Bachelor’s thesis title

Information and Communication Technology

2022

Mani Bhattarai

USE OF SUPERVISED MACHINE LEARNING ALGORITHMS FOR PREDICTIONS OF VALUES THEORETICALLY

TURKU AMK
TURKU UNIVERSITY OF
APPLIED SCIENCES

Bachelor’s Thesis | Abstract

Turku University of Applied Sciences

Information and Communication Technology

2022 | 28

Mani Bhattarai

USE OF SUPERVISED MACHINE LEARNING ALGORITHMS FOR PREDICTIONS OF VALUES THEORETICALLY

The work has discussed the use of machine learning algorithms in the development of automated models such that they will be able to provide a great deal of assistance to humans and remove the dependency on the human being. Among the different machine learning algorithms present this work has focused on the supervised machine learning algorithm, for the process of data modeling and implementation in a real-world application. Different stages that would be required for the process of the model development, as well as the implementation, have been discussed in this report. With this work, determination of usefulness with supervised machine learning algorithm has tried to be achieved. For this, using the theoretical approach to make predictions and determination of the values has been performed with supervised algorithms.

Keywords:

Machine learning, supervised learning, Data Science, Visualization, and automated models.

Content

List of abbreviations 6

1 Introduction 7

1.1 Scope 8

1.2 Limitation 9

1.3 Resources to collect information 10

1.4 Literature review 10

2 Data Science and tools 12

2.1 The role of Data Science 12

2.2 Python 14

2.3 Numpy 15

2.4 Matplotlib 15

2.5 Seaborn 16

2.6 Pandas 16

2.7 Python for Data Science 17

3 Machine Learning 18

3.1 Supervised Machine Learning 18

3.1.1 Linear Regression 19

3.1.2 Logistic Regression 20

3.1.3 Decision Tree 21

3.1.4 Random Forest 23

4 Method 24

4.1 Capturing of Data 24

4.1.1 Data Entry 25

4.1.2 Data Extraction 25

4.2 Maintaining of Data 26

4.2.1 Data Cleaning 26

4.2.2 Data Architecture 26

4.3 Processing of Data 27

4.3.1 Data Mining 27

4.3.2 Data Predictive Analysis 28

4.3.3 Qualitative Analysis 29

4.4 Data Reporting 29

4.5 Data Visualization 30

4.6 Decision Making 30

5 Conclusion 31

References 34

List of abbreviations

CART	Classification And Regression Trees
ID3	Iterative Dichomister 3
ML	Machine Learning

1 Introduction

Machine learning algorithms can be used to provide services related to model development for business organizations and help in their service delivery. The digitalization of businesses and services has been responsible for generating a larger amount of data in the last decade and for this reason machine learning algorithms have been used to handle those data and develop suitable systems. (Akinsola, 2017)Using traditional methods for the handling of data was possible only with the volume of such data being low, however, the volume of those data has increased rapidly with digitization. This is caused as data have been coming larger number of sources, and handling them is nearly impossible. Such a huge volume of data, often called big data, will need special technologies to handle them (Wang, Yang, Wang, & Sherratt, 2020) and among these technologies, using machine learning is one of the widely used. New technologies such as machine learning handle those larger volumes of data accurately and remove the requirement of depending on humans to handle such data. In this work particularly, the machine learning technique with the supervised approach has been discussed. The objective of this thesis is to analyze different supervised machine learning algorithms, and use them to determine values as well as make predictions of missing values in the given dataset using the theoretical approach.

Machine learning includes the training of machines with the data such that any models built out of them will be suitable to perform the desired task by those models. So, feeding the data to the models will make them intelligent, and based on the features learned out of such data they will be able to handle the related process. To make use of the machine learning approach, three major approaches are available, which include the use of the supervised approach, the unsupervised approach, and the semi-supervised approach. (Liu, Athanasiou, Padture, & Sheldon, 2020) Supervised machine learning is the one where teacher criteria will be present in the model such that it will be guiding for the development of the model. The labeled data can act as the teacher criteria, and based on the classification of data can be done. It can be understood as the presence of a target attribute out of the larger number of attributes so that the model will try to learn based on the given attributes and the target attributes. This process can be simpler as the model can be aware of the attribute which needs to be learned and those which need to be set as the target. In particular, this thesis work aims to use a supervised approach to machine learning and its use the use of the supervised approach will also be effective in the case of data classification where different data will be classified based on the given target variable. Another approach of machine learning is the use of an unsupervised approach, where the target attribute will be absent due to which model itself needs to analyze the whole data and based. The model will be self-learning on basis of available data with it to produce the final model. The computation with this type of data analysis can be difficult as it requires the model to determine the target variable with the analysis of the complete data during the training process of data. Likewise, another approach for model development will be the semi-supervised approach in model development. This model developed with the use of a semi-supervised approach will utilize both the features of supervised as well as unsupervised to develop the learning model. The complexity of such a semi-supervised model will depend on the type of algorithm which will be followed.

1.1 Scope

Determination of the theoretical values and prediction of the outcomes in the model development is very common practice with the use of the supervised machine learning approach, and this is because the provided target variables will allow the models to be aware of the features which should be targeted. This makes the process of model development quick and it also allows the development of a reliable model since all the required attributes will be collected during the model development. In the current time, the presence of big data has created the situation where the use of a manual approach for data analysis and model development is nearly impossible thus, with the approach selected which includes the use of the supervised machine learning approach it will be simpler. Thus, supervised machine learning has its scope and application across the larger areas including data classification and value prediction. This work attempts to focus on the prediction of the values using the theoretical approach.

The thesis aims to make use of the supervised approach of machine learning so that an effective model can be developed to address the requirement of value prediction. The thesis will also evaluate different supervised learning algorithms and theoretically analyze the outcomes which will be obtained. The major objectives which will be fulfilled with the completion of this research-based work have been listed as:

- To analyze the different supervised approaches of machine learning
- To analyze the different steps which will be used for the machine learning
- To analyze and discuss the areas where machine learning could be used
- To discuss the tools which are available for the handling of the machine learning process

1.2 Limitation

The use of machine learning is one of the latest technologies and has been widely used across the many sophisticated works in the past as well as the current time. This has helped the model development and drastically helped to improve the efficiency of models developed with the use of machine learning. Apart from these benefits, there are several issues and these act as limitations to the process. Among them, a major limitation with the use of the machine learning process for model development and information handling is that it will only be able to handle data up to a certain range. Even though larger data as compared to the manual model will be possible, still, this will not be able to handle larger volumes of data and obtain information from them. In such a situation use of deep learning, models should be implemented only data could be analyzed and the model will be possible to be developed. The deep learning models are learning algorithms that will be used to handle the larger data sets, and features from those data can be extracted by processing them for a larger number of the epoch. The epoch in such process will be related to the time's data will be trained for the model, and with a larger number of time training, better features will be learned by the model developed based on the deep learning algorithm.

1.3 Resources to collect information

In the working of this work, a lot of resources have been used, and since this work is the research-based work that has been performed in the theoretical approach lots of past papers that have been produced in this field has been used. Apart from these sources, online sources and search engine like Google scholar has been used to collect the information required for this research-based work.

1.4 Literature review

The use of machine learning techniques has increased as they have been effective in handling a larger amount of the data and providing better results to the users, and organization. The use of big data, which is being produced currently due to the digitalization of most of the applications has been done and based on which the use of machine learning techniques has developed suitable models for the process. Without the development of the models, it would not be possible to manage big data only with the manual approach, hence the process of the model development has been simplified with the use of different machine learning processes. Apart from this, since data related to any process will be used for the model development and hence outcomes obtained from the process are effective and can provide useful information regarding the future needs of the companies and even the individuals. There are different tools and algorithms which will be required to make use of machine learning, and these will be discussed in the following chapter. The tools are especially useful for the initial stage of the data pre-processing such that finally obtained processed data will be ready to use and implement in the desired type of machine learning algorithm. These tools are very essential since they will be directly responsible for the preparation of the data, and such data will be used to model the system with the use of a machine learning algorithm. So, the correctness of the data should be considered due to which finally developed models will also be reliable and could produce the correct outcomes. Some of the tools for data handling are discussed in the next chapter.

2 Data Science and tools

Data science deals with the data and its analysis, such that when they will be evaluated better results could be obtained. In the digital world, data plays a major role in any model development as well as the decision about the company in the future times. The use of traditional approaches cannot be fulfilled to address the modern requirements where there are huge numbers of data, and also effectively handling all those data is necessary. The accuracy of the collected data is very important, as all the processes of the further model development and thus the accuracy of the model will be dependent on that information. (Aalst, 2016)Modern tools and resources that will come with the use of data science will be able to address all those requirements and provide a reliable system for the users.

2.1 The role of Data Science

Data science is a very broad concept as the possibilities which come with this are a lot more. Each of those roles needs to be performed in a systematic manner such that desired model could be developed and out of which suitable results can be obtained. Different steps that can be followed in data science have been discussed in the following section. (Berti, Zelst, & Aalst, 2019)

- Data collection
  This will be the major and first step in any work that is associated with data science and will include the collection of the data from the sources. This is very important because the collected data should be accurate, as a result of which future works involving the model development and implementation of those results in the real world will depend on the quality and accuracy of the collected results. During this process, data should be collected only from the trusted sites and with prior permission from the owners of the data.

- Data processing

  This will be the next step in the process of the data science-related project and is an essential step. Most of the collected data will be present in their raw format meaning they will not be possible to be used in the model development process. The occurrence of any errors in the data, presence of a null value, missing value, etc. is the major cause that inhibits the direct use of those data directly in any model development process. In this case, pre-processing stage of the data will come handy and all this needs to be done with the use of technology. The manual approach to data handling would only be possible if there would be a limited number of data, but in the current time big data will be present and this will make them nearly impossible for the manual handling of those data. Once the data will be passed through this process they will be suitable for being used in the model and implemented for the real-world situation.
- Model development

  This stage involves the development of the actual model and for which different algorithms will be used. Using the particular algorithms will have its specific results and this will depend on the type of data that is present and the nature of the result which will be required. So, the suitable machine learning algorithm should be selected such that it will be useful in the development of the desired model which will be dedicated to the specific case.
- Model testing

  This will be the testing phase of the model and is very essential as all the developed models will not be guaranteed to work effectively. For this, the outcomes obtained from such results should be analyzed. This is done in the testing of the model. If the model will be able to make a reliable analysis of the given data then it can be said to be suitable for the testing process and even could be used for the further stages.

- Implementing in the real world
  This will be seen as the final stage of any work and model which will be developed with the use of data science and the technologies related to it. All models will be targeted to automate the day-to-day activities in the business such that constant human requirements will be avoided. During this process, the model which will be giving a reliable amount of correctness during the testing phase will be used for system implementation. As the real world will be very challenging as the slight error could be causing loss of significant resources so, it should be analyzed before the implementation that the mode will be providing fitting results to the users. On the other hand, if the developed models will not be possible to develop a reliable amount of accuracy to the user's requirements, then such models should not be considered for the implementation in the real world.

2.2 Python

Python is the scripting language, which supports the general purpose and high-level programming environment. As the general language, it is possible to use python for a wider range of services and these include web development, scientific research, data analysis, game development, software development, etc. There are lots of libraries and modules that have been available for the language. The syntax for this is very short and is very powerful due to which complex operations in the data science and scientific process will also be able to be addressed in a shorter time frame. This has made this language much more popular for work related to data science and scientific experiments.

Currently, the language has two major variants which are python 2 and python 3, and different versions of each of them have been coming. Currently, only version 3 has been developed the updated version.

2.3 Numpy

This is one of the fundamental packages in python which will be used mostly for scientific computation. This package consists of a set of a large number of libraries that can be effective in the analysis of the numerical data in the system development as well as data processing. This package comes with the tools which have to support the numerical computation and the functions which will be able to handle the complex mathematical tasks.

Before using this library it should be installed separately, as this will not come with the default installation of python. This installation is very easy and everyone can install it with the basic skill in the python library installation. Installation of the libraries with the pip installer is a very common and easier approach to installing third parties libraries in the system. This is open source and has been publicly maintained on Github due to which there are a larger number of active communities who use it. Thus, in any case of difficulties faced during the installation and use all of them can be easily addressed and users can use them in their projects.

2.4 Matplotlib

This is the next library that is available with python and will be extensively used for various processes such creation of the visualization which will be static, animated as well as interactive. This makes it one of the popular libraries to be used for data analysis, especially for the visualization of the data. When the analysis of those data about the data evaluation as well as visualization would be planned it would be very difficult, however, the use of the Matplotib library has simplified the whole process. Some of the major works that can be done with the use of this library are the creation and publication of the plots which will be of standard quality, development of the figures which will be interactive with features like zooming in and out as well as an update once the initial work will be complete.

The library also allows the customization of the visual style and layout during the dataset representation. The obtained results can even be exported to a larger number of file formats and among them all they will be possible for the embedding in the Jupyter notebook and also for different graphical user interfaces. Due to all these features, a larger number of third-party libraries have made use of this library. Likewise, all of these processes are very essential concerning the data and model development hence they will be suitable to use in the data analysis processes.

2.5 Seaborn

It is a widely used library for data visualization and has been based on the matplotlib library. This library allows the users to achieve a higher-level interface, which will render the attractive drawing as well as informative graphics related to the statistics. Since this is the library that has been developed by third parties it should be imported to the file before its use and it can be easily installed with the use of the pip installer. The library has also a lot of active communities due to which both new as well as experienced users can easily use this library for the project.

2.6 Pandas

It is the python library, which is used for the data analysis, its manipulation due to which analysis of the data can be performed in the time scale. It is a very powerful and open-source library, and it has been built on top of the python programming language. The ability to analyze a larger amount of data and perform efficient analysis of those data has made it one of the popular libraries among data analysts. The library is simple to use as the learning curve is straight. There are lots of free resources that provide information about the library and it has made the library much more popular. Like other third parties library, it also should be installed separately before it could be used for the data analysis. For this pip, the installer will provide an easier approach to installing the library and making its use in the file.

2.7 Python for Data Science

All the libraries which have been developed on top of the python language, which is related to the data analysis and their handling have made this language much more reliable for the data science-related project. These libraries will help to provide quick results even with the complex process and also will be able to address a larger number of data files. It will be time-consuming and difficult to achieve the entire task related to the data analysis if they need to be developed from scratch work.

Apart from this, the syntax is simple and one needs to write a little number of codes to execute the same work which would have taken lots of code if written in any other programming language. In the current time scope of this language has grown rapidly and this is due to its ability to handle the larger data, which have been produced in huge volume in the current time.

3 Machine Learning

Machine learning involves the use of machines such that any process could be automated and eliminates the dependency on human beings. This process of automation relies on the data and those data will be used to make future decisions about any future process regarding the companies. As all the processes of the model will be depending on the data, during the process of the data collection it should be considered that all the data collection will be done from a valid source because of which it will be possible to gain a reliable model and use them even for the real-life cases at the day to day life.

To develop any model with the use of the ML approach there are three major approaches, these include the use of the supervised learning approach, unsupervised learning approach, and semi-supervised approach. In this work, we are working with the case of the supervised learning approach and some of the popular algorithms that will be based on the supervised learning approach have been discussed in the following section.

3.1 Supervised Machine Learning

Supervised machine learning algorithms refer to the learning process where the model development process will be provided with the target attribute such that any model will be focusing on the attribute. (Hendrycks, Mazeika, Kadavath, & Song, 2019)This makes the supervised learning approach similar to the guided approach where the teacher seems to guide and makes the overall process of the model development simpler and more effective. In the case where there will be a lack of such a target set to the model, there will be extra work that needs to be done by the model during the time of model learning and this could impact the effectiveness of the model. This will also avoid the chances of major attributes being missed in the learning process. Since the target attribute will be initially set it will be learned by the model and this helps the model to gain better knowledge about the data and later will be providing effective results for the system. The use of supervised learning algorithms is much more common with the task that should be working in the classification of the data, and when supervised learning will be used during the feature learning stage it will help to obtain better results during the implementation of the model in the real cases. Some popular supervised learning algorithms have been discussed in the following section.

3.1.1 Linear Regression

Regression is the technique that is widely used for the two major applications in data analysis, and this is related to the prediction of the data and the forecasting of the data. These areas are very common in the project which will be related to machine learning and hence regression is commonly used in the ml project that deals with these areas. Likewise, the next area where the regression analysis will be used extensively includes the area that will be dealing with the determination of the causal relations among the variable of dependent and independent nature, as well as the dataset of the fixed nature having the different variables. In all the regression models there will be the use of independent variables that will be predicting the dependent variables. (Schmidt & Finan, 2017)

Linear regression is the regression model in which, there will be the model development process and dependent variables will be determined based on a single independent variable. (Bartlett, Long, Luosi, & Tsigler, 2020)These models help to determine the impact of the independent variables on the given dependent variables. In any model development process, the linear regression will be making use of the straight line such that it helps to minimize the discrepancies which will be existing between the predicted values and the actual output value. With the use of a linear regression model, it will be possible to estimate the data based on the actual observations from the random sample size of data. Based on this it will be possible to determine if there is a linear relationship between the input and output. Given the set of data, this linear regression method will be used for the model development in which the target variable that will be fitted under the given condition will be selected, whereas other data will be discarded. It can be seen that it will help to provide the working model in which it will be possible to group the data of similar nature in one group and other in the next group.

3.1.2 Logistic Regression

Logistic regression is one of the machine learning algorithms that will be suitable for the evaluation of the dependent variable in the time when it is having a binary outcome. For such a dataset it will help to develop the predictive analysis, such that the model developed out of this algorithm will be able to predict suitable outcomes. By using this algorithm it will be possible to describe the data, and also determine the existing relationship between one dependent variable along with the other nominal, ordinal, or interval variables. (Ranganathan, Pramesh, & Aggarwal, 2017)The logistic regressions are of different types depending on the nature of the data, a suitable method will be used for the classification and prediction of the data. Some of the common types of logistic regression analysis include the presence of binary logistic regression, multinominal logistic regression, and ordinal logistic regression.

In the case of binary logistic regression, there will only be two possible categorical responses, and all the present data will be classified into one of the two possible values. Likewise, with the multinominal logistic regression, there will be more than two categories, and all the present data will be classified in one of the available classes. Such classes will be present without any ordering and suitable classes will be selected based on the feature matching. Finally, with the ordinal logistic regression, there will be more categories and these categories will be ordered. The incoming data will be then provided to the suitable class where there will be a feature match.

Thus, the majority of the situations that will be coming in the real-life cases will be addressed by the regression analysis, and hence the mode training can be done effectively. Depending on the type of dataset available a suitable type of logistic regression will be used for the model development.

3.1.3 Decision Tree

This is one of the widely used and powerful supervised machine learning algorithms, which can be used for the prediction and classification of the available dataset. The development of the decision tree will be done as the normal tree, where the initial point can be taken as the root of the tree, and depending on the type of the variables they will be further developed with the addition of new leaves which will be represented by the nodes in the developed tree. (Arora, 2017)With this decision tree model the decision analysis, which will be done visually and explicitly representation the data and their representation can be done such that they will help in the overall decision making.

For the development of the tree and classification of the available data, there are various algorithms in the given data. Each of these algorithms to develop the decision tree and some of the popular algorithms have been listed and discussed in the following section.

ID3 stands for the iterative Dichotomiser 3 and is used for the development of the decision tree to the given dataset. The development of the decision tree with this algorithm will begin with the root node, and with each iteration, the algorithm will calculate the entropy of all the available features, and the one feature which will be having higher entropy will be selected as the first node to the given tree level. With such selection of one attribute in each of the nodes, this will help to develop the partitioned classes, and classification of the data can be done. The tree will be then repeated until all the attributes will be classified. Since the entropy will be the major basis for the creation of the decision tree, the entropy of all the features will be created for the development of the decision tree. (Phu, Tran, Chau, Dat, & Duy, 2017)

C4.5, this is the extension to the ID3 algorithm, decision tree developed with this algorithm will be used for the classification and hence this algorithm is often referred to as the statistical classifier. The algorithm uses a similar concept as with the case of the ID3 process where the concept related to the information entropy will be used for the development of the algorithm. At each node, this algorithm will select the data attribute which has the highest normalized information gain. This information gains the difference between the entropy. With this process, the algorithm will be repeated to obtain the final decision tree until all the factors will be used to develop the tree structure. (Damanik, Windarto, Wanto, & Poningsih, 2019)

CART is the abbreviated form for classification and regression trees which is the non – parametric decision-making approach used for the development of the classification and the regression-based trees. This will be dependent on whether the input data are of a categorical or numeric nature. The decision trees will be formed with the collection of several rules and variables that will be used in the modeling of the dataset. (Ghiasi, Zendehboudi, & Mohsenipour, 2020) It is based on the selection of the variables out of the rules that are used to split the observation out of the dependent variable. When the particular rule will be selected, then a single node will be split into two nodes, and this will be applied to all available features until the complete decision tree has not been developed.

In this manner, it is possible to make use of the decision tree for the development of the tree such that classification, as well as the prediction of the variables, will be possible.

3.1.4 Random Forest

These are another supervised learning algorithm for the development of the model for the classification, regression, and other tasks which will be operated with the development of the multiple decision trees at the time of the training process (Rigatti, 2017)The concept of the random forest is seen as the collection of the multiple decision trees, which means trees will be used to the creation of the forest. So, the decision class will be selected on the basis that has been developed by the larger number of trees in the forest. It can be seen that, as the results will be selected from the results obtained out of the majority of the trees the selected classes in the random forest will be effective as compared to the single decision tree. It is also seen that the overfitting of the training data that will be seen in the single decision tree will be corrected while a random forest will be used to make the selection of the decision class. In the general sense, the results from the random forest will outperform that of the decision tree. In obtaining the effectiveness of the model which will be coming from the random forest the characteristic of the incoming data will be making a significant impact.

They are often seen as the BlackBox model, for the process of model development as they will be able to generate reasonable outcomes for the wider range of the input data, and meanwhile, they will be requiring a minimum amount of the configuration. With all these features that will be coming to the random forest, it can be used for the model developed in the supervised approach, and even the data can be classified as well as predicted with a reliable amount of accuracy.

4 Method

For any research-based work, this will be related to the analysis of the data and prediction of the data different methods should be followed such that final results will be obtained. The use of the systematic process for the project evaluation will help to develop the results which will be required for the solving of the project. The major phases in the work which has been performed have been discussed in the following section.

4.1 Capturing of Data

Data plays a major part in any work which involves the development of the model, and such data should be collected only from reliable sources such that they could be used for the model development and obtain the reliable outcomes of it. Two major approaches to data collection are present for any work which involves the use of data for model development. These two methods involve the primary collection of the data and the second approach to collect the data. In the primary approach, the data will be collected from the first-hand experiment, due to which specific results that will be suitable for the project will be obtained. Thus, for the projects that need to deal with critical issues, this will be an effective method of data collection. As the downside to this method of data collection, it can be seen that the use of the primary method of data collection will not be suitable for the project which has a short time for the release to the market. The next approach to the data collection is the second approach and this involves the use of the public made available results and data from the past works by the researcher and academician. Since the results will be made available publicly, the use of this method will be very easy and even takes a little time to bring out the project to market. The restriction with this approach is that the results obtained will be general in nature and hence cannot be used for the project which needs accurate results and deals with the critical nature of work.

This work is the evaluation of the theoretical results, due to which any information and results that have been used are obtained in the second approach. The findings made by past researchers and academicians in a similar field have been selected.

4.1.1 Data Entry

This is the initial stage of the work analysis, which will be related to any model evaluation based on the available data. The selected and entered data at this stage will be making an impact on all the future processes and work, which this process should be performed with great care. This can be done by the selection of the data only from reliable sources, which have validated and trusted the standard practice. Unable to maintain this requirement in any process development will result in models which are unreliable.

4.1.2 Data Extraction

The selected data should be further explored and analyzed to extract useful information from them. Before this process, there could be some necessary steps depending on the nature of the data being selected, which includes the pre-processing of the data if the data contains some error values, missing values, etc. once this process will be completed, the data could be used for the extraction of the features. This is usually done during the training process and the model will be allowed to learn features in the data. Such a trained model will be then expected to provide suitable results in the evaluation phase, and only the models that can provide reliable results will be used for the application in daily usage.

4.2 Maintaining of Data

Another essential stage in any data-based model, which will be working in the prediction of the values and estimation of the values is the maintenance of the data. This step ensures that data will be preserved in the standard quality and will be ready for use in the training phase of the model development. This stage consists of the following works.

4.2.1 Data Cleaning

In the process of data cleaning, data will be evaluated for the missing values, and incorrect values if there are in the dataset. In the case such invalid data will be existing they will be cleaned and data will be made suitable for the further process. This can be done with the filling of average value to the missing values or replacing them with 0, due to which they will not be showing any unusual behavior for the model development. Once all these issues will be addressed the data can be ready for use at the model training stage.

4.2.2 Data Architecture

This is the next stage, where the selected data will be evaluated to ensure that it will be suitable for the training process of any particular model. There will be a specific requirement for each of the models, and to address those requirements data should be developed accordingly for that particular architecture. With these entire factors considered the data will be then suitable for the model training stage.

4.3 Processing of Data

With the data collection will be done, the processing of the data will be followed. In this stage, the raw data will be processed and meaningful information from those data will be obtained. For such information extraction and processing of the data, various techniques are available and these have been discussed in the following section.

4.3.1 Data Mining

Data mining is the process of using the raw data, which have been collected out of the various data sources. In this mining process value out of those data will be obtained, and patterns in the larger data sets will be extracted with the use of different machine learning statistics and many more. With this, it can be seen that data mining includes a wider range of processes and some of the major works have been discussed in the following section.

- Detection of the anomaly
  The larger set of data will consist of a huge number of anomalies and the presence of such anomalies will impact the training process of the data. Due to this, it will be essential to detect such anomalies and remove them before the dataset will be used for the data mining process. (Iakovidis, Georgakopoulos, & Vasilakakis, 2018)These anomalies in the dataset can be seen as the major sources of the error in the model. Using the average value can be seen as the way to remove the anomaly in the dataset once they will be detected.
- Association based learning
  One of the major tasks in any data mining process will be the determination of the association among the dataset, and such dependency will be developed based on the existing rules in the database. Some of the commonly used rules include the presence of a relationship existing

between the products which will be frequently brought and its assumption that it will be popular among a large number of the customers. So, in the process of data mining, where the rule-based association will be determined some of the commonly used processes include the market basket analysis customer churn analysis, etc.

- Clustering

  For the case of a larger number of data files clustering is another data mining process and this will be related to the use of datasets and grouping them based on some common features which will be available within them. With this process, it will be possible to estimate the hidden patterns that will be existing in the dataset.

- Classification

  Another common stage that will be done in the case of data mining is the classification of the data into either binary or multiple classes. This classification will be done by initially training the dataset during which the model will be learning useful features out of the model.

All these data mining processes will help extract useful features out of the selected dataset. To get the correct evaluation of the features, data should be mined and processed and then only the mining process should be followed.

4.3.2 Data Predictive Analysis

Data predictive analysis will be done from the models developed with the use of a supervised learning algorithm. Once the model will be trained, it will be used for the predictive analysis of real-world data. Only the model that has a reliable amount of accuracy will be used for the predictive analysis.

4.3.3 Qualitative Analysis

Obtained results need to be analyzed for their use of them in practical situations. Two major approaches can be used for the analysis of those data these are with the use of quantitative analysis and qualitative analysis. Quantitative analysis will deal with the data handling approach, where data are related to the experiments. The results obtained from this quantitative data analysis approach will be effective for all the cases. These results can be hard and fast as they will not vary with the instances of the experiment performed. On the other hand, the qualitative analysis includes the interpretation of the collected information and on basis of the final result will be obtained. This process of the analysis can be subjective as results obtained can vary from the instances of the interpretation and the person performing such evaluation.

This work has tried to perform the theoretical analysis of the past works and the results obtained from them. Which use of the qualitative analysis will be suitable for this work in communicating Data

This stage involves the communication of the data to the outsiders and based on which they can use the result as per the requirement. The information needs to be conveyed more simply and effectively, due to users who will use the data can grab the essential information within a short period. Some suitable data communication processes are discussed in the following section.

4.4 Data Reporting

The report will be drafted, and this will contain all the methods used for the analysis and this also includes the algorithms used for the data mining process. The report will also contain the final results from the analysis process. With all this information contained in the report, it can be used as the source of guidelines for new users as well as for the owner of the report for future references. Such a report will be meant to obtain detailed information within a shorter time frame.

4.5 Data Visualization

It is very common to receive data in the raw format; however, interpretation of those raw data will be difficult. This might even take a longer time to extract the insight out of those data. To avoid such issues and obtain valuable information out of such data use of the visualization technique can be an efficient method. When this visualization tool will be used it provides the ability to gain detailed insight into the data and information can be retrieved in a shorter instance of time.

4.6 Decision Making

Decision-making is the final process, which should be done only after the complete evaluation of the system. Once the data and their results will be evaluated by the management team they will reach a situation where they can make an accurate decision. The results when obtained with the complete steps followed regarding the data mining will be able to give accurate results and help the decision-making process in the company.

5 Conclusion

This thesis aimed to analyze machine learning techniques and various supervised algorithms and their use in the prediction of the values in a theoretical manner. The work has analyzed the significance of the data analysis, and it requires in the current time. It can be seen that with the digitalization of most of the services, an effective process of data analysis should be followed. A traditional approach to data analysis cannot be suitable for modern times since the volume of data is very large. When a limited number of data is present, manual methods can be used for handling the data but this approach will be ineffective as the volume of data increases.

To address these issues, it is essential to make use of big data handling tools and among many of the available options, using machine learning is one option and this can help to address the problem with big data analysis. This work has also analyzed the use of the machine learning approach for data analysis. With the use of machine learning techniques, it is possible to handle the data with the use of several algorithms and these include the use of a machine learning approach with supervised, unsupervised and semi-supervised approaches. The use of the supervised method has been discussed and analyzed throughout this thesis. In the supervised algorithms, there is the prior selection of the target variable due to which models can focus on those attributes during the learning process. The model learns the feature related to that attribute and uses it in the implementation phase.

Different machine learning algorithms have been discussed and during which the working of those algorithms has been analyzed in this thesis. During the evaluation of the result, it is clear that it is n possible to estimate the missing values or classify the data into suitable classes. These are major requirements for the data analysis and using them for their real-world application of the process.

Initially, the standard data are collected, and they undergo the process of cleaning such that any anomalies that would be in the data could be avoided. Some of the common anomalies include the presence of a missing value, the values that are incorrect as per the given context, and many more. Such presence of abnormality is not suitable during the practice of model development. So, the first activity should be the removal of such anomalies in the given dataset and this has been done at the pre-processing stage of the machine learning approach in the model development and system analysis. The next step would be to select the suitable appropriate machine algorithms for the model development. With the use of the machine learning approach the model development, three major types of algorithms can be used for the model development, and out of the supervised learning, an algorithm has been selected in this work. With the supervised approach to training the model, it is seen that model is possible to be developed where the target variable will be initially selected due to which model will be especially focusing on those attributes. This makes the model effective in learning the parameter which makes a major impact on the given dataset. The processed data are used to learn by the developed model regarding the features present in the given dataset and this phase is called the training phase. Once the training is complete then the model should be tested and verified such that it will be able to produce a reliable amount of accuracy. This is done in the testing phase of the model, if the model gives a reliable amount of accuracy it can be used for real-world implantation. This ensures that decision of the data classification and prediction will be possible with the selected model.

Finally, the thesis has discussed the various useful attributes related to the data analysis for which different tools and techniques have been discussed. All the tools and techniques in the data analysis will be working to produce the information which can be useful in the decision-making process. Various libraries have been discussed in this report, and most of them developed in the Python language, due to the large number of Python libraries l, Python is seen as the powerful language for the data science and analysis of those data.

The collected results are not the final stage in the data analysis and evaluation, and this process should be completed with the process of documentation and visualization of the obtained results. The various methods of documentation as well as visualization have been discussed in the thesis Those results and documentation will provide detailed insight into the data clearly and effectively.

References

Aalst, W.V.(2016).Data Science In Action. Springer.

Akinsola, J.E.(2017).Supervised Machine Learning Algorithms: Classification and Comparision.Research Gate,12

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, *163*(8), 15–19. https://doi.org/10.5120/ijca2017913660

Bartlett, P. L., Long, P. M., Luosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. PNAS.

Berti, A., Zelst, S. J., & Aalst, W. (2019). Process mining for python(PM4Py): Bridging the gap between process - and data science. arxiv.

Damanik, I. S., Windarto, A. P., Wanto, A., Poningsih, Andani, S. R., & Saputra, W. (2019). Decision tree optimization in C4.5 algorithm using genetic algorithm. *Journal of Physics. Conference Series*, *1255*(1), 012012. https://doi.org/10.1088/1742-6596/1255/1/012012

Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A. (2020). Decision tree-based diagnosis of coronary artery disease: CART model. *Computer Methods and Programs in Biomedicine*, *192*(105400), 105400. https://doi.org/10.1016/j.cmpb.2020.105400

Hendrycks, D., Mazeika, M., Kadavath, S., & Song, D. (2019). Using self-supervised learning can improve model robustness and uncertainty. In *arXiv [cs.LG]*. http://arxiv.org/abs/1906.12340

https://www.geeksforgeeks.org/clustering-in-machine-learning/.(2021). Retrieved from https://www.geeksforgeeks.org: https://www.geeksforgeeks.org/clustering-in-machine-learning/

https://www.javatpoint.com/data-mining.(n.d.). Retrieved from https://www.javatpoint.com/

Lakovidis, D. K., Georgakopoulos, S. V., & Vasilakakis, M. (2018). *Detecting and locating Gastrointestinal anomalies using deep learning and iterative cluster unification*. IEEE.

Liu, X. (2020). *A machine learning approach to fracture mechanics problems*. https://doi.org/10.26226/morressier.5f5f8e69aa777f8ba5bd603e

Phu, V. N., Tran, V. T., Chau, V. T., Dat, N. D., & Duy, K. L. (2017). A decision tree using the ID3 algorithm for English semantic analysis. Springer Link*, 593-613.*

Ranganathan, P., Pramesh, C. & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. PMC *, 148-151.*

Rigatti, S. J. (2017). Random forest. Journal of insurance medicine*, 31-39.*
Schmidt, A. F., & Finan, C. (2017). Linear regression and the normality assumption. Elsevier*, 146-151.*

Wang, J., Yang, Y., Wang, T., & Sherratt, R. S. (2020). Big Data Service Architecture: A survey. Journal of Internet