|  |
| --- |
| **nORTHUMBRIA UNIVERSITY LONDON CAMPUS** |
| Computing and Digital Technologies Project LD7083 |
| Dissertation |
| **Title: Predictive analysis for the Immigration data of Nepal** |
| Name: Subash Pandey |
| ID: W20055480 |

|  |
| --- |
|  |

**Word count: 12009**

# **Declaration**

I declare the following:

1. That the information in the dissertation is a product of my own effort, and that all sources—printed, electronic, and personal—have received acknowledgement in the bibliography and in the references.
2. the word count of this dissertation is 12009.
3. If it is deemed acceptable, I consent to an entire electronic copy of the dissertation—or portions of it—being posted on Blackboard to provide future students the chance to see samples of previous dissertations, unless the confidentiality of this dissertation has been assured. I am aware that pupils could print off copies or download them if they were exhibited online for no more than five years. The authorship would be concealed.
4. I consent to have my dissertation submitted to a service that looks for plagiarism. The service will sort my dissertation into a database and compare it to work from this department, from any other department, and from other organizations that also use the service. In the case that the service finds a high degree of resemblance between content inside the service, it will notify my supervisor and second marker, who may elect to do additional research that could ultimately result in disciplinary action if plagiarism claims are made by students.
5. After reading the UNN/CEIS policy statement on research and consultancy, I can vouch for the fact that ethical concerns were considered, assessed, and properly resolved in this study.

**Signature: SUBASH PANDEY**

**Date: 07/09/2022**

# **Acknowledgement**

I'd want to express my gratitude to Parvez Jugon for serving as my supervisor and for his constant counsel, direction, encouragement, and criticism during the project. I also want to thank Rose Fong for assisting me in choosing the dissertation's topic.

Table of Contents

[**Declaration** 1](#_Toc113745157)

[**Acknowledgement** 2](#_Toc113745158)

[**Abstract** 4](#_Toc113745159)

[**Introduction** 4](#_Toc113745160)

[**Research Purpose** 5](#_Toc113745161)

[**Research Questions** 5](#_Toc113745162)

[**Aims and Objectives** 5](#_Toc113745163)

[**Scope** 6](#_Toc113745164)

[**Ethical Considerations** 6](#_Toc113745165)

[**Literature Review** 7](#_Toc113745166)

[**Research Methodology / Description of Practical Research Work Undertaken** 11](#_Toc113745167)

[**Research Philosophy** 12](#_Toc113745168)

[**Research Approach** 12](#_Toc113745169)

[**Discussion of practical work done** 14](#_Toc113745170)

[**Discussed ideas that were rejected** 15](#_Toc113745171)

[**Results and Analysis** 16](#_Toc113745172)

[**Predictive analysis techniques** 17](#_Toc113745173)

[**Neural Network** 17](#_Toc113745174)

[**Time Series** 18](#_Toc113745175)

[**Clustering** 19](#_Toc113745176)

[**Regression models** 19](#_Toc113745177)

[**Decision Tree** 20](#_Toc113745178)

[**Why use Linear Regression and Decision tree?** 21](#_Toc113745179)

[**Using Linear Regression** 22](#_Toc113745180)

[**Correlation coefficient** 24](#_Toc113745181)

[**For Total Number(TN)** 25](#_Toc113745182)

[**For TN by Land** 29](#_Toc113745183)

[**By Average stay(days)** 31](#_Toc113745184)

[**By growth rate** 32](#_Toc113745185)

[**Using Decision Tree** 33](#_Toc113745186)

[**Critical Evaluation, Conclusions and Recommendations** 38](#_Toc113745187)

[**Contributions to wider academic body of knowledge** 38](#_Toc113745188)

[**References** 43](#_Toc113745189)

[**Appendices** 45](#_Toc113745190)

**Abstract**

Data analysis appeared and started around the early 90’s. This can be termed as the process in computers for discovering the patterns among the very large and huge datasets. Using various ways and statistical methods results can be obtained which can be further used in decision making and planning. For the analysis of data storing of data was very necessary and thanks to the evolution of data warehouse and database technologies. So, due to which the companies can forecast and predict the sales and purchasing patterns. However due to the increasing use of the analysis in all forms and sectors of data this is very essential to go on depth of data analysis. Data analysis is used in all sectors such as health, military, sales, government data, etc. For our research and analysis, we are taking the data from 1964 to 2019. The data will have the number of people arriving in Nepal in a tourist visa. We will take the data such as total number, growth rate, Total number by air, Total number by land and average length of stay for our analysis. We will use these data to predict the upcoming people in next few years. The main objective is to predict the incoming people from the data collected using the predictive analysis. We will also visualize the data from the analysis made which should help in formulating further strategies. Positivism philosophy is followed while doing this research. This means only the accurate information is used as the observation.

# Introduction

New technologies are emerging rapidly which is also changing the behaviour and consuming pattern of the customers. Businesses, companies, and organizations are changing their way of analysis and thought. They are adapting the digitization and new trend i.e., data analysis. In the past few decades businesses and government organizations have gathered a lot of data. Now with the help of that data trends should be analysed and decisions should be made with the help of those. Without proper analysis and use the data, those data are just useless and their storage just consuming a lot of space. This research is based on a predictive analysis for the Immigration data of Nepal collected by the government which is available in the public domain. Immigration is one of the organizations where there are a lot of data regarding the number of people travelling from one place to another. Making decisions and developing strategies based on the facts rather than assuming what the data will show you allows you to make sense of the past and forecast future trends and behaviours. Businesses and organisations are much more able to comprehend their audience, industry, and firm when equipped with the insights gleaned from the data. As a result, they are better able to make decisions and establish long-term plans. There is a huge chance of analysing data in the Immigration field all over the world. The data ranges from number of people to airlines to types of visas, nationalities, etc.

## **Research Purpose**

The main purpose of the research is to help the decision makers giving the broad sense of the collected data. Decision makers may not be technical people and sometimes they don’t have idea on how to play with data, so the purpose and objective of research is to mainly give the stakeholders to take wise decision looking at the trends and patterns occurring in the data. Also, the predictive analysis should help them to understand data properly. Pictorial representation and visualization such as graphs make the decision makers easier to understand and make useful insights from them. Current knowledge synthesis should always be approached with rigour that is methodical, repeatable, and responsible in nature (Brooker, 2010). The main purpose is to understand and identify trends and patterns also to find out and remove the outliers if necessary. This is quantitative based research where the numbers of people travelling are used. In such types of research, it is relatively easier and consume less time to collect data as most of the time secondary data is used where some else has collected data. In our research the data, conclusion and findings are moreover straightforward rather than generalized. Statistical tests and analysis are used. This research offers some deep analysis and helps to find out more about the context of the research being done. Our research is also great for the purpose of further exploring the data. These qualitative data being researched are also the elements of prediction for the continuous data which are changing and adding up every certain time. The main numerical factors that are being researched is the numbers of people arriving in Nepal. It consists of data i.e., number of people arriving in certain year through land and air. Also, the growth percentage and average days of stay is analysed.

## **Research Questions**

There could be various questions regarding the following research being done. However, this research will focus on the following questions to address in the conclusion.

* How does the predictive analysis help in forecasting Immigration data?
* What influence the predictive model output can have in the decision-making process?

## **Aims and Objectives**

The main aim of the project is to analyse the historical data of people arriving in Nepal on tourist visa and predict the number of people who will arrive in next two years using predictive analysis technique including linear regression and decision tree. The aim is also to assist different organizations in decision making by using our insights, charts, and analysis.

Objectives of this research can be broken down into following:

1. Collecting the data: Data will be collected from the authorized government sites.
2. Data Cleaning: In the extracted data also, there are so many missing information we must adjust or remove to gain the maximum precision for our research.
3. Data Cleaning: These data extracted is a huge data. The data are so huge and divided in so many components that we must extract data which we are analysing. In the extracted data also, there are so many missing information we must adjust or remove to gain the maximum precision for our research.
4. Data Modelling (Visualizing): For the communication between data points and structures, data modelling is the process of developing a visual showing of an entire information system or certain components of it.

## **Scope**

The scope must explain and describe about the area that will be explored during the work progress of the project. Following is the scope of the research:

* Extracting data of people arriving in tourist visa from 1964 to 2019 in Nepal via land and via air.
* Analyse the trends using various statistical analysis like slope, intercept, linear regression, correlation coefficient between two variables.
* Visualizing all those analysis in graphs and make a detailed report for further analysis.

## **Ethical Considerations**

Even if any research is valuable to society, it is not fully justified if those are violating the human rights. On this research during all the processes and data collection process, all the ethical concerns were considered properly. All research studies should prioritise safeguarding human participants by implementing relevant ethical norms (Arifin, 2018). A set of rules that direct your study designs and procedures are known as ethical considerations in research. When gathering data from people, scientists and researchers must always abide by a set of ethical principles. The integrity of science, respect for human rights and dignity, and cooperation between science and society all depend on research ethics. These guidelines guarantee that study subjects' involvement is free, informed, and secure. You'll strike a balance between pursuing significant research objectives and employing morally upright research techniques. Whether intentional or not, it is always required to protect participants against long-term or extreme harm. The credibility of your study will also suffer if you violate research ethics because it will be difficult for others to believe the results of your work if your techniques are immoral. Even if a research hypothesis is important to society, this does not allow you to violate the dignity or human rights of your study participants. This document doesn’t publish any personal information of the organization data. All the preventive measures are taken. Even the personal data are not used. Only data used are quantity of peoples travelling i.e., their numbers not the actual information. Scientific integrity, rights of humans and their dignity is maintained in our research. There are also no legal issues as we are not violating any legal matters during the research. The data is also justified as it is extracted from government portal itself which is available in the public domain and no restrictions for research. Also, the social issues are not violated as the outcome of the research is good for the society without violating the human rights.

# **Literature Review**

There are many areas in data analysis among them predictive analysis is most significant areas of research. In the subject of data analytics, predictive algorithms are frequently acknowledged as one of the most significant areas of research (Dominguez, 2019). These findings can be employed in many different contexts, with education being one of the most crucial and being projected in a variety of ways. Findings that estimate educational achievements using various sorts of student-related data, with a focus on higher education settings, most notably universities, uses predictive analysis methodologies. Early warning systems (EWS), which are apps that can forecast potential threats like the probability of the enrolled students passing or failing from a certain course and notify stakeholders of such hazards with the aim to make any corrections in them if needed. An assessment of the status of predictive analytics for the immigration statistics of Nepal is what this literature review aims to do. The RAND Corporation suggested conducting comparable research and experiments in the EU. Scholars were invited from five organisations employed together to develop miniature for the flow of migration in European union at exact given real-time with the data available in social media. They got some results by implementing the designed miniature, but the technique for predicting EU mobility flows is still under development. Our research has a distinct area and is based on a similar model. Although the technologies employed are comparable, there may be a few slight variations. The results of the following research reveal that tactics are becoming more popular not only in the government sectors like immigration but also in the telecommunications sector and other common sectors like health, sales, gambling, stock market, etc. There are still few in-depth literature evaluations and classification schemes, despite the expanding usage of predictive analytics techniques in several industries. For filling a knowledge gap in the literature and directing future research, this study developed framework for classification. The study that has been conducted is very much associated to the computing environment. It illustrates how various computation methods are used widely in practical contexts. These predictive analytics and its outcome explain its applications and why it’s important, and how it may benefit your business. Additionally, you’ll learn about the methods and tools employed in predictive analytics, as well as examples of their application across a range of industries. The results of the research have a significant influence since they can have an impact on the entire decision-making process. Making future and judgments can be a very challenging job in any of the professional sector. The top-level management can benefit greatly from a sneak preview of what will happen or at least they can have a rough estimate. Predictive analysis process begins with the thought of data exploration. To make the good extraction of insights, analysts will use a variety of statistical approaches, processes and apparatus and the visualisation software as well. We will extract and examine information from the official website of government. Data cleansing sometimes also referred to data cleaning, next step in getting data ready for predictive modelling or analysis. These statistics will be used to forecast how many people will be around in the next few years. Using a variety of aspects, the new assumptions are determined to be superior to the other estimators for determining and forecasting need and economic gains (Michael D. Creel, 1990). Predicting the arriving persons utilising the data gathered through predictive analysis is the primary goal.

Though there has been much research around this domain, single research cannot fulfil all the requirements needed. Many data and insights can be extracted from much research. Similar research has already been done in this field, but we felt some limitations that’s why we are doing this research and study. Similarly, our research will also have some limitations. Data given in the portal has data across all the attributes such as nationality, number of people coming each month, purpose of those people to visit. Even the data is given for what locations those people are visiting. The good news is that you can tweak your measurement variables over time and contrast your analysis or prognosis with what actually transpired. You can grow closer and closer to a model that accurately captures the subtleties of your company's real-world operations through a process of continual development. You can also engage specialists and take use of their industry-specific knowledge to determine whether your model is missing any essential components. Data for the average expense people do and people from what gender and age group are also included in the data. Links between tourists and immigration are significant yet understudied in the literature on tourism demand modelling (DWYER, 2012). But we are only analysing the data like total number of people arriving, people arriving by air, people arriving via land route, increase rate by total number and average days of stay per person. There is no plan for those limitations to make them exist, but single research cannot justify doing many studies. It will have the impact of the efficiency of researcher as well and the research paper can be vague and un-understandable. So, to stick on one point and make it perfect as possible those limitations must exist. Also, this will make path for further researchers to work on. Whatever techniques researchers use, they should be aware of the underlying hypotheses and constraints of their strategy in relation to the study’s design and any inferences that might be made from the results (Adrian Kirkwood, 2013). Also because of the vagueness that can appear because of analysing all the data and short of researchers those limitations exist. Also, in the future we will have more data related, so it will be good to analyse those rather than doing everything now. This type of study related to data analysis has been done a lot in many fields but in context of government organizations like immigration in Nepal, this research doesn’t happen much. All the related stakeholder organizations depend on reports from media or private firms for their source. So, this will make them realize the importance of these types of research and motivate them to invest in such research. They can outsource these types of projects to other companies or even sponsor such research that are happening. Some organizations with sufficient budget can also even open their own department for such. The research will also help those organizations working on small scale who can’t afford to invest or buy such data reports. They can take the output of this research as a reference for their use. As it is already discussed the many organizations like government and private can use this output for their decision making. Tourism departments, immigration departments, travel agencies, airlines are examples of some organizations that can use the data for their good. Analysis and forecasting don’t need to be precise as it depends on a lot of other factors. The output is assumed to occur and happen if everything remains the same. This will also help to pave the path for other researchers and organizations and motivates. If they can see the impact and analyse the effect of such research helping other bodies, it paves a path for further research.

The technique and basics for the predictive analysis is almost for all types of research. It is not necessary and mandatory that research happening in different types of fields may or should utilize different techniques. The statistical analysis and data mining algorithms utilised in the prediction models are used to find the data patterns and trends (Mittal, 2019). Because of the flow of steps and processes most of the research looks alike because after all, all that needs to be done is the data which is in text or numerical form. So, there has been a lot of research which used those techniques. The research has been in different field which accumulated a lot of data in past decade. Various insights and techniques can be learned from those studies that can be used in our research as well. Sectors like health, banking, sales, weather forecasting have been subjected to such scenarios. There are many journals and government published papers which can be found online. I also have been studying and analysing a lot of those accessing them through portals like Ethos and google scholar. These studies have been done by many institutions like privity companies related to those business and government departments for the right to information for the public. These all the research have been useful to all of them as they are very different on its own. The statistical analysis provides the arbitrary numbers meaning, so giving the data some life (Ali, 2016). Not all that research are similar types so after studying and analysing the study they can learn the techniques related to the research which will help them to improve the efficiency for the precision of output. Predictive analytic techniques most frequently employed include linear regression. To forecast the future of the target, it leverages linear relationships between a dependent variable (the target) and one or more independent variables (the predictors). The forecast is based on the presumption that there is a causal or dependent relationship between the target and the predictors. This research lay a path and motivates other researchers as well the sponsors to conduct another research. For our own research also, those records have been very useful. We learned various techniques and methods which we can use in our own research. The flow of doing research is also learned from that already existing research. We can also learn about the importance of finding good resources from those. Without a good resource it is very difficult to start and finish any research. A clear roadmap can be drawn which also helps in making mind map. Mind map is very necessary during any research. We will have a basis idea from where to start and finish. It will also help us to keep in track because in these vast resources we can easily get off the track if we don’t track ourselves properly. This is also an important lesson we can learn from past research. Our research in comparison to the previously done research is almost similar in a sense that the motives are same to analyse the data to visualize them which eventually will be looked by the managers to take better decisions. We learned about the various techniques for the research and accuracy of all the different techniques. We also had the glimpse of which technology to use because we will know which is easy and more precise. Different technologies and different statistical approaches used in such predictive analysis will have effect on the efficiency and output. So, we can say one of the main learning from the previously done similar research is that we concluded which technologies and strategies we will use on our research.

These types of research have significantly developed overtime. This is because of its need and its importance has been felt by many stakeholders over the last decade. Data analysis is the act of discovering knowledge by studying massive amounts of data from time to time based and condensing it into critical data (Biag, 2013). Technology and communication have impacted our society in very deep-rooted way. Many people are being technology friendly and educated as well. These types of research are also developed because of the advancement of education. People are studying more of these data analytics related subjects. Also, the businesses are being globalized and booming so more of these research needs to be done which will help those businesses to progress more. More analysis friendly technology is also being evolved which will make these types of research easier than it would have been few decades ago. These types of research also need technical human manpower, funding, and its real time application. Without the application of its output after its research, it is of no worth. People were also less motivated due to these reasons as well. Due to the intense interest in and discussion surrounding the subject of job creation and firm size, size of a company is without a doubt the aspect that has been examined the most in relation to its contributions to growth (Per Davidsson, 2019). But in current present scenario these findings are evaluated and applied in real business scenario. So, it is very evident that the research has developed significantly overtime and it will even increase more because of its need and demand. Every thesis has its reader and audience related to the field in some way. People read research papers either because of its need for some information or to do the similar research. This is very well-done research using some specific statistical technique for a specific scenario. This research can be studied by the students for their knowledge as well as by the suggestions can be studied by the concerned authorities like tourism businesses and government agencies if needed. Different readers can view this paper in different way. Students can see the statistical techniques and technologies that are used in the research. They can compare why this technology was used and how is it different from other ways. Very huge datasets are usual now thanks to the invention of computers, the analysis of the accompanying data and/or the process of extracting relevant information from their vast networks is less commonplace (Diday, 2011) . While the agencies like the government organization and tourism business can analyse the output to make its business decision or strategic planning. Students can also find the gaps in this research. The gaps will help others to find the appropriate objective for the other research they will be doing. This research also gives the context of overall tourism scenario in different dates in Nepal, also how the mathematical statistics and programming technologies can be integrated for analysing data, and it paves a path for learners who aspires to do similar research.

# **Research Methodology / Description of Practical Research Work Undertaken**

The study is solely based on the data which is extracted from the official Immigration Department and Tourism Ministry’s web portal. Here we will discuss about the research philosophy and the approach we are taking during the research. We will also discuss about the strategies adopted during the research. Also, the data collection methods whether it is primary or secondary is discussed. We have used the secondary data, but it is discussed that the how the secondary data was collected and why it can be trusted.

## **Research Philosophy**

There is a limited number of theories we can adhere to when conducting any type of research. Type of research philosophy to use can depend on the type of study being conducted. I will adhere to the positivist concept in my situation. According to the positivist worldview, only data that is "factual" and derived from the actual observation (in field), including the measurement of numerical form of data, is trustworthy. In positivist studies, the researcher's role is limited to the collection of data collection and interpreting the objective of primary research. Results from these kinds of experiments are frequently audible and quantifiable. The foundation of positivism is quantifiable, measurable data. The philosophy of positivism, according to one source, agrees with the empiricist idea that knowledge derives from human experience. It perceives the world as being made up of distinct, observable things and things that happen, and it sees those things interacting in expected, repetitive, and periodic ways. The sort of research being conducted, and the researcher's style also influence the research approach. I'll use a deductive strategy for this investigation. Testing the current idea is the main goal of the deductive approach. According to the literature on consumer research, "positivistic social science" places a strong emphasis on the search for casual connections or casual explanations (Hunt, 1991). We cannot use the deductive method if we don't have a theory. We'll employ a quantitative research approach. The information comes from a secondary source and is presented numerically. When such a specific phenomenon occurs, we can investigate. Huge amounts of our data are shown numerically. Quantitative data is required because we are creating a predictive model. The technology produces more accurate results the more data you feed it. Simply said, conducting studies on such many individuals makes action research, experiments, and other strategies useless.

## **Research Approach**

Statistical analysis is frequently used in quantitative research to link what is already known to what might be discovered through investigation. As a result, knowing the relationships between variables using any inductive or descriptive statistics is necessary for evaluating data using quantitative methodologies. Estimating the parameters and drawing conclusions about populations are both aided by descriptive statistics. A method for testing objective hypotheses by looking at the relationship between variables is quantitative research. To enable statistical analysis of numbered data, these variables can be measured, often using instruments. The intro, literature and theories, methodology, results, and commentary make up the predetermined format of the final written report. We have also followed the same approach in our research. The accuracy of the model developed depends on the quantity of the data feed. The more data is fed, the more accurate the model will be. Like qualitative researchers, people who use this method of inquiry make assumptions about the ability to generalise and repeat the results, adjust for possible explanations, and deductively test theories. Deductive reasoning or deduction is the basis of quantitative research, which employs a wide range of quantitative analysis techniques, from simple descriptive of the variables to advanced statistical modelling to demonstrate statistical correlations between the variables (Khalid, 2012). So, this research was conducted in phases i.e., Collecting, cleaning, and analysing the data from the source, developed and trained a model using the linear regression and decision tree algorithms to create predictive models, finally testing the model using the untrained data to find the accuracy of the model.

In the research the data used are either the primary or the secondary data or may be the other resources. Primary data are the data which are collected by the researcher themselves first hand. Various surveys are done regarding the necessary data. Data can be collected by questionnaires or various other methods like observations, experiments, and interview. These types of data are the real time data and researcher are very much involved in it. This can also be time consuming and cost inefficient as compared to the secondary data collection. But the data collected are very specific to the need of the researcher and researcher will get the high quality of data they need. Whereas the secondary data is the data which is readymade and already collected by someone. These can be extracted from various sources like government website, publications, books and so on. The data may not be spot on as required for the researcher, but data can be handled and make it accordingly by cleaning or various other processes. In our case the data was collected from the government website. The data are genuine and accurate. Data here is collected in the ports of entry. Ports of entry can be land ports or airports. When people arrive in certain port they must pass through immigration. So, when they pass through immigration desk their passports need to be scanned, that’s when the data is recorded. In our data extracted there were many data including the purpose of visit, time of stay, month when these people arrived. The data was also categorized by which nationality, purpose, sex, places where they are visiting and various other as well. But we extracted only the data which we needed that was the total number of people arriving by specific year, their average stay and rate of increase of incoming people.

## **Discussion of practical work done**

Practical work can be explained in step by step. There are various steps that were involved while doing the practical work and analysis. First the data was collected and from that data we have extracted the data which were needed for this research. We have explained about the possible techniques for the predictive analysis. Out of those techniques we choose two of them for our analysis i.e., linear regression and decision tree method. In our data we first find out all the correlation coefficient of all the variables through the correlation coefficient table. Through that we can observe the relations of the variables with each other. Through the observation we see have a basis idea some variables have direct relation with some other variables whereas some are negatively correlated. Year and the total number have positive correlation as well as with the average day of stay. Correlation coefficient varies from 0 to 1 i.e., higher value means higher correlation. There is also a negative correlation between year and growth rate. We can observe that in our graph as well. Then we proceeded to our analysis part.

Firstly, we did the analysis through the linear regression method. First the data needed to be cleaned and there were also some missing data which needs to be removed or filled with other suitable data. So, we have considered some outliers in our dataset and few missing data also needs to be filled. Last two rows of the data are considered as outliers as they are highly varied from other data. Those data are more than quarter more, so we removed those. And in the variable average stay first few cells are empty but we can fill them with average value as the value is constant throughout the year from start to finish. With the variable Total Number, we will visualise the linear regression line (TN). Finding the intercept and slope is the major goal. By eliminating the final two rows that we deemed to be outliers, we will produce a new data frame, designated df3. Then, we'll add a new column called "Numbers" where we'll list values from 0 to the data frame's length. The values of column "Numbers" will now be kept in an array called "X." The variable Total Number(TN), on the other hand, will be kept in a variable called y as an array. lin model = LinearRegression will then be used to generate the linear model that will be used to train our data set (). The intercept and slope will be determined using the function fit(X,y). The procedure described above is now used to visualise data in a linear format. The anticipated total numbers from the model will be obtained, and they will be kept in the new variable designated as y prediction. Then, a new column called Prediction is created and the predicted value is saved there. Using the plot() function, we will now plot the data for the graphical representation. The chosen title for the graph is appropriate. The data will then be displayed on a graph using the function plot (). The red line above is a graph that is plotted for the year and the Total Number(TN) variable, and the blue line is a linear line. We can see that the relationship is linear, and it also suggests that as time goes on, more individuals are arriving overall. The R2 score, also known as the coefficient of determination, will now be computed. The effectiveness of our linear model is evaluated using this. This number represents how much the independent variable, which is the input, can predict the fluctuation in the output, or dependent variable. This allows us to assess how well the outcomes are. That’s how we can find the accuracy of the model. Therefore, we utilize the linear regression formula, adding one unit to the length of the value of the data frame, to get the prediction value for the following year. So, we use the same process for other variables and find out the predicted value of the next year.

Then we did the analysis by decision tree. We did the same as in the linear regression to remove the outliers and filled the empty cells. Then, In order to forecast our final two numbers for the variable Total number, we established a new variable called Prediction. As we attempt to forecast for the next two years, we will move the values in the new column by two values to the top side. The remaining five data are then displayed using the tail() method, and the last two cells are clearly empty. The next step is to create a new dataset called X, convert it to array format with numpy, and remove the previous year's value, which is represented by the future years variable's value of two. With the exception of the x years, which are future years, we now build a new data set called y and convert it as well to a numpy array to obtain all the target values required. We have now divided the data into training and testing sets. We set aside 75% of the data for the model's training purposes and 25% for testing. The image for the upcoming challenge is seen above. We employ the function train test split. The data can also be fixed for training purposes. Now, with the function fit(), we develop a decision tree model. Now that we have the x years of the future data sets, we can store them in the array seen in the above image. displaying the expected values at the end.

## **Discussed ideas that were rejected**

A lot of ideas and insights could have been gained from the data which were originally extracted from the site. There were probably many variants of data which could be used to gain the insights, but we only took the specific data for the specific objective in this research. In our main dataset which was extracted the total number of people arriving was given month wise. That was the idea under consideration but after critical review we rejected the idea for predicting number for each month. This was very much possible to do but there are many reasons to reject the idea. Main reason is that it was not feasible to find numbers per month. When the strategies are made it is planned for the yearly budget and yearly income. So, predicting monthly wise was not so reasonable in this case. Also, because the numbers when divided per month are too less as compared to year. If the people arriving monthly was very high around at least half million that would have been considered. Another idea we discussed but rejected in the end was to predict the destinations people would visit. It would have been a good idea to predict the number of people that would visit the specific area, but the data was not sufficient for this analysis. Data were present for only for around a decade and it would be very inefficient model. The accuracy of the model would be very less as the total number of people was also so varying every year. It was also discussed to take the data regarding the income generated by those tourists. But this thing was kept under future scope and recommendations. More research is required regarding the financial analysis. So, these are few ideas that are discussed but were rejected.

# **Results and Analysis**

The data is taken from a report of tourism ministry published in 2019 before covid. The data consists of various charts and analysis regarding the arrival and departure of passengers as various categories. All the data are represented in tabular format in the report. The data is also publicly available for all, and anybody can access the data anytime for any use under the constitutional act called RTI i.e., right to information. We took the data which gives the information about the total number of people coming in a tourist visa year wise from 1964 to 2019. Also, the data is further divided into two as Total number by land and air because there are two routes from where people can arrive. There were also various categorial data like arrivals in each month, arrivals by nationality, arrival for what purpose but due to the vagueness of the research we excluded the data. However, the other data like growth rate and average days of stay are taken for consideration. The data we are using is a clean data. There is no fear and worry for the accuracy of the data as this is collected by the government body itself. This data is also verified and approved by various other departments before publishing. In many research data samples are taken on guess for the analysis but in this research the data is 100% accurate data. Though there may be few errors as we are dealing with number of people, it is assumed this is so low that it can be rejected as all the data are collected digitally by scanning the passports.

## **Predictive analysis techniques**

There are various ways by which the predictive analysis can be done, and predictive models can be made. Some of the used techniques are described below:

**Neural Network**: Neural networks were discovered in a concept that the predictive modelling can be done by imitating a human brain. Neurons and other parts of brain are really imitated, and the process of human brain working is also made. Neural Networks can be applied to the problem of optimal signal processing (Yu Hen Hu, 2002). The model can do tasks like finding relations between complex data and recognizing patterns. In the domains of AI, machine learning, and deep learning, neural networks enable computer programmes to identify patterns and resolve common issues by mimicking the behaviour of the human brain. Training data is essential for neural networks to develop and enhance their accuracy over time. However, these learning algorithms become effective tools in computer science and artificial intelligence once they are adjusted for accuracy, enabling us to quickly classify and cluster data. It can be used for huge dataset or in a condition it is difficult to recognize which formula to use for the analysis. It is suitable for situations where we need just only predictions rather than the theoretical explanations. The design of the neural network is useful because it simulates in the form of human brain’s patterns during the decision making. This is mostly used for the recognition of image and image processing related projects. They have layers mostly categorized in three i.e., input layer, hidden layer, and output layer.

Diagram, schematic, bubble chart

Description automatically generated

Fig 1: Simplified Neural Network

**Time Series**: This model focuses on capturing data from different point. So, this is basically series of data where time is the input. A numeric metric is developed and by using last year’s data it can predict the data for next few days or weeks. It is also used in various sectors like small scale businesses and medical sectors to analyse the time for occurrence of symptoms of any disease. It is also an important way to understand a situation where the prediction can be done over time. Same model with short time of data is less accurate than the model developed over years feeding more data. For example, if a small business owner wants to predict how many customers will arrive tomorrow, they can take average for past 6 month, so this is a simple way to predict customer. Some characteristics of nonlinear time series models are challenging to extract analytically, but by creating fictitious series and using nonparametric techniques, accurate numerical approximations may be produced (Robinso, 1983). However, the growth is not always static or a linear sometimes it can be also exponential. So, this model is better for the exponential growing data.

Graphical user interface, text, application

Description automatically generated

Fig 2: Time series model

**Clustering**: This model depends on the similar qualities of data, the clustering model divides the input data into distinct, layered agile groupings. For example, if an online shoe retailer wants to conduct focused marketing campaigning for its clients, it can search many records develop individualised plans for each customer. Market segmentation has long relied on cluster analysis since it improves the effectiveness and efficiency of a company's marketing efforts (Higuchi, 2021). Is it better spending time doing that? Probably not. They may swiftly classify clients into similar groups based on shared traits using this clustering method, and then develop tactics for individual assembly on a broader scale. This method can also be used in banks. For example, this predictive modelling method may also be used to classify applications for loan into a bucket system. It is also efficient to distinguish criminal activity group of people and so on. It is searching for fixed numbers of clusters in a dataset. All the data points must be allocated to some cluster by decreasing the sum of squares.

Chart, scatter chart

Description automatically generated

Fig 3: Clustering

**Regression models**: One of the most important techniques for predictive modelling is regression modelling. Here the correlation between the variables is found to make a linear expression. This expression helps to find the strength between these variables. A constant is added in the formula due to which it can represented as a function. A regression line can also be drawn to visualize the scenario graphically. For example, analysing the house price can be done using the linear regression model. One of the most crucial statistical and data mining methods used by statisticians and researchers for the study and classification of binary and proportional response datasets is logistic regression (LR) (Maalouf, 2011). Correlation can be shown between the prices of the house and interest rates. A future prediction for house price can be done by given interest rate. The variable which is being predicted is called as the dependent variable whereas the reference variable taken is called as independent variable.

Chart, scatter chart

Description automatically generated

Fig 4: Linear regression

**Decision Tree**: Decision tree can be termed as the graphical representation for the choices that we can make during the decision-making process. This can be the best technique if we want to know what factors leads to some decisions. On this method it places data to different sections analysing certain variables. There are individual branches and leaves in the decision tree made. Those branches denote the decisions and choices made. They are very easiest to understand so they are one of the simplest techniques in predictive analysis. A decision support system called a decision tree employs a tree-like graph to represent decisions and their potential outcomes, such as utility costs, chance event outcomes, and resource costs (Dr. Neeraj Bhargava, 2013). They can be really handful when the decision must be made in short time. The branches are called the nodes and they represent the outcomes of the decisions that are possible. These are the tools that must be used as reference for making decision but not the actual solution.

Diagram

Description automatically generated

Fig 5: Decision Tree example

## **Why use Linear Regression and Decision tree?**

There are many ways and algorithms we can use in the predictive analysis but in our research, we have used the linear regression and decision tree method for the prediction. Each and all methods have their own pros and cons for their use. Regression analysis is best suited to understand the relationships between the variables. By using the regression analysis like r-squared we can check how much data in total is explained by the model we made. It is like the accuracy. It is also comparatively easier in regression model to find out the features which are good predictors. For example, if we have 20 functions, we can find out which function suits best for which variable. Also, the coefficient for each variable can be estimated as well as the level of confidence for them. This model is very interpretable as well with being simple. The related business stakeholders will be happy with the proposition as all the businesses are very much interested in what logic is being implemented in the model and this model is comparatively easy to understand.

Also, the decision tree is used in our analysis. This is a very good way because it is easier to make the visual representation. It is also easy for the data set with non-numeric values though in our case it is only numerical data. There is less amount of data cleaning needed for this method. We can also make different scenarios for this like best, worst, or likely. It can also be combined with other decision-making processes. It takes less effort in the decision tree for preparation of the data however it is a necessity to have a ready-made information for creating the variables. It also requires less data cleaning than any other techniques in decision tree.

## **Using Linear Regression**

On this part the data analysis part using the linear regression is shown with all the codes, analysis, and visualization.

Text

Description automatically generated

Fig 6: Importing libraries

Various libraries from python needs to be accessed to perform our analysis. We used NumPy, pandas, scikit-learn and matplotlib in our analysis. NumPy is one of the most important libraries in python. It is used while dealing with array data sets. It also offers various functions to work with linear algebra, matrices, and Fourier transform. And pandas are used for analysing, clean, explore and to manipulate the data. These processes make the data readable and clean. Scikit-learn also is a library that has various algorithms like random forest, vector machine etc. And matplotlib is a visualization utility used for the visualization.

A picture containing table

Description automatically generated

Fig 7: loading data

In the above picture we can see we loaded the data in data frame i.e., df using read\_csv function. We also visualized the top 5 data using the head() function. As we can see there are some empty cells in the function. Those empty cells need to be treated before analysing the data. Either they are removed or filled with some data.

Table

Description automatically generated

Fig 8

We have some empty cells in our one column i.e., average stay. This column data is not a changing data. The values are same throughout the year. So, it can be filled with the average value in the empty cells. As we can observe the last two data in our data set is too much different from other data so it can cause value problems, so we consider it as outlier. We also didn’t consider the data during the covid period as the number of arriving tourist were almost negligible which is around 5% of normal period. We considered those ones also outliers as well. So, we will remove this for our analysis. And then we have a new data frame that don’t have any empty cells and outliers.

### **Correlation coefficient**

A picture containing table

Description automatically generated

Fig 9: correlation table

Correlation coefficient indicates the strength between the two different entities in terms of the relationship which is linear. Greater than zero value of correlation coefficient indicates a relationship which is positive that means the increase in value of one variable also leads to the increase in the value of another variable. In our case we have the correlation coefficient table above which show the coefficient values between different variables i.e., year, growth rate, Total Number(TN), TN by Air, TN by Land and Average stay(days). Correlation coefficient of each of the variable with another variable can be found with the table. Correlation coefficient is a very difficult calculation to find so we use the function corr() to find it. The formula to find out the correlation coefficient is given below.

Text

Description automatically generated

r = correlation coefficient

**xi** = values of the x variable

**x̄ = mean of the x variables**

**yi** = values of y variable

ȳ = mean of y variables

As we can see variable year has a positive correlation between Total number, TN by land, TN by Air, and average day of stay. The values closer to 1 mean they have a direct relation between them. One variable increase significantly as the other variable increase. Correlation coefficient of the year and average stay day is 0.075 that means the value is close to zero which means they are not much correlated. We can also see in the data that the average stay day doesn’t increase like other variables like Total Number. We can see that in our graph as well which will be discussed further. As we can also observe the coefficient between the growth rate and Total number is negative. It is because of having a lot of negative growth of people in many years. We have also visualized the linear relation between these variables where the slope is going downwards.

### **For Total Number(TN)**

Text, letter

Description automatically generated

Fig 10

We will plot the linear regression line with the variable Total Number(TN). The main objective is to find the intercept and slope. We will create a new data frame i.e., df3 by removing the last two rows which we considered as outliers. Then we will create a new column ‘Numbers’ where we will list the values ranging from 0 to the length of the data frame. We will now store the values of column ‘Numbers’ in an array name as X. Then again the variable Total Number(TN) will be stored in a variable called y as an array. And then the linear model to train our data set will be created as lin\_model = LinearRegression().fit(X,y) which will be used to find the intercept and the slope. These values will be used at the end to predict the value for the next year.

Text

Description automatically generated

Fig 11: preparing for data visualization

Now the above process is for the visualization of the data in linear form. We will get the predicted total numbers from the above model, and we will store them in the new variable called as y\_prediction. Then the predicted value is stored in a new column called as Prediction. Now we will plot the data for the graphical representation using the plot() function. Title for graph is selected as suitable.

Chart, line chart

Description automatically generated

Fig 12

We can see the graph plotted for the year and Total Number(TN) variable in the red line above and the blue line is the linear line. We can see the relation is linear and it also gives idea that as year passes by the total number of people arriving has also increased.

Now we will calculate the R2 score which is also called as the coefficient of determination. This is used to find how well our linear model performs. This value is the amount of variation in the output or the dependent variable which is predicted from the independent variable which is the input. By this we can check how well the results are which are obtained from the model.

Graphical user interface, application, Word

Description automatically generated

Fig 13: coefficient of determination

This also means how well the model we have made explains the data which are observed. For an example 60% value of r-squared means that the 60% of the observed input is explained by our model. In our case for the Total Number(TN) the coefficient of determination is 87.9% which is very good. It can also be described and termed as the accuracy.

Graphical user interface, text, application, Word

Description automatically generated

Fig 14: predicted value

So, for the next year prediction value we use the linear regression formula adding one unit to the length of the value of the data frame.

**For TN by Air**

**Text, letter

Description automatically generated**

Fig 15

We will plot the linear regression line with the variable TN by Air. The main objective here is also to find the intercept and slope. We will create a new data frame i.e., df3 by removing the last two rows which we considered as outliers. Then we will create a new column ‘Numbers’ where we will list the values ranging from 0 to the length of the data frame. We will now store the values of column ‘Numbers’ in an array name as X. Then again the variable TN by Air will be stored in a variable called y as an array. And then the linear model to train our data set will be created as lin\_model = LinearRegression().fit(X,y) which will be used to find the intercept and the slope. These values will be used at the end to predict the value for the next year. Similar process for all the variables for all the calculations.

Text

Description automatically generated

Fig 16

Now the above process is for the visualization of the data in linear form. We will get the predicted TN by Air from the above model, and we will store them in the new variable called as y\_prediction. Then the predicted value is stored in a new column called as Prediction. Now we will plot the data for the graphical representation using the plot() function. Title for graph is selected as suitable.

Chart, line chart

Description automatically generated

Fig 17

We can see the graph plotted for the year and TN by Air variable in the red line above and the blue line is the linear line. We can see the relation is linear and it also gives idea that as year passes by the total number of people arriving by air has also increased.

Graphical user interface, text, application, email

Description automatically generated

Fig 18

For this variable 86.9% of the data is justified by this model. Or we can say this is the accuracy for this variable using this model.

### **For TN by Land**

Text

Description automatically generated

Fig 19

Chart, line chart

Description automatically generated

Fig 20

Graphical user interface, text, application

Description automatically generated

Fig 21

### **By Average stay(days)**

Text

Description automatically generated

Fig 22

Chart, line chart

Description automatically generated

Fig 23

As we can see in the graph above the line is almost perfect straight. We can also observe in the correlation coefficient table that the coefficient between the average stay and year was almost 0.

### **By growth rate**

Graphical user interface, text

Description automatically generated

Fig 24

Chart, line chart

Description automatically generated

Fig 25

As we can observe in the above picture the line is going down. From the correlation table also, we can observe that the correlation coefficient was negative between year and growth rate.

## **Using Decision Tree**

Here in this part, we will show codes, analysis and visualization using the decision tree algorithm.

Text

Description automatically generated

Fig 26

To complete our investigation, we must access several Python libraries. In our analysis, we used NumPy, pandas, scikit-learn, and matplotlib. One of the most crucial Python libraries is NumPy. When working with array data sets, it is utilised. Additionally, it provides a few functions for working with matrices, Fourier transform, and linear algebra. Additionally, pandas are used to analyse, purify, explore, and change data. These procedures clean up and make the data legible. There are several other algorithms in the Scikit-learn library, including random forests and vector machines. And the visualisation tool used for the visualisation is called matplotlib.

Table

Description automatically generated with low confidence

Fig 27

We now loaded the data in the data frame using the function read\_csv(). Also we can see the top five data are displayed using the function head(). It can be observed that there are some empty cells which needs to be taken care of before analysing the data. They can be removed or filled with some data. In our case we will fill it with average value as it is the average day which is almost same every year. Deleting entire row will also delete other data variables which can cause less accuracy in the developed model.

A picture containing text

Description automatically generated

Fig 28

One of our column’s variables, which is the average stay, there are a few empty cells. This column's data doesn't change. The values remain the same all year round. Therefore, the empty cells can be filled with the average value. As we can see, the last two data in our data set are too drastically different from the other data, which can lead to value issues. As a result, we classify them as outliers. Additionally, we excluded the data from the COVID period because there were only a very small number of tourists arriving—about 5% less than during the usual period. These were also considered as outliers by us. We then obtain a new data frame that is free of outliers and empty cells.

Chart, line chart

Description automatically generated

Fig 29: visualizing the Total Number(TN)

We can visualize the data here by keeping year in x-axis and total number in y-axis. We can see it is an increasing trend that the total number of tourists are increasing every year.

Graphical user interface, text

Description automatically generated with medium confidence

Fig 30

Here we create a new variable called Prediction to predict our last two number for variable Total number. We will shift the new column values by two values upper side as we are trying to predict for the next two years. The data then is visualized using tail() function to see he last five data and we can see the last two cells are empty.

Graphical user interface, application

Description automatically generated

Fig 31

We now create a new dataset called X and convert it into array format using numpy and we also remove the last years value which is future\_years varible i.e., two.

Text

Description automatically generated

Fig 32

Now we create a new data set named as y and convert that also to numpy array to get all the needed target values except the x years which is future years.

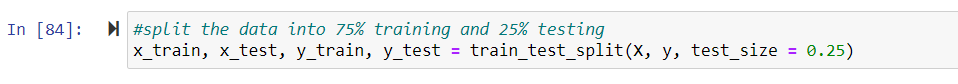


Fig 33

Now we split the data in training set and testing set. We allocate the 75% of data for training purpose of the model and 25% for testing purpose. Above is the picture for the following task. We use train\_test\_split function. We can also fix the data for training purpose.

Graphical user interface, text, application

Description automatically generated

Fig 34

Now we create a decision tree model using the function fit().

Graphical user interface, text

Description automatically generated

Fig 35

Now we get the x years of the future data sets and store them in array which is displayed in the above picture.

Graphical user interface, text, application

Description automatically generated

Fig 36

Finally displaying the predicted values.

# **Critical Evaluation, Conclusions and Recommendations**

The main objective of the research has been obtained and the research questions also have been answered. The research has been done in different steps from researching to data manipulating to analysing. All the steps have been done according to the plan and it is tried to justify according to the objective and aims. We explained about the steps and all the plan of our research in the beginning. Explanation of the techniques and technologies used are also explained properly in the paper. Research could have been done in different way in terms of the techniques and tools used which could have resulted in different outcome and ways of doing things. But we also explained about the selection of the tools and processes with its advantages and reasons to choose it. Different methods would have led to different results in terms of accuracy, but we must choose the best method depending on the type of research and the accuracy also needs to be taken care of. All the steps of the research process have been explained and described from data obtaining to cleaning to filling up the empty cells to removing the outliers. Evaluation of research paper can be a highly political activity (Scott Reeves, 2015). All the screenshots of the process involved are also kept in the repost. The code part and the calculations part also kept making it clear about the process of research. All those are also described accordingly with their steps and reasons for using them. There could be many cases of analysis and predictions for our data set. In our first process using the regression method we have predicted all the variables like total number, total number by air and land, average stay and by the growth rate. But in the case where we used the decision tree we only evaluated for the variable total number. This is more about the process rather than evaluating all the available variables. We also found the accuracy of the model designed by the regression method which basically gives how much percentage of data is justified by the model. And in case of the decision tree model, we split the data and used three quarter as training model while other remaining one quarter as a testing data.

## **Contributions to wider academic body of knowledge**

The finding contributes to the area of knowledge as well as its applicable fields. There are many areas where the finds of this research can be applied. The findings apply to the educational area as well as the business and organizational community. The research obviously lays a path in this sector. It gives a rough idea on what techniques and tools are useful for what types of research. Not only the research but also the understanding about the type and amount of data that is being used in the research is equally important. This also makes a base for the future researchers and motivate them to undertake another research. We also came to know about the flaws of the way of designing such models. It can be observed that the models are not fully accurate. There are some flaws which can be because of many reasons. It can be because of data, or the data processing techniques, or the impurities of data or many other reasons. So, we can always find a better way to make the model more accurate. It should be tried to justify more amount of data from the derived model. Also, the visualization process can be made better using various other business intelligence tools as well. It is always a better way to visualize using such tools as there are multiple options to express the data. Academic achievement inequalities between children and schools of various socioeconomic level may be narrowed, according to educational scholars and practitioners, if supportive school and classroom climates are present (Moore, 2016). It serves for the junior research in terms of guiding them and giving idea on how to carry out some similar types of research. They can learn in what sense the research is important and what are the actual reasons to carry out research. If the researcher knows why they are carrying out the research and on what field their research will be helpful they will spend enough time on planning how to carry out the research. This will help them find the correct techniques and tools which they can use during their research. This also helps in the more accurate and error free research to publish. So, this findings from this research have some significant contribution to academic body of knowledge for both students, research, and institutions.

This research can also be compared with another similar research done previously. Yes, there are much similar research done previously with similar types of data and techniques used. Every research has its own objective and aims. Even if the techniques can be same and the data being analysed can be of similar types, and the tools can be similar it is not necessary the research work cannot be compared. There could be many other things on which they can be compared. They can be compared on their different objective they serve or with the scope and research questions. Yes, there were similar research done previously with the same context of immigration but not from Nepal. We did this for the first-time using Nepal’s data. And the data taken is also related only to travellers getting tourist visa. Previously done research were focused on analysing the historical trend rather than the predicting for the future. Our research is based on the predictive analysis model which we are predicting the numbers for coming few years. Our research is based on conclusion that the insights can be used by many institutions such as government as well as private business entities. So, the result from our research is different from previous research is in a sense that we have done research on numerical based predicting the values i.e., number of people arriving whereas in most of the other research it is a conclusion based giving some recommendations what steps to do for its improvements. The findings of the research carried out has many strengths as well as it has some drawbacks. It is a good thing to know about the assumption of numeric prediction happening on next unit of time. In our case we predicted the number of people coming the next year. By using the same model, we can predict not only for a year but for few years. The model is also very accurate in a sense that it satisfies more than 80% of the data, that is what the main strength of the finding is about. Not only for the total number but also other variables can be predicted using the same method. Though there are some limitations of the model. No model is cent percent accurate. This model can be used for the assumption and the data can be used for the analysis and make plans accordingly for the future. But believing that this will happen for sure will be a mistake. Various external factors are not accounted during the analysis. This output is supposed to be happening if the other factors remain the same. So, this research cannot be justified as applicable in all scenarios. Recently the covid pandemic hit and there were significantly low people arriving. So, if the plans were made according to the analysis, it would be a different thing because covid couldn’t be predicted. But these events are so rare so we should focus on the good parts. It is very useful for the planning regarding the tourism industry, it can also be useful for the institutions doing business related to tourism. This is also useful for other researchers as it paves path for them in a sense that it a gives idea on various terms, tools, and technologies useful for similar types of research.

Our main objective was to build a predictive model and find out the predicted number of people for next few years. We been successful on our objective. Hence there could be many amendments that could be made to make our model more accurate. Data cleaning process could be more efficient and different technique could have been chosen. But we can say our objective has been achieved. The research question was also to find out how the predictive analysis will help in forecasting the data and how will it influence in the decision-making process. This fulfilment of the main aim also answers these questions. Predictive analysis is obviously useful for forecasting the immigration data. We used the suitable statistical technique i.e., regression model and decision tree for forecasting. There is no doubt that the prediction made by the model is helpful in decision making. Plans and preparations can be done accordingly. Budgeting and strategies can also be discussed according to the prediction by the model is the accuracy is satisfactory enough. More research could have been done and extended from same data. Different more techniques can be used for more accuracy and data could be extracted for not only next two years but also for few more years. But overall, the achievement has been satisfactory. Any research, regardless of how beneficial to society it may be, cannot be fully justified if it violates people's rights. All ethical issues were appropriately considered during this research's data collecting and processing phases. There are certain principles and limitations that we should follow doing any work, this applies in the research also. There may not be any specific definition of being ethical, but it is a thing of general understanding. Not any sentiment of a personal being or certain group of people or ideology has been disturbed. This in fact works in favour of them. By putting into practise the necessary ethical standards, all research investigations should place a high priority on protecting human participants. There are no personally identifiable company data included in this document. The whole range of precautions is used. Not even the personal information is used. The data utilised here is the quantity of travellers, but not the personal and actual data. In our research, we uphold the principles of science, human rights, and human dignity. Additionally, there aren't any legal problems because the research is done without breaking any laws. Additionally justified by the data. All the legal regulations are also followed properly. Data extraction was also legal as it was publicly available and can be used by anyone for information or research. Though there are not any specific legal requirements made for this specific research, but we have fulfilled all the generic rules that fits all the research and studies. A lot of learnings can be made from this research. Though our objective has been met and research questions has been answered there are still a lot that could be done from this research. There are still things that can be researched more for more insights. We can use various other techniques and methods to observe the same output. By this we can find how accurate is the model. We can also compare the accuracy and techniques by using different methods to find the best way for this type of research. We can use different tools for the coding and visualization. It gives the bright idea and sense of which tool is best for this type of research. We can also do this research to find out some more research questions. Analysing in terms of finding in helpful for Immigration and decision-making process is good but it can also be used in perspective of business. The analysing variables can be set suitable for the business institutions. Results and analysis can be more of philosophical approach rather than a practical approach for the businesses. It is also recommended to collect more data regarding the number of people for the more accuracy of the model.

As study was carried forward, we gained some additional knowledge and additional discoveries were made regarding the importance of such research works. We came to know how this not only helps in decision making but also in various other sectors. It can help in the crime investigations as well which can lead to preventing crimes.

# **References**

Adrian Kirkwood, L. P., 2013. Examining some assumptions and limitations of research on the effects of emerging technologies for teaching and learning in higher education. *British Journal of Educational Technology,* 44(4), pp. 536-543.

Ali, Z., 2016. Basic statistical tools in research and data analysis. *Indian Journal of Anesthesia,* p. 662–669.

Arifin, S. R. M., 2018. Ethical Considerations in Qualitative Study. *International Journal of care scholars,* 1(2).

Biag, M. M. A., 2013. Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis. *IOSR Journal of Computer Engineering,* 10(6), pp. 01-06.

Brooker, R., 2010. Case study in the contemporary world of research: Using notions of purpose, place, process and product to develop some principles for practice. *International Journal of Social Research Methodology,* 3(1), pp. 49-61.

Diday, L. B. &. E., 2011. From the Statistics of Data to the Statistics of Knowledge. *Journal of American Statistical Association,* pp. 470-487.

Dominguez, L., 2019. Systematic Literature Review of Predictive Analysis Tools in Higher Education. *Journal of Applied Science,* pp. 1-26.

Dr. Neeraj Bhargava, G. S., 2013. Decision Tree Analysis on J48 Algorithm for Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering,* 3(6).

DWYER, N. S. &. L., 2012. Immigration and Tourism Demand in Australia: A Panel Data Analysis. *An Internatinal Journal of Tourism and Hospitality Research.*

Higuchi, A., 2021. A factor-cluster analysis profile of consumers. *Journal of Business Research,* Volume 123, pp. 70-78.

Hunt, S., 1991. Positivism and Paradigm Dominance in Consumer Research: Toward Critical Pluralism and Rapprochement. *Journal of Consumer Research.*

Karimi, M., 2016. Modeling, analysis and multi-objective optimization of twist extrusion process using predictive models and meta-heuristic approaches, based on finite element results. *Journal of Intelligent Manufacturing,* p. 463–473.

Khalid, K., 2012. GET ALONG WITH QUANTITATIVE RESEARCH PROCESS. *International Journal of Research in Management,* 2(2).

Maalouf, M., 2011. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies,* 3(3), pp. 281-296.

Michael D. Creel, J. B. L., 1990. Theoretical and Empirical Advantages of Truncated Count Data Estimators for Analysis of Deer Hunting in California. *American Journal of Agricultural Economics,* 72(2), pp. 434-441.

Mittal, S., 2019. An Insight into Machine Learning Techniques for Predictive Analysis And Feature Selection. *International Journal of Innovative Technology and Exploring Engineering,* 8(9s).

Moore, H., 2016. A Research Synthesis of the Associations Between Socioeconomic Background, Inequality, School Climate, and Academic Achievement. *Review of Educaitonal Research,* 87(2).

Per Davidsson, B. K., 2019. Empirical Analysis of Business Growth Factors Using Swedish Data. *Journal of small business management,* pp. 332-349.

Robinso, P. M., 1983. NONPARAMETRIC ESTIMATORS FOR TIME SERIES. *Journal of Time Series Analysis,* 4(3), pp. 185-207.

Scott Reeves, S. B., 2015. Interprofessional Education and Practice Guide No. 3: Evaluating interprofessional education. *Journal of Interprofessional Care,* pp. 305-312.

Yu Hen Hu, J.-N. H. a. S. W. P., 2002. Handbook of Neural Network Signal Processing. *The Journal of the Acoustical Society of America.*

# **Appendices**

