A project report on

Disease Prediction Using Machine Learning

Submitted to

AMITY SCHOOL OF APPLIED SCIENCES

AMITY UNIVERSITY

LUCKNOW CAMPUS, UTTAR PRADESH



SUBMITTED BY

ADITYA SRIVASTAVA

MASTER OF STATISTICS

A8979220002

IIIrd semester

Under the guidance of

Dr Gunjan Singh

Amity university

Lucknow campus

Lucknow

2021

**Declaration**

I do hereby declare that the project entitled Disease Prediction Using Machine Learning was carried out by me at Amity University from 6th July 2021 to 6th August 2021. This project has not been submitted elsewhere for any other.

Date: 27th August 2021 Aditya Srivastava

Place: Lucknow Master of Statistics

Amity University

**Certificate**

This is to certify that project entitled “Disease Prediction Using Machine Learning” submitted to Amity University as a part of the project embodies the work done by Mrs Gunjan Singh under my supervision and guidance at Amity University.

Dr Gunjan Singh

Amity School of Applied Sciences

**Acknowledgement**

I express my deep sense of regard and gratitude to my teacher and guide, dr Gunjan Singh, Amity school of applied sciences, under whose supervision I had been carrying out this work. I will always be grateful to her for his continuous encouragement, valuable suggestions, and generous support. I wish to acknowledge my sincere thanks to all respected teachers, especially Mrs Shailja Pandey, Department of Statistics, Amity School of Applied Sciences, Amity University for her cooperation and guidelines throughout this course. All the members of the library, data centre, my sincere thanks.

Finally, I take this opportunity to thank Almighty God and my parents for providing me with the opportunity to be a part of Amity University.

**Abstract**

The pharmaceutical sector is huge nowadays and the research in this field is getting expanding day by day so data is also created in huge numbers. Managing a huge number of data cannot be handled manually. Machine learning helps to manage the data efficiently. My paper is constructed on the disease prediction based on symptoms using machine learning algorithms which is decision tree regression algorithms. The process is based on the symptom’s disease can be predicted through the machine. We are going to predict the disease in a particular body according to the different symptoms.

**Introduction**

Every person is suffering from some type of disease and handling every is very easy. Disease prediction is the way through a person understood the disease from their symptoms and then took the medicines or the treatment is required for the disease

Machine learning is a subsidiary of artificial intelligence that helps to work through algorithms that will make upgrades automatically through experience by the use of data. Machine learning can assist the disease prediction process in many ways.

When a person has a complaint about the problem or symptoms then it is going to be tested and if the symptoms are not stopped and more symptoms are showing disease is predicted according to the symptoms not by seeing. What the results say are defined the diseases.

**Dataset**

It is the assembly of data. Everything around is the data. What we use in mobile has a collection of data. What we do in a day is also data. Data is the most important thing to do in any inspection. If we do not have the data then we cannot make any predicted conclusions with any data. Prediction with data helps to understand the outcomes of the particular objective.

In this project, we have the data of some diseases with their symptoms. Data has 500 people for the disease prediction that can be done through symptoms. the data has some symptoms such as itching, chills, nodal skin eruptions, sneezing shivering, stomach pain, joint pain according to their symptom’s disease can be predicted. The whole data was collected by the company which is QtPi Robotics and then the analysis is done on the dataset.

**Review of Literature**

**Machine Learning**

Machine learning is a very demanding technology nowadays. It is a subset of Artificial Intelligence. It was first named by **ARTHUR** **SAMUEL** in 1959 who is an expert in computer gaming and artificial intelligence which stated that **“IT GIVES COMPUTERS THE ABILITY TO LEARN WITHOUT BEING EXPLICITLY PROGRAMMED”** but the first definition was stated by **TOM M. MITCHELL** which said that

In our real-world examples such as score prediction, spam detection, gaming we all use in our daily life. We as a human is very independent on machine learning methods. From morning to night, we use these methods but do not know that we are using them. We are so much dependent on this system that we do not know how to do that kind of work if we do not have that type of help. Machine learning helps mankind for a better future but with so much helps around even if we do not want that our nature and we cannot think without their help. Everything has merits and demerits so does machine learning which we discuss later.

Machine learning is within computer science but it is different from traditional computational approaches. In traditional computational approaches, algorithms are explicitly programmed instructions given to the computers whereas in machine learning computers are permited to train on data inputs and use statistical analysis in building the model for the sample data.

In machine learning when the input the data it will automatically interpret the data to analyze the data using some machine learning algorithms.

According to Wikipedia, in the next ten years machine learning can generate up to one billion dollars in the pharmaceutical company and it is estimated that it can replace 25% of the jobs which is a great disadvantage.

There are some chief aspects that show the importance of machine learning –

1. The rapid increase in the production of data
2. Solving complex problems which are difficult for human beings.
3. Decision making in various sectors including finance.
4. Finding hidden patterns and extracting useful information from data.

**Machine learning is mainly seprated into three types**

1. Supervised machine learning
2. Unsupervised machine learning
3. Reinforcement machine learning

**Supervised machine learning** is an algorithm in which the model is instructed using the labelled data that is they have the input and some predicted output. It needs supervision to train the model. there are many algorithms in supervised machine learning such as regression analysis. In this paper, we are using decision tree regression which is used when we have to take a decision-related problem.

**Application of Machine Learning**

There are some applications in machine learning such as: -

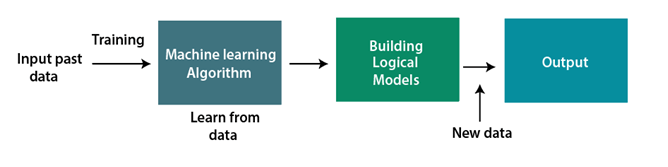
1. Netflix uses machine learning in recommendations. When someone opens Netflix app for the preliminary it shows infamous recommendations on TV shows and movies but the data is assembled when the person is using the app and after then recommendations in their app.
2. A facial recognition system is present in every mobile phone in which through patterns it recognizes the person which is an application of machine learning and essential language processing.

**How does machine learning work?**

Machine learning learns from historical data, build prediction models, it also modifies the model as the new data comes and predicts the updated outputs. The accuracy of the data depends on the huge amount of data. If the data is big then the model is also better and the prediction is also more accurate.

If we have a complex problem and we have to find some predictions, so rather than writing a whole code, we just need particular algorithms and new data to build an efficient model.

The block diagram of how machine learning implied is as follows

in the block diagram shown above, we have to first input the whole data which was collected through the past experiences after then data is split into training data and testing data. In the training, the data model is created and in testing, data prediction is made. Now through the training data model is created different machine learning algorithms and output is then created but if the new data is added then it modifies them and then again it gives the output.

**History of Machine Learning**

The Early History of Machine Learning (pre – 1940)

In **1834**, Charles Babbage the father of the computer come up with a device that was arranged with punch cards. However the machine was never constructed, but today all modern computers relied on that logical structure.

In **1936**, Alan Tuning had a theory of how can machine can decide and execute a set of instructions.

The period of Stored Program Computers

In **1940**, the first automatic handled computer, **ENIAC** was invented which was the initial common computer. After this **EDSAC** was also made in **1949** and **EDVAC** in **1950** to the stored-program computer.

In **1943**, a person semantic network was modelled with an electrical circuit. From now on scientists started applying the idea of human neurons by analyzing that how they work.

**Computer Machinery and Intelligence**

In **1950** Alan Tuning published a paper **“Computer Machinery and Intelligence”** on the topic of artificial intelligence. In this project, he asked the audience that how can machines work.

**Machine Intelligence in Games**

In **1952**, A program was created that helped an IBM computer to play

a checkers game. It was created by Arthur Samuel, who was the

pioneer of Machine Learning. It performed better more it played.

The term “Machine Learning” was first spelt by Arthur Samuel in 1959.

**The First “AI” Winter**

The duration **1974** to **1980** is called “AI” Winter because it was the hard time for AI and ML researchers.

**“In this duration, failure of machine translation took place, and humans had lower their interest in AI, which led to lesser funding by the government to the researches.”**

**Machine Learning from Thesis to Actuality**

To remove reverberating over phone lines the first neural network was

applied to a actual life troubles using an modifying filter in **1959**.

A neural network NETtalk was invented in **1985**, which was able to

instruct itself how to accurately spell **20,000 words in one week**. It

was invented by **Terry Sejnowski** and **Charles Rosenberg**.

In **1997** IBM invented the intelligent computer named deep blue, who secure the chess game in opposition to the chess specialist **Garry Kasparov**. It became the first computer that had beaten a human chess specialist.

**Machine Learning in 21 st century**

In the year **2006,** a new name has given to neural net research as **“deep learning”** by computer scientist **Geoffrey Hinton**. Nowadays, it has became prime moving technologies.

A deep neural network was created by Google in **2012** which learned to acknowledge the representation of persons and cats in YouTube videos.

 In **2014**, Chabot &quot; Eugen Goostman&quot; cleared the Turing Test. It was the first Chabot who convinced 33% of human magistrates that it was not a gadget.

In **2014** a new deep neural network DeepFace was created by Facebook and declared that it could acknowledge a person with.

**Machine Learning at Present**

Machine learning has not a current development in research. It is present everywhere in engineering, medical fields, business, technology etc. such as self-driven cars, amazon Alexa, transportation and commuting etc. it has to mow many ways of algorithms such as supervised machine learning, unsupervised machine learning, support vector machines, reinforcement machine learning, clustering analysis, association etc.

**Disease Prediction Using Machine Learning**

Disease prediction can be done for ages but the procedure is the time taken to make the less amount of time and less amount of money it should be used through machine learning. If a person is having a fungal infection, then one symptom through another is going to be matched then the conclusions should. Previously it is done manually so it can be predicted another disease they do not have any idea regarding some serious symptoms which is not showing on the body and after it is showing it leads to the seriously ill person.

**Decision tree regression**

A decision tree is the prime common practical approaches for supervised learning. It is familiar with decode both Regression and grouping tasks with the end being more into experimental approach. The decisions or the test are executed based on the characteristics of the given dataset.

A Decision Tree is a tree-structured classifier. It has three types of nodes. The Root Node is the opening node that represents the complete specimen and may get break further into further nodes. The Interior Nodes represent the quality of a data set and the subdivide shows the conclusion rules. Finally, the Leaf Nodes represent the result. This design is very handy for resolve conclusion-related difficulties.

Diagram

Description automatically generated

With a particular data point, it is run fully through the total tree by replying True/False untill it outreach the leaf node. The final prediction is the mean of the rate of the response variable in that specified leaf node. Through different duplication, the Tree can forecast a actual rate for the data point.

Diagram

Description automatically generated

The illustration is a portrayal of the execution of a Decision Tree algorithm. Decision trees have an upper hand that it is uncomplicated to recognize, minor data washing is needed, non-linearity does not influence the model’s showing and the number of hyper-variables to be adjust is nearly void. Although, it can have an over-connection issue, which can be solve using the ***Random Forest***algorithm.

In this specimen, we will go through the execution of ***Decision Tree Regression***, in which we will predict the profit of an frozen yogurt shop based on the climate in a locality for 300 days.

**Why do we use Decision Trees?**

There are various algorithms in Machine learning, so choosing the finest variable for the given dataset and question is the main point to remember while making a machine learning model. Below are the two motive for using the Decision tree:

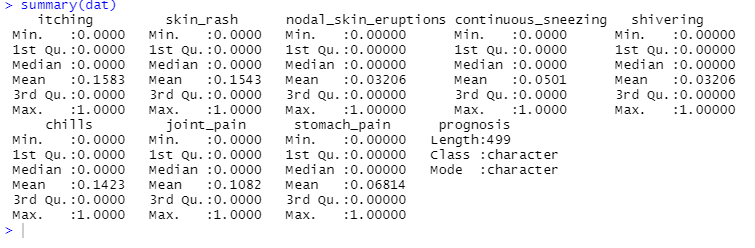
* Decision Trees generally copy human thinking ability while making a decision, so it is easy to acknowledge.
* The idea behind the decision tree can be easily acknowledged because it shows a tree-like structure.

**Decision Tree Terminologies**

1. **Root Node:** The root node starts from where the decision tree begins. It shows the whole dataset, which further gets divided into two or more similar sets.
2. **Leaf Node:** Leaf nodes are the last output node, and the tree cannot be separated further after getting a leaf node.
3. **Splitting:** Splitting is the procedure of splitting the decision node/root node into sub-nodes as stated in the given statements.
4. **Branch/Sub Tree:** A tree formed by breaking the tree.
5. **Pruning:** Pruning is the process of separating the unwanted branches from the tree.
6. **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

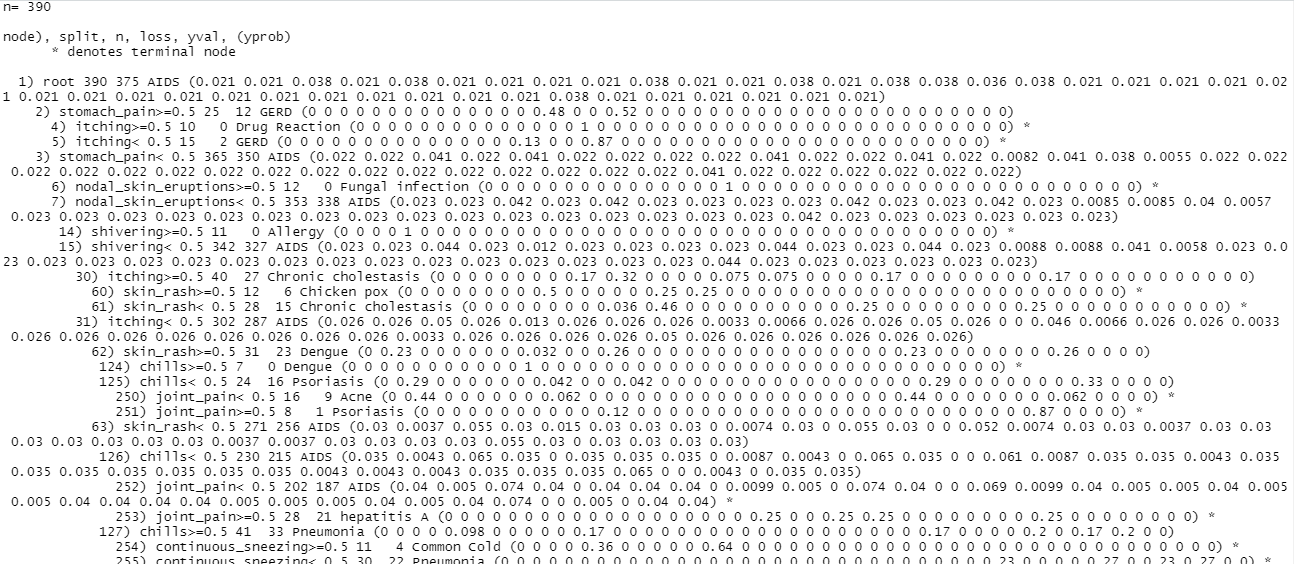
**Analysis**

**Table 1**

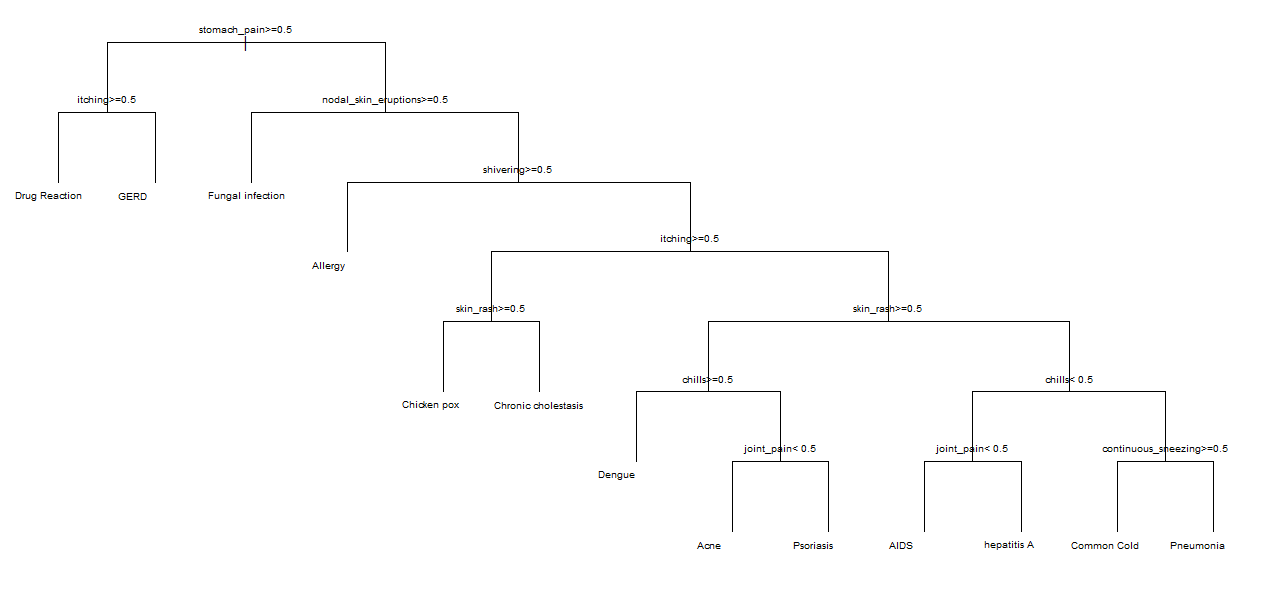


This table shows the outline of every variable. The numerical variable has their minimum, maximum, median, 1st quartile and their 3rd quartile whereas the categorical variable is showing classes of a particular set.

**Table 2**

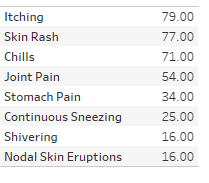


This table showing is the decision tree model. The data used is the training dataset which is created by splitting the whole data in the training set and the test set and the model is shown in simpler form in the next graph.



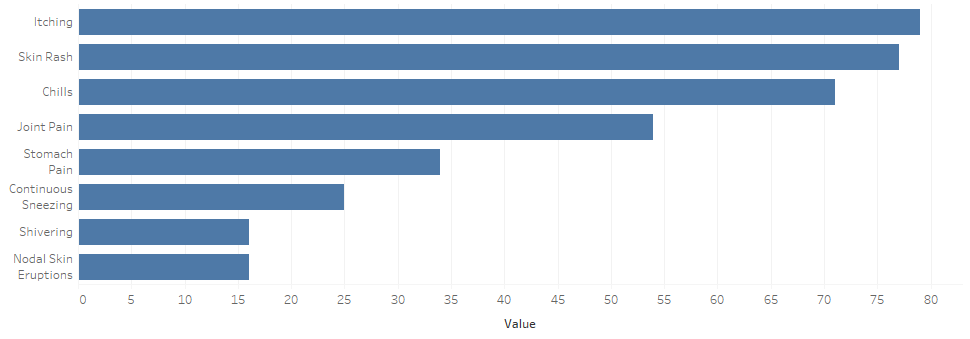
In this graph at the top, the root node is the stomach pain. First, the person has started the stomach pain and if the person shows another symptom such as itching or nodal skin eruptions then according to their symptoms their disease is predicted.

**Table 3**



This table conveys the result of symptoms on the disease. It notifies us that the itching has a 79% effect which is entirely high and then it is skin rash which is 77% and the least is nodal skin eruptions which are 16% as it started earlier but it has less effect on the human body.

**Graph 2**



This graph convey how much effect the human body of disease by symptoms.

**Conclusion**

We concluded that machine learning helps to predict disease in a great manner. Predicting any disease with the aid of the decision tree algorithm of machine learning. Researchers get great help from machine learning so that they don’t have to follow lengthy methods. A person who gets the symptoms of any disease gets readily detected by the computer. Some symptoms can be visible and they are in high percentage in our body to predict any disease. Our traditional system of predicting disease is outdated and it needs to be replaced as it will waste a huge amount of time, money and mainly services that will use somewhere where it is also needed.

**References**

<https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda>

<https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>

Mir A, Dhage SN (2018) Diabetes disease prediction using machine learning on big data of healthcare. In: 2018 fourth international conference on computing communication control and automation (ICCUBEA) [Google Scholar](https://scholar.google.com/scholar?q=Mir%20A%2C%20Dhage%20SN%20%282018%29%20Diabetes%20disease%20prediction%20using%20machine%20learning%20on%20big%20data%20of%20healthcare.%20In%3A%202018%20fourth%20international%20conference%20on%20computing%20communication%20control%20and%20automation%20%28ICCUBEA%29)

Singh M, Bhatia V, Bhatia R, Big data analytics solution to healthcare [Google Scholar](https://scholar.google.com/scholar?q=Singh%20M%2C%20Bhatia%20V%2C%20Bhatia%20R%2C%20Big%20data%20analytics%20solution%20to%20healthcare)

Koppad SH, Kumar A, Application of big data analytics in the healthcare system to predict COPD [Google Scholar](https://scholar.google.com/scholar?q=Koppad%20SH%2C%20Kumar%20A%2C%20Application%20of%20big%20data%20analytics%20in%20healthcare%20system%20to%20predict%20COPD)

Ray S (2019) A quick review of machine learning algorithms. In: 2019 international conference on machine learning, big data, cloud and parallel computing (Com-IT-Con), India, 14th–16th Feb 2019 [Google Scholar](https://scholar.google.com/scholar?q=Ray%20S%20%282019%29%20A%20quick%20review%20of%20machine%20learning%20algorithms.%20In%3A%202019%20international%20conference%20on%20machine%20learning%2C%20big%20data%2C%20cloud%20and%20parallel%20computing%20%28Com-IT-Con%29%2C%20India%2C%2014th%E2%80%9316th%20Feb%202019)

DhomseKanchan B, MahaleKishor M (2016) Study of machine learning algorithms for special disease prediction using principal component analysis. In: 2016 international conference on global trends in signal processing, information computing and communication. IEEE 20168 [Google Scholar](https://scholar.google.com/scholar?q=DhomseKanchan%20B%2C%20MahaleKishor%20M%20%282016%29%20Study%20of%20machine%20learning%20algorithms%20for%20special%20disease%20prediction%20using%20principal%20component%20analysis.%20In%3A%202016%20international%20conference%20on%20global%20trends%20in%20signal%20processing%2C%20information%20computing%20and%20communication.%20IEEE%2020168)

Shetty D, Rit K, Shaikh S, Patil N (2017) Diabetes disease prediction using data mining. In: 2017 international conference on innovations in information, embedded and communication systems (ICIIECS)[Google Scholar](https://scholar.google.com/scholar?q=Shetty%20D%2C%20Rit%20K%2C%20Shaikh%20S%2C%20Patil%20N%20%282017%29%20Diabetes%20disease%20prediction%20using%20data%20mining.%20In%3A%202017%20international%20conference%20on%20innovations%20in%20information%2C%20embedded%20and%20communication%20systems%20%28ICIIECS%29)

Thirunavukkarasu K, Singh AS, Irfan M, Chowdhury A (2018) Prediction of liver disease using classification algorithms. In: 2018 4th international conference on computing communication and automation (ICCCA)[Google Scholar](https://scholar.google.com/scholar?q=Thirunavukkarasu%20K%2C%20Singh%20AS%2C%20Irfan%20M%2C%20Chowdhury%20A%20%282018%29%20Prediction%20of%20liver%20disease%20using%20classification%20algorithms.%20In%3A%202018%204th%20international%20conference%20on%20computing%20communication%20and%20automation%20%28ICCCA%29)

SanthanaKrishnan J, Geetha S (2019) Prediction of heart disease using machine learning algorithms. In: 2019 1st international conference on innovations in information and communication technology (ICIICT)[Google Scholar](https://scholar.google.com/scholar?q=SanthanaKrishnan%20J%2C%20Geetha%20S%20%282019%29%20Prediction%20of%20heart%20disease%20using%20machine%20learning%20algorithms.%20In%3A%202019%201st%20international%20conference%20on%20innovations%20in%20information%20and%20communication%20technology%20%28ICIICT%29)

<https://www.researchgate.net/publication/347381005_Disease_Prediction_Using_Machine_Learning>

<https://in.springboard.com/>

<https://www.besanttechnologies.com/>

<https://www.stat.cmu.edu/>

<https://stackoverflow.com>

<https://www.datamentor.io/>

<https://www.forbes.com/>

<https://www.investopedia.com/>

<https://www.techfunnel.com/>

<https://www.sciencedirect.com/>

<https://www.digitalvidya.com/>

<https://marketbusinessnews.com/>

<https://www.tutorialspoint.com/>

<https://in.springboard.com/>

<https://www.javatpoint.com/>

<https://towardsdatascience.com/>

<https://www.topechelon.com>

<https://www.geeksforgeeks.org/>

<https://books.google.co.in/>

<https://science.sciencemag.org/>

<https://www.adacovid19.org/>

<https://www.sciencedirect.com/>

<https://www.analyticsvidhya.com/>

<https://www.edureka.co/>

<https://www.digitalocean.com/>

Introduction to machine learning by Nils J Nilsson

<https://www.udemy.com>

<https://en.wikipedia.org/>

<https://machinelearningmastery.com/>

<https://www.tutorialspoint.com/>

<https://www.careeraddict.com/>

<https://recruiteze.com/>

<https://www.candidatemanager.net/>

<https://get.fountain.com/>