vector machine in addition to the linear regression algorithm. In this dissertation, there are attributes containing numerical and polynomial values. Linear regression algorithm helps to determine the nature of relationship between two attributes via an equation. A simple prediction of any $y$-value is possible using linear regression in the range of respective maximum and minimum observed $x$-values (Inna Boldina, 2015). Computational efficiency, ability to generate clear rules and ease of interpretation are some of the benefits in using decision tree algorithm (Yen-Liang Chen, 2016). Neural network algorithm is adaptable in mapping experimental functions even if there are non-linear negative issues inside these functions.

Network testing and network training are two standard methods provided by the neural network algorithm to build a model (Rita Fontanella, 2018). Similarly, support vector machine algorithm is very much flexible as it performs well in many cases like for datasets with many attributes and cases with only few contributes to the training process. On realizing the merits of algorithms like neural network, decision tree etc. it was decided to pursue these algorithms in this dissertation to predict job features which will be beneficial for the users. Another reason to replace naïve Bayes algorithm and KNN is the nature of complexity and improper assumptions that result in errors (Anchal Tomar, 2016). Research works on computing domain require an ability to realise algorithms that address certain defined problems and to recognize their strengths and weaknesses. Accuracy, efficiency, coefficient of determination and root mean squared error are some of the factors to be considered in the science of computing. Selection of appropriate algorithms required for the specific purpose and the application of those algorithms is a significant part of computing researches (Hassani, 2017).

Machine learning techniques used for training are train or test split, cross validation and k folds cross validation. In train or test split method, dataset is split into training set and testing set in a proportion like 70 percent and 30 percent respectively. Model built tries to study the training data which contains a known output. After that test data is passed into the model to determine the performance of the model. In cross validation method, data is split into k subsets and training takes place on k-1 subset. Cross validation is like train/test split but applied to further subsets. K folds is one type of cross validation method in which k-1 subsets are used to train the data and last subset is left as test data (Bronshtein, 2017). Evaluation criteria of a model depends on certain factors like mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), coefficient of determination etc. where mean absolute error measures the closeness of predictions to the actual values, root mean squared error summarizes error in the model, relative absolute error gives relative absolute difference between the actual and expected values and coefficient of determination is the measure of predictive power of the model respectively (azure, 2018). Data visualization is enabled using trending tools like tableau and charts.js. Using drag and drop, required attributes can be placed on the tableau for plotting. One significant advantage of tableau is the
easy to use interface built on VizQL which enable to describe plenty of visual presentations of data like tables, maps, time series etc. Users can filter and sort their data, drill into deep information using queries, perform sophisticated calculations and perform statistical analysis using the tableau desktop. Rapid analysis is possible using the in-memory data engine of tableau (Tableau software inc, 2015).

3.5 STRENGTHS AND WEAKNESSES
During the proposal stage of this dissertation, it was decided to use PHP as the coding language and tools like WampServer or adobe Dreamweaver were planned to deploy for the development and execution of the web application. Some disadvantages like variation in the quality of PHP frameworks, unavailability of options to modify core characteristics and patterns, negative impacts on the speed and performance of websites etc. were observed on some works after using PHP framework (Mindfire solutions, 2017). On realizing these demerits, it was planned to use angular platform with basic html 5 and CSS for coding. Angular technology is a framework managed by Google, enables smooth performance of web and fully extendable. One significant advantage of angular technology is that it works better with other libraries and features can be modified as per the needs of developer. Angular follows a Model-View-View Model(MVVM) architecture in which the application is split into different parts for easy development and understanding. Angular is also provided with built-in dependency injection subsystem that aid the developer to build and test applications with a proper understanding of outputs and errors. These features enhance further progressive development of applications in the future (Ambulkar, 2016). Similarly, at the initial stage of this research work MATLAB and RapidMiner were chosen as the prediction tools. But when I thought of developing an application for future prediction, I understood that there is a need to employ a much more promising tool that offer advanced features. As a result, it was decided to use Microsoft azure studio as the main prediction tool. Another reason to switch to azure studio is its ease of use, setting up this research as a starting point for all the present and new researchers who are keen on exploring more on machine learning domain. A major strength of this dissertation is the methodology adopted to reach to the data analysis stage. The purpose of the research was taken into consideration at first. Both inductive and deductive approaches made a better progress in the journey. Positivistic nature
highlighted true knowledge from the experiments. The entire process is illustrated in figure 2 (Abhishek Kaushik, 2016).

Visualization tools like Tableau and charts.js deliver a graphical representation of the future data. After data collection, a proper analysis of data took place on finding the aggregate meanings (Abhishek Kaushik, 2016). This significant feature of the web application provides a clear picture of required information for the user. General steps used in the visualization stage is shown in figure 3. RapidMiner used in this dissertation also contributes a leading role in both visualization stages as well as prediction stages. Another feature that makes this application interesting is the introduction of API’s. Azure API’s and google map API make the application more realistic in nature. One major weakness of this dissertation is that the research work is limited to a single district in Kerala, Trivandrum. Another factor that adversely affect this research work is the unavailability of sufficient data. Due to the security reasons, only limited amount of data was obtained.
3.6 SAMPLE
This section of the research work is also another instance of iterative operations. During the initial stages of this dissertation, it was assumed that there was a clear source for the required data. Many IT companies, resorts, servicing companies, science academies, travel agencies etc. were identified to collect data for the prediction purposes. Unfortunately, a few cyber-attacks occurred in India created huge financial damages (PTI, 2018). A recent report on ‘The Indian Express’ stated that financial damages of more than $500,000 occurred after the attack. As a result, data security became the primary concern of each company. Most of the companies invested large capital in their cyber security departments and tried to make their security systems stronger to avoid all possibility of breaches (PTI, 2018). This situation has adversely affected the availability of data required for prediction purposes. On realising that procedures to get data from companies became more complicated, a detailed mail containing the need of data and willingness to follow ethical factors was sent to the companies. A list of 600 rows of data was obtained from ‘Advanced Institute of Maths and Science’ located on Trivandrum. Similarly, another list of 809 rows of data was obtained from a ‘Amgo Tour and Travels company’ in Trivandrum. To predict job features for several departments, it was required to get data of different departments. A sampling frame was constructed with the data obtained.
3.7 ETHICAL PROCEDURES
Ethical factors to be considered in this dissertation were the following:

1) Confidentiality of data: The need to meet this factor is very much essential in this dissertation. All the data of several departments collected from different sources are used only for this research work. The collected data is not transferred or copied to any other devices.

2) Data privacy: Once the data containing company details was obtained, the data was taken care properly. Proper privacy measures were maintained for the retrieved data. There is always a heavy competition in the job industry, so it is very essential to maintain data privacy.

3) Informed consent: The companies were informed regarding the nature of research and permission was asked to use their data for the prediction purpose. Security clearance forms are supplied in the appendix.

3.8 CONCLUSION
This study is based on predictive analytics in which machine learning algorithms are used to predict certain job prospect features in the Trivandrum city, Kerala. Data comprises mainly numerical values, so linear regression algorithm and classification algorithms like decision tree algorithm, decision forest algorithm, support vector machine algorithm and neural network algorithm are used for the prediction purposes in this dissertation. Artefact design involves a web application using angular platform, where user chooses a year. Then job features like number of vacancies, salary, number of intakes, years of experience required are predicted for the future years that the user inputs. Visualization effects are added using tableau and charts.js to impart a good understanding of future data. A positivistic approach is necessary to determine the relationship between variables. Similarly, both inductive and deductive approaches are required to build a solid framework in this study. The methodology and approach mentioned in this chapter is very essential to achieve the objectives of this dissertation
CHAPTER 4: ARTEFACT DESIGN AND DEVELOPMENT

Figure 4 illustrates the use case diagram of backend technical processes in the application. A user enters the web application, each algorithm is used to determine a prediction model and a forecasting model. Prediction model gives the current results and the forecasting model gives results for the future. User is provided with a form and options are provided to select different job prospect features. User understands the results clearly in the visualization stage. Technical tools used in this study are described in the following sections.

Figure 4 Use case diagram of the process
4.1 RAPIDMINER
RapidMiner has three main windows: design window (space to design a model), results window (to view the results) and an auto model window (to get a prior idea of prediction). In this study, data in the form of excel sheet is given as input using ‘Read Excel’ operator. This data is then given to ‘Filter Examples’ operator, where missing values under an attribute are either replaced by a value or ignored. Next step is to set label for an attribute which is to be predicted. This is done by ‘Set Role’ operator. After selecting label attribute, an operator called ‘Select Attributes’ is used to remove certain polynomial attributes from passing into the prediction stage. Then comes the splitting stage where the data is split using ‘Split Data’ operator into training and testing sets in 70 percent and 30 percent respectively. Seventy percentage of the training data is given to the model for training and the rest thirty percentage of the data is given to operator ‘Apply Model’. At this time, required algorithm operators like decision tree, support vector machine, linear regression etc. can be introduced into the design page.

The operator ‘Apply Model’ is used to test the built model. To increase the efficiency of model, predicted output in real data type are converted to integers using ‘Real to Integer’ operator. Evaluation of the model is done using ‘Performance’ operator. A model shown in figure 5 is built using ‘decision tree’ operator, gaining a root mean squared error value of 0.647 and squared correlation of 0.897. Similarly, a model built using ‘support vector machine’ operator shown in figure 6 gives a performance vector containing root mean squared error value of 0.715 and squared correlation value of 0.867. ‘Forward Selection’ operator is used in this design to optimise the results.

![Figure 5 RapidMiner model using decision tree algorithm](image-url)
After creating the models, predicted results were stored in an excel sheet using ‘Write Excel’ operator in RapidMiner. This output from the RapidMiner was given as input to the Tableau. The purpose of this method was to visualise predicted results from RapidMiner in a more interactive way for the user. Department wise classification is made possible using both statistical and machine learning methods satisfying the objective of this dissertation. Tableau public is an open service to publish data visualizations. The visualizations that are published into Tableau Public are embedded in the web application. Embed code for the tableau work containing predicted results are used in coding section of this dissertation to
display an interactive visualization for the user. Figure 7 shows a tableau workbook containing department wise classification of the predicted results from RapidMiner.

![Figure 7 Tableau workbook](image)

### 4.3 MICROSOFT AZURE STUDIO

To bring RapidMiner API into the web application, RapidMiner server was required. RapidMiner server is not a free service provider, license was compulsory to make use of this server. On realising that license was very expensive, it was decided to use Microsoft azure studio at this stage for forecasting purposes.

![Figure 8 Azure studio model using decision forest regression operator](image)
Microsoft azure machine learning studio is a drag and drop tool just like RapidMiner. The similarity of azure studio with RapidMiner made me interested to learn more about azure studio. Azure studio is the trending tool around the world in data analytics industry. It provides an interactive and visual workspace for fast building and testing of models. Models were created using the operators in Azure studio. Firstly, input dataset is loaded into azure studio in the form of a CSV file. This dataset is then loaded into ‘Select columns in dataset’ operator to select the attributes to pass for prediction purposes. After the selection procedure, dataset is passed to ‘split data’ operator, where the dataset is split into training and testing sets in the proportion 70 percent and 30 percent respectively. In addition to this splitting technique, dataset is also passed to ‘cross validation model’ operator where model training takes place in cross validation technique. Another method used in this tool is the RF technique in which random splits are given at the time of training stage. The outputs of training techniques are given in the findings section of this dissertation.

‘Train Model’ operator is where the special attribute is selected for prediction. Algorithm operators like linear regression, decision forest regression etc. are employed at this stage. Figure 8 shows an azure studio model built using decision forest regression. After training the model, testing stage arises. ‘Score Model’ operator is used to test the model with testing dataset. After testing the model, results from different techniques are combined using ‘Join Data’ operator, which is optional. ‘Evaluate model’ operator is then applied to determine performance of the model. This operator delivers some values like Mean absolute error, root mean squared error, relative absolute error, relative squared error and coefficient of
determination which are crucial factors to evaluate the model performance. Azure studio allows to publish these models as web services which are used in the web application. Figure 9 shows a web service experiment created on azure studio using linear regression algorithm.

Azure portal was used to create API management gateways, through which advanced interactive options were discovered for the user. ‘200 series’ ok messages were enabled in the created services. The

![Figure 10 Azure API management services window](image)

Figure 10 shows an azure API management Service window created by the name ‘job prediction API gateway’ containing services for linear regression, decision tree etc. Azure application gateway controls traffic in the web application. Requests made from html page will be send to the gateway for processing. Based on request, gateway routes the traffic to appropriate services created in the portal. Response from the portal will be delivered to the gateway and hence reaches the application as shown in the figure 11.

![Figure 11 Gateway data flow diagram](image)
4.4 ANGULAR PLATFORM

Angular acts as a basic framework for the creation of a dynamic web application. Angular is a hybrid HTML or JavaScript framework that allows to create web applications in a simple manner. The graph shown in figure 13 highlights the pace of surging demand for angular around the world.

![Graph showing increasing demand for Angular technology](image)

**Figure 13** Increasing demand for angular technology (Chauhan, 2015)

‘Visual studio Code’ is the tool used in this dissertation as the text editor for angular. One significant feature that this tool provides is the code formatting option. On selecting ‘format code’ from the menu, all the codes get arranged in an organised manner. No extensions are required for this feature. Normally, angular uses typescript for the development tasks. To run on any web browser, this typescript is then
compiled to JavaScript. This marks the need to have a quality typescript tool for developing angular applications (Wycliffe, 2017). The figure 12 shows ‘Home page’ of the web application opened in visual studio code. To create a new page for the web application, it is required to create a new component each time in the visual studio code. A component is created using the command ‘ng generate component [component name]’ (Larsen, 2018). The data flow diagram shown below depicts authentication of a user who enters the web application. User must type his/her username and password to login to the web application. If the username and password entered by user are valid, then he or she gets logged into the home page of application. Figure 15 shows the login page of the web application.

![Login data flow diagram](image)

Figure 14 Login data flow diagram
The figure 15 shows the login page of the web application where the user enters after logging in.

Figure 15 Login Page of the web application

The figure 16 shows the home page of the web application where the user enters after logged in.

Figure 16 Home Page of the web application
On clicking ‘MODELS’ on the navigation bar, second page of the application is loaded which is shown in figure 17. In this page, options are provided to visualize the predicted models.

Visualization from Tableau is embedded in this page. The embed code of each data visualization is added.
under each algorithm. As a result, on clicking buttons on this webpage respective visualizations can be observed by the user. The figure 18 shows predicted results based on the number of vacancies using decision tree algorithm. This figure in Tableau shows an interactive graphical representation of the predicted results.

Similarly, on moving to the third option ‘FORECAST’ shown on navigation bar as in figure 19, a window appears where buttons are provided to go to the prediction pages under each algorithm. Prediction forms are displayed under each algorithm. In a prediction form, options are provided in the form of a drop-down menu, where the user can select special attributes like number of vacancies, number of intakes, salary, years of experience required to predict for the future years. Calendar is provided in the input form to select the year for which user wants data predicted. Some space is provided in the prediction form itself to visualize instant prediction results. Figure21 shows predicted data visualizations via charts.js.

These features fulfil the main purpose of this dissertation. Users are provided with a very interactive web application. This web application is responsive in nature, so that users can use the application in different devices, as seen in figure 20.
Responsive characteristic of this application delivers the predicted results in every form of screen-width. The application is user-friendly and visualization effects provide a clear understanding to the user. Calendar provided in the prediction input form enables user to select any year for which he or she wanted the data to get predicted. All the outputs delivered by this application is presented in the findings section of this dissertation. It also includes the pictorial representation of the predicted results via charts.js. Users get future information about the number of vacancies, number of intakes, salary etc. Proper career planning in terms of choosing courses to study and developing job specific technical skills for the demanding sector are some of the changes that user can adopt after using the application. In this way challenges faced by user in getting a job can be minimized. The entire unemployment crisis cannot be wiped out all sudden by using this application. But this application can bring a significant helping hand for the users who are suffering from unemployment.
API stands for ‘Application Program Interface’. API is defined as a set of routines, protocols and tools for developing software applications. API allows several programs, devices and applications to communicate with each other to enhance the automation processes (Johansson, 2016). Google map API is used in this web application to provide the location as shown in figure 22. The coordinates of Trivandrum, latitude: 8.5241 and longitude: 76.9366 are added in the coding section. Wide communication skills, unified access for communication channels and responsive customer service are some of the benefits that can be achieved in this web application by using furthermore API’s in the future (Johansson, 2016).
CHAPTER 5: FINDINGS

5.1 INTRODUCTION
In this chapter, findings of the dissertation are briefly presented, and further discussions will be provided in the following chapter. Primary results of the designed machine learning models, test results, predicted data visualizations etc. are shown in this section.

5.2 DATA FLOW DIAGRAM OF TEST RESULTS

Figure 23 Azure data flow diagram containing tests

Figure 23 illustrates the data flow diagram built in azure to determine the test results.
Figure 24 shows the data flow diagram built in azure to determine the prediction accuracy for different algorithms. Subsequent test results are displayed in this chapter and the description of each figure is also presented in this section:

Figure 24 Azure data flow diagram for algorithms
Figure 27 Chi-square test results

Figure 27 shows the chi-square test results obtained in azure studio. A detailed description of chi-square test is provided in the summary of this chapter. Figure 25 and figure 26 shows the split and cross validation training test results for a sample dataset containing 100 rows. Split technique uses ‘split rows’ mode to divide the dataset into two parts namely training set and testing set in a specified proportion like 70 and 30 respectively. Then the model is trained using training set and tested using the testing set. In cross validation technique, dataset is divided into a few folds (10) and a model is built for each fold. Nine out of ten data is used for training the model and the remaining one is employed for testing (Roope Astala, 2018).
5.3 LINEAR REGRESSION TEST RESULTS

Linear regression algorithm is used in RapidMiner and azure to build models based on different job attributes. Histograms, test results, plots etc. are displayed in this section:

Figure 28 Linear regression on Number of Intakes
Figure 29 Linear regression on Number of vacancies

Figure 28, figure 29 gives the mean, median, standard deviation values for a sample dataset on number of intakes and number of vacancies respectively. Figure 30 gives a comparison between results obtained through split technique and cross validation technique.

Figure 30 Scored label from split technique Vs Scored label (2) from cross validation technique
Figure 32 shows training and testing scores obtained in azure using the linear regression algorithm based on number of vacancies using split and cross validation techniques respectively. Coefficient of determination (COD) obtained via split technique based on number of vacancies is 0.834052 and via cross validation technique COD is 0.835158. Root mean squared error (RMSE) secured for the training score is 3.222899 and RMSE obtained for testing score is 3.175298. There is less variation between the values.
Figure 33 shows the statistical prediction done in tableau and figure 34 highlights the prediction done by the application on user demand.
In the tableau visualization shown in figure 33, a sector wise prediction results can be seen based on salary and user is able to view only the required sectors by selecting them on the right-side panel. Management sector is found as the emerging sector in terms of salary. In the application, visualization tool charts.js enables to pop out the instant predicted values on pointing to different coordinates on the graph. Figure 36 and figure 35 gives the predicted results visualization using split and cross validation training techniques in azure. Standard deviation (SD) obtained via split technique and cross validation technique is 7.3807 and 7.4216 respectively. Taking the square of SD values give variance. Median for the used dataset is found to be 14.
Figure 37 presents a scatter plot between the attributes obtained in tableau. Figure 38 gives the 3D structure of scatter plot for number of vacancies. Red dots represent predicted values in 3D plot.

Figure 37 Scatter plot on Tableau of linear regression on vacancies

Figure 38 Scatter plot 3D on RapidMiner for linear regression on intakes
Figure 40 and figure 39 give the density box plot or whisker plot graphs obtained in tableau and azure respectively for two labels namely vacancies and salary respectively. Median on salary plot lies between 400k and 200k, minimum value is 0k and the maximum is just below 1000k. Median on vacancy plot lies between 15 and 10, minimum value is 0 and the maximum is above 30.
5.4 DECISION FOREST TEST RESULTS

Decision Forest algorithm is used in azure to build models based on different job attributes. Histograms, test results, plots etc. are displayed in this section:

![Histogram](image1)

Figure 41 Decision Forest on Salary

![Box Plot](image2)

Figure 42 Decision forest on salary in Azure

Figure 41 gives the mean and median on salary for a sample dataset. Figure 43 highlights the prediction done by the application on user demand. Figure 42 gives the density box plot obtained in azure for the label salary.

![Application Prediction](image3)

Figure 43 Application prediction
Figure 45, figure 44 gives the predicted results visualization using split and cross validation training techniques respectively in azure. SD obtained via split technique and cross validation technique is 7.7271 and 7.6169 respectively. Taking the square of SD values give variance. Median for the used dataset is found to be 15.
Figure 46 Error Statistics of Decision Forest on Number of vacancies using split technique (Training score)

Figure 46 and figure 47 shows training and testing scores obtained in azure based-on the number of vacancies using split and cross validation techniques respectively. COD obtained via split technique based on number of vacancies is 0.854323 and via cross validation technique COD is 0.861207. RMSE secured for the training score is 3.01965 and RMSE obtained for testing score is 2.913622. There is less variation between the values.

Figure 47 Error Statistics of decision forest on Number of vacancies using Cross validation technique (Testing score)
5.5 NEURAL NETWORK REGRESSION TEST RESULTS

Neural Network algorithm is used in RapidMiner and azure to build models based on different job attributes. Histograms, test results, plots etc. are displayed in this section:

Figure 48 application gives tableau result on number of intakes giving sector wise statistical analysis for a user.

Figure 48 shows the statistical prediction done in tableau. Figure 49 gives the 3D structure of scatter plot for salary. Green dots represent predicted values in 3D plot.

Figure 49 Scatter plot 3D in RapidMiner
Figure 50 highlights the prediction done by application on user demand. Figure 51 shows an instance of the network model obtained on RapidMiner.

Figure 51 Neural network regression model
Figure 54 gives the mean and median on years of experience. Figure 53 and figure 52 gives the density box plot or whisker plot graphs obtained in tableau and azure respectively for two labels namely vacancies and year of experience respectively. Median on vacancy plot lies between 200 and 100, minimum value is 0 and the maximum is just below 600. Median on year plot is 4, minimum value is 1 and the maximum is 6.
In the tableau visualization shown in figure 48, sector wise prediction results can be seen based on number of intakes and user is able to view only the required sectors by selecting them on the right-side panel. Management sector is found as the emerging sector in terms of number of intakes. Figure 55, figure 56 shows training and testing scores obtained in azure based on years of experience using split and cross validation techniques respectively. COD obtained via split technique based on years of experience is 0.949481 and via cross validation technique COD is 0.945949. RMSE secured for the training score is 0.316922 and RMSE obtained for testing score is 0.331279. There is less variation between the values.
Figure 57, figure 58 gives the predicted results visualization using cross validation and split training techniques respectively in azure. SD obtained via split technique and cross validation technique is 1.4674 and 1.4959 respectively. Taking the square of SD values give variance. Median for the used dataset is found to be 4.
5.6 DECISION TREE TEST RESULTS

Decision Tree algorithm is used in RapidMiner and azure to build models based on different job attributes. Histograms, test results, plots etc. are displayed in this section:

Figure 59 Tableau showing predicted results on salary giving sector wise statistical analysis for a user

Figure 60 highlights the prediction done by the application on user demand. Statistical analysis of the predicted results on salary using tableau is shown in figure 59.

Figure 60 Application Prediction
In the tableau visualization shown in figure 59, a sector wise prediction results can be seen based on number of vacancies and user is able to view only the required sectors by selecting them on the right-side panel. Management sector is found as the emerging sector in terms of number of vacancies. Figure 63 and figure 61 gives the density box plot graphs obtained in tableau and azure respectively. Median on number of intakes plot lies between 200 and 100, minimum value is 1 and the maximum lies between 600 and 500. Median on years plot lies between 12 and 10, minimum value is 0 and the maximum is above 24. Figure 62 gives the mean, median, standard deviation values for a sample of dataset on years of experience.
Figure 65 Error Histogram of decision tree on Number of vacancies using split technique (Training score)

Figure 64 Error Histogram of decision tree on Number of vacancies using Cross validation technique (Testing score)

Figure 65, figure 64 shows training and testing scores obtained in azure based on number of vacancies using split and cross validation techniques respectively. COD obtained via split technique based on number of vacancies is 0.861851 and via cross validation technique COD is 0.870084. RMSE secured for the training score is 2.940594 and RMSE obtained for testing score is 2.818913. There is less variation between the values.
Figure 66 presents a scatter plot between the attributes obtained in tableau. Figure 67 shows the model of decision tree secured.

Figure 66 Scatter plot on Tableau

Figure 67 Decision tree model in RapidMiner
Figure 68 gives the predicted results visualization using split and cross validation training techniques in azure. SD obtained via split technique and cross validation technique is 7.6595 and 7.6361 respectively. Taking the square of SD values give variance. Median for the used dataset is found to be 15. Figure 69 gives the 3D structure of scatter plot for number of vacancies. Red dots represent predicted values in the 3D plot.
5.7 COMPARISON TEST RESULTS
Training scores for different algorithms based on number of vacancies using cross validation training technique is displayed in the following table:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mean Absolute Error</th>
<th>Root Mean Squared Error</th>
<th>Relative Absolute Error</th>
<th>Relative Squared Error</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECISION FOREST</td>
<td>2.216926</td>
<td>3.235113</td>
<td>0.309718</td>
<td>0.151668</td>
<td>84.8 %</td>
</tr>
<tr>
<td>LINEAR REGRESSION</td>
<td>2.164397</td>
<td>3.156273</td>
<td>0.302379</td>
<td>0.144366</td>
<td>85.5 %</td>
</tr>
<tr>
<td>POISSON REGRESSION</td>
<td>2.763619</td>
<td>3.8929</td>
<td>0.386094</td>
<td>0.219615</td>
<td>78.0 %</td>
</tr>
<tr>
<td>DECISION TREE</td>
<td>2.209144</td>
<td>3.085834</td>
<td>0.308631</td>
<td>0.137994</td>
<td>86.2 %</td>
</tr>
<tr>
<td>NEURAL NETWORK REGRESSION</td>
<td>2.107977</td>
<td>2.975091</td>
<td>0.294497</td>
<td>0.128267</td>
<td>87.1 %</td>
</tr>
</tbody>
</table>

Table 1 Test results summary
The label ‘number of vacancies’ is the total number of vacancies available in a company and the label ‘number of intakes’ is the number of candidates selected for the job. Training scores for multilayer perceptron based on other attributes is given in the table 2.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Mean Absolute Error</th>
<th>Root Mean Squared Error</th>
<th>Relative Absolute Error</th>
<th>Relative Squared Error</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER OF INTAKES</td>
<td>2.10053</td>
<td>2.866903</td>
<td>0.342897</td>
<td>0.155799</td>
<td>84.4 %</td>
</tr>
<tr>
<td>YEARS OF EXPERIENCE</td>
<td>0.255482</td>
<td>0.339599</td>
<td>0.209563</td>
<td>0.056355</td>
<td>94.3 %</td>
</tr>
</tbody>
</table>

Table 2 Multilayer perceptron for other labels

Multilayer perceptron neural network based on different algorithm parameters is tested in this study. Default parameters on MLP determined coefficient of determination as 87.1 %. The maximum coefficient of determination was obtained when the number of hidden nodes were 10 and the random seed was 11. This alteration in the number of hidden nodes produced COD as 94.3 %. The results are displayed in the table 3.

<table>
<thead>
<tr>
<th>Number of hidden nodes</th>
<th>Learning rate</th>
<th>Number of learning iterations</th>
<th>Initial learning weight</th>
<th>Momentum</th>
<th>Random number seed</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Use range builder 0.01,0.02,0.04</td>
<td>20,40,80,160</td>
<td>0.1</td>
<td>0</td>
<td>7</td>
<td>89.4 %</td>
</tr>
<tr>
<td>100</td>
<td>0.005</td>
<td>100</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>91.0 %</td>
</tr>
</tbody>
</table>
Accuracy measures for the sector wise prediction using cross validation training technique is displayed in the following table:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mean Absolute Error</th>
<th>Root Mean Squared Error</th>
<th>Relative Absolute Error</th>
<th>Relative Squared Error</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINEAR REGRESSION</td>
<td>0.3501</td>
<td>0.59177</td>
<td>0.31121</td>
<td>0.2030</td>
<td>79.6 %</td>
</tr>
<tr>
<td>POISSON REGRESSION</td>
<td>0.651</td>
<td>0.8073</td>
<td>0.579</td>
<td>0.377</td>
<td>62.2 %</td>
</tr>
<tr>
<td>DECISION TREE</td>
<td>0.464</td>
<td>0.685</td>
<td>0.413</td>
<td>0.2707</td>
<td>72.9 %</td>
</tr>
<tr>
<td>NEURAL NETWORK REGRESSION</td>
<td>0.0137</td>
<td>0.046</td>
<td>0.0122</td>
<td>0.0012</td>
<td>94.4%</td>
</tr>
</tbody>
</table>

Table 3 MLP for different parameters

Table 4 accuracy measures for sector wise prediction in azure
CHAPTER 6: DISCUSSION

6.1 INTRODUCTION

In this section implication of the findings are discussed and the results are interpreted. Research questions are reviewed in this chapter to show the relation and contribution of work. Limitations of the research are also considered in this section with proper demonstration.

6.2 ADDRESSING RESEARCH QUESTIONS

How accurately the machine learning techniques of analysing and predicting job features of a company in Trivandrum city can be implemented and linked to a web application? Examine the prediction accuracy based on different sectors on the end to end user demands?

During the proposal stage of this dissertation, the proposal was to create a web application. There was no clear thought of making an application with upgradable options for future use. Interestingly, emerging tools like azure, RapidMiner, Tableau and charts.js have a crucial role in this dissertation, making this application a valuable tool for the future use. Deployment of angular platform in this application has made this web application more relevant. Both the research questions are answered statistically as in the findings section. Mean absolute error (2.107), root mean squared error (2.975), relative absolute error (0.294), relative squared error (0.128) and coefficient of determination (87.1 %) obtained for multilayer perceptron proves MLP as the best algorithm in this study. Test results of other machine learning algorithms are presented in Table 1. Similarly, accuracy measures of sector wise prediction are root mean squared error (0.046) and coefficient of determination (94.4 %). These findings illustrate the extent to which machine learning techniques can be used in analysing and predicting job prospect attributes through a web application. In the application, user can select a sector like IT, management, administration etc. for which the user needs prediction. The web application provides data regarding the various job prospect features for the future.

Difference between experimental values and expected values can be determined using chi-square test. Figure 27 shows chi square values presented in a table. If the chi square value is greater than critical value, then null hypothesis is rejected. The critical value considered is 0.05, here the chi square values are greater than 0.05. Therefore, null hypothesis is rejected. A dataset containing 100 rows of data is taken, split into 80 rows and 20 rows for training data and testing data respectively. This is the split technique. Similarly, another model is built for cross validation technique. Histograms for cross validation technique and split technique with cumulative distribution and probability density are shown in Figure 26 and Figure 25 respectively. Data visualization of predicted models in tableau and the results obtained from application which give values for different quarters on pointing to the graphs are presented in the findings.
chapter. Histogram of split technique and cross validation technique for each algorithm is also displayed in the findings section. Overall, the study results reflect a better implementation of predictive analysis using machine learning techniques. Prediction of job prospect features for a user based on different job sectors enable users to understand the future trends in each sector and acquire suitable skills for a job in the most prospective sector.

During this dissertation, it was very interesting to note the large number of literatures on machine learning domain. Literatures are always an indispensable building block for further discussions on a topic. As highlighted in the literature section, the four assumptions of simple linear regression model are adopted in this study and ARIMA forecasting method used by Dudek mentioned in literature section is replaced by azure studio forecasting. The decision tree algorithm findings reflect what was defined by Lei Ren in 2016 that to capture non-linearities and other effects efficiently, it is required to build interpretable models through decision tree. In this study, input dataset that affects the output data is used in building neural network regression model. This perspective is like the approach identified by Malgorzata Pawul in 2016.

**6.3 DEMONSTRATING DISSERTATION LIMITATIONS**

In this study, an attempt is done to minimise the challenges faced by job aspirants in getting information regarding the future job opportunities. This study tries to bring machine learning technology into the hands of users through a web application. Now it is time to discuss the limiting factors of this dissertation.

Firstly, the prediction was limited to only a single district, Trivandrum, Kerala can be seen in figure 70. There are 14 districts in Kerala and the total population of Kerala is about 3.47 crore. The employment crisis is a serious issue throughout the state (indiaonlinepages, 2018).
Secondly, unavailability of enough data was a major limitation. A list of 600 rows of data was obtained from ‘Advanced Institute of Maths and Science’ located on Trivandrum. Similarly, another list of 809 rows of data was obtained from ‘Amgo Tour and Travels company’ in Trivandrum. The amount of data obtained was very less. As a result, some prediction graphs showed less variation for longer years as shown in figure 71. If more data is given to this application in training, then a user obtains better sector wise results regarding future job opportunities.

Figure 71 prediction graphs showing less variation
CHAPTER 7: CONCLUSIONS

The main purpose of this dissertation was to analyse and predict job prospect features in the capital city of Kerala, Trivandrum using machine learning techniques and to link them into a web application. The web application created for this study using angular platform provides an interactive interface for the job aspirants based on azure machine learning neural models. After training and testing five algorithms, an interactive session based on the prediction models from each algorithm was incorporated in the web application of the thesis. Multilayer perceptron neural network was observed to be the best algorithm meeting the requirements of this study.

Visualization tools used in this dissertation ease the understanding of the job aspirants with charts and graphical representations. As highlighted in the literature review section, that the orientation of data points in scatter plot determines the relationships (Massimo Guidolin, 2018). The scatter plots in this thesis deliver both attribute comparison and 3D visualizations of the predicted results. Error histograms of each algorithm provided in the findings section illustrate the performance of the model built in this study. The research shows that statistical analysis provided in the web application will be helpful for the job aspirants to understand current and future trends in the job market for different sectors. Application also delivers sector wise information regarding future job opportunities with visualization. Scored label visualizations in the findings section give a comparative study of machine learning training techniques. As mentioned in the literature review, regression analysis used in this study enabled to figure out the independent variable that influence on the dependent variable (Jeon, 2015). Similarly, universal approximation property and robustness were some of the advantages of using neural network in this thesis work (Dudek, 2015). Presently, unemployment is one of the crucial issues faced in Kerala. This dissertation reveals meaningful results that will help job aspirants to do proper planning and develop adequate technical skills for the jobs relevant in the future. In this way, I hope that this study can uplift a small change among the users to create a big positive impact on their lives. The findings in this dissertation answer the research questions and thereby meet the objective of this study.

At the initial stages of the study, I was not aware of critical thinking skills, but this dissertation has taught me a new way of approaching a project. This thesis helped me to develop a macro view and better understanding of a project work. Learning was more effective when I raised the requirements of the thesis progressively. I could also use more complex software’s like charts.js, azure studio etc. The study boosted my confidence level and would enhance the project management skills for the future works. During the proposal stage, it was planned to create a simple html application. Introduction of angular platform in this dissertation has made this work worthy. Angular platform is an emerging technology and it has enhanced the productivity of this application. Interactive nature of the web application creates a positive impact among users as they can choose the year and job sector to be predicted. There are some limitations in this
dissertation like insufficient data availability and limitation in the geographical area covered. Despite these limitations, this study constitutes a valuable contribution in the field of machine learning. I should collect more data to get better results. When the data is more, training dataset increases, and the prediction accuracy rates also get surged. I also must expand the prediction to other districts of Kerala. This is potentially be an area for the future research. Employment crisis is present throughout the state. Therefore, it is very essential to broaden this work for all the districts. Another recommendation for the future work is the conversion of this web application into an android application using PWA builder and android studio. This dissertation helped me to study a lot about machine learning. I would like to thank all the Dublin Business School faculties for providing the knowledge required to travel in the journey of thesis.
Available at: https://ac.els-cdn.com/S187770581830095X/1-s2.0-S187770581830095X-
main.pdf?_tid=69a3713b-73de-4007-8451-
db2dbdb88b07&acdnat=1533507236_aaa338fc464a45a50fa0c96718817fd4f
[Accessed 2 august 2018].

Available at: http://www.enggjournals.com/ijcse/doc/IJCSE16-08-08-012.pdf
[Accessed 15 August 2018].

Available at: http://www.enggjournals.com/ijcse/doc/IJCSE16-08-11-011.pdf
[Accessed 16 August 2018].

[Accessed 15 august 2018].

Available at: https://www.ijser.org/researchpaper/Angular-JS.pdf
[Accessed 7 august 2018].

[Accessed 6 august 2018].

Available at: https://pdfs.semanticscholar.org/455f/15c8e25875e3c5d5d5c74991588a62ed0439.pdf
[Accessed 8 august 2018].

Available at: https://timesofindia.indiatimes.com/business/india-business/number-indian-internet-users-will-
[Accessed 21 June 2018].

Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-
model
[Accessed 7 august 2018].

Available at: https://www.indiatoday.in/india/story/digital-india-internet-users-967920-2017-03-27
[Accessed 9 july 2018].

Available at: https://medium.com/technology-nineleaps/basics-of-statistics-for-machine-learning-engineers-ii-
d25c5a5dac67
[Accessed 6 august 2018].

Available at: https://www.researchgate.net/publication/277623045_MULTIPLE_LINEAR_REGRESSION_MODEL_OF_INFLATION_RATE_IN_SUDAN
[Accessed 15 July 2018].

Available at: https://www.predictivanalyticstoday.com/what-is-predictive-analytics/
[Accessed 15 July 2018].

Available at: http://www.indiaonlinepages.com/population/kerala-population.html
[Accessed 13 August 2018].

[Accessed 18 July 2018].

[Accessed 5 August 2018].

Available at: https://pdfs.semanticscholar.org/50a9/a4a1cf87575bbb83b43419102d09fc89f942.pdf
[Accessed 16 July 2018].

Available at: https://unified-communications.tmcnet.com/topics/unified-communications/articles/428094-how-apis-revolutionize-companys-communication-strategy.htm
[Accessed 8 August 2018].

Available at: https://www.scribd.com/doc/305930845/Article-on-Unemployment-in-Kerala
[Accessed 2 July 2018].

Available at: http://www.rroij.com/open-access/a-survey-on-machine-learning-conceptalgorithms-and-applications-.pdf
[Accessed 16 July 2018].

Available at: https://www.cs.cmu.edu/~aarti/pubs/AkshayThesis.pdf
[Accessed 6 August 2018].

Available at: https://github.com/angular/angular-cli/wiki/generate-component
[Accessed 8 August 2018].
Available at: https://ac.els-cdn.com/S187661021631760X/1-s2.0-S187661021631760X-
main.pdf?_tid=45b806ed-0b67-4cbb-9a2d-7003ac0074be&acdnat=1531830545_22073a1c0cb00200f592b22b60626f96
[Accessed 15 july 2018].


Małgorzata Pawul, M. Ś., 2016. APPLICATION OF ARTIFICIAL NEURAL NETWORKS FOR PREDICTION OF AIR

Available at: https://eclass.teicrete.gr/modules/document/file.php/DLH105/Research%20Methods%20for%20Business%20St-
udents%2C%205th%20Edition.pdf
[Accessed 6 august 2018].


Available at: https://www.researchgate.net/publication/318814623_Forest_PA_Constructing_a_Decision_Forest_by_Penalizi-
ing_Attributes_used_in_Previous_Trees
[Accessed 10 august 2018].

Available at: https://www.sciencedirect.com/science/article/pii/B978012809638203373
[Accessed 16 july 2018].

Available at: https://medium.com/@mindfiresolutions.usa/advantages-and-disadvantages-of-php-frameworks-
c046d50754e5
[Accessed 7 august 2018].

Available at: http://journals.sagepub.com/doi/pdf/10.5772/60429
[Accessed 29 july 2018].

Available at: http://neuralnetworksanddeeplearning.com/about.html
[Accessed 14 july 2018].

Available at: https://www.researchgate.net/profile/J_E_T_Akinsola/publication/318338750_Supervised_Machine_Learning-
Algorithms_Classification_and_Comparison/links/596481dd0f7e9b819497e265/Supervised-Machine-Learning-
Algorithms-Classification-and-Comparison.pdf
[Accessed 7 august 2018].


APPENDICES

Security Clearance

Dublin Business School
Company Security Clearance

Name: AKKU GEORGE SAJU
Student Number: 10376281

Company Security Clearance
Please initial as appropriate

✓ We agree that the student(s) may undertake a dissertation of the nature indicated above and that he/she/they will be given access to appropriate information sources within our organisation.

✓ We agree that copies of the finished project will be made available for assessment by staff of Dublin Business School and External examiners.

Company Name: AIMS-ADVANCED INSTITUTE OF MATHS AND SCIENCE
Signed: AKASH PRATHAP
Position: HR MANAGER
Date: 18/06/2018
DATE: 18/06/2018

TO WHOM IT MAY CONCERN

This is to certify that Mr. AKKU GEORGE SAJU has been provided the following data 2015-2017 for the purpose of completing his thesis program at Dublin Business School.

a) Department wise number of trainees / interns / freshers hired each year.

b) Department wise total number of vacancies.

c) Department wise employee strength each year.

The source data provided to Mr. AKKU GEORGE SAJU by email is true and correct.

[Signature]

HR Manager.
DATE: 18/06/2018

TO WHOM IT MAY CONCERN

This is to certify that Mr. AKKU GEORGE SAJU has been provided the following data 2015-2017 for the purpose of completing his thesis program at Dublin Business School.

a) Department wise number of trainees / interns / freshers hired each year.
b) Department wise total number of vacancies.
c) Department wise employee strength each year.

The source data provided to Mr. AKKU GEORGE SAJU by email is true and correct.

[Signature]

MANAGING DIRECTOR
Security Clearance

Dublin Business School
Company Security Clearance

Name: AKKU GEORGE SAJU
Student Number: 10376281

Company Security Clearance
Please initial as appropriate

✔ We agree that the student(s) may undertake a dissertation of the nature indicated above and that he/she/they will be given access to appropriate information sources within our Organisation.

✔ We agree that copies of the finished project will be made available for assessment by staff of Dublin Business School and External examiners.

Company Name: Amgo Tours and Travels
Signed: Mukesh M P
Position: MANAGING DIRECTOR
Date: 18/06/2018