# PREDICTION OF DISEASE WITH MINING ALGORITHMS IN MACHINE LEARNING

**A V V Sai Pranav*[1],N Sai Susmitha Naidu[2],N Manishanker[3],P Dolavanya[4],T Gowtham Srinivas[5]**

*[1,3,4,5]Student, Department of Computer Science and Engineering,GITAM UNIVERSITY, Vizag, Andhra Pradesh 530045, India*
*[2]Assistant Professor, Department of Computer Science and Engineering, GITAM UNIVERSITY, Vizag, Andhra Pradesh, 530045, India*

---

**Abstract -***Technology these days is pivotal for better human livelihood. In technology, if it is particularly about a specific domain then Machine learning has a colossal effect on the human lifestyle these days. Fields like Banking, Software products, Information technology, Agriculture, Defence, Manufacturing, Education, Marketing e.t.c., which healthcare sector is no exception in it. Many of the subsidiary fields in the healthcare sector like a prediction of chronic, infectious, physiological diseases and the diagnoses, tracking and monitoring the patients in real-time scenarios, and many more. With the help of the most advanced algorithms and training, data sets and machine learning models are designated to many fields and subsidiaries. Prediction of diseases based on the symptoms that are being felt by a human makes more sense in synthesizing a model with significant datasets and perceiving machine learning algorithms. The dataset which is being trained will have routine symptoms that one would ever suffer with so that the disease that is going to be predicted will make more sense and be relatable. This application lasted by illustrating the proper usage and working of most admired algorithms like Decision tree classification algorithm, Random forest classification algorithm, Naïve Bayes classifier algorithm. On a whole, this paper dispenses the working mechanism of the algorithms in the prediction of illness(or disease).*

***Key Words*: *Machine Learning, Classification, Naive Bayes Algorithm, Decision Tree Algorithm, Random Forest Algorithm.***

## 1. INTRODUCTION

The healthcare or pharmaceutical sector has a great need for technology to achieve more advancement and to provide more sophisticated treatment and facilities for mankind. It is quite obvious that these fields need more data mining algorithms in a correct manner which will help physicians or surgeons to give appropriate treatment and will be accommodating to doctors [1].

Diseases or illnesses such as Dengue, COVID -19, malaria, Ebola, Yellow fever, HIV / AIDS, Viral gastroenteritis, Varicella, Viral hepatitis, Chikungunya, Jaundice, etc., could affect a drastic change in an individual's health or sometimes turns out to be a life-threatening problem and cause death if disregard. Classification algorithms viz. Decision Tree, Random Forest, and Naïve Bayes algorithms can hint us an antidote to that particular circumstance [2]. Any given sector, even the healthcare sector, definitely makes a drastic change by mining the huge database and exploring the hidden patterns in investigating the case's future

## 2. OVERVIEW

A quick overview can give a conceptual idea and brief design language of how the model works in real-time. Few flow charts, and a brief about algorithms , tools used in developing the model helps to build a broad idea over the model working.

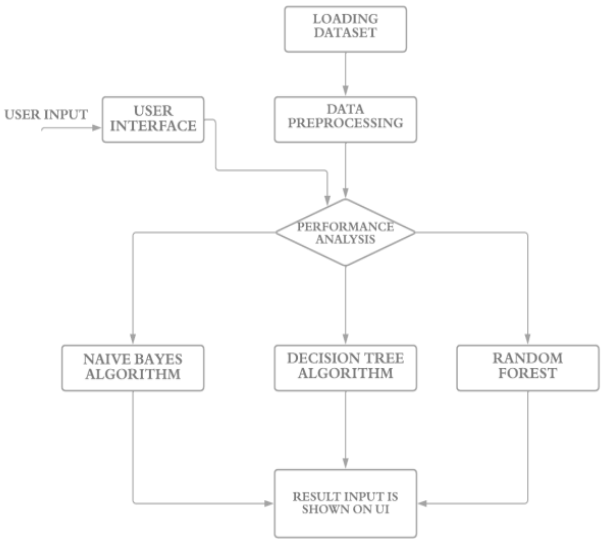


**Fig.1 A Simple flow chart for the Machine Learning model**

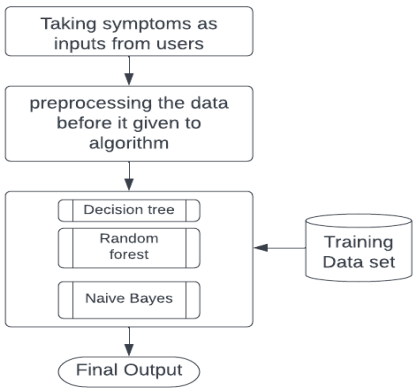


**Fig.2 Illustrating the work flow of the model**

### 2.1 Input

In order to predict and expect an appropriate disease as an output, the user must have to choose the symptoms, and the User Interface(UI) accepts the inputs from the user which acts as a barrier in between the machine learning classification algorithms and the user. There are almost 90+ symptoms provided to the user for choosing between the right symptom, in turn, giving the most precise disease that the user is suffering from given symptoms.

### 2.2 Preprocessing of given data

This is a crucial step that converts the given raw data into a format that can be simply interpreted by the algorithm is generally known as preprocessing of the given user data. This also contains sub-parts like data cleaning, data integration, and data reduction as well.

- ***Data cleaning:*** This is a step, where the data set contains empty cells or NaN values or data of any other different data type. By this, the training of the model goes flawlessly without any interruptions [3].
- ***Data integration:*** This step contains a method in which different kinds of data or formerly called heterogeneous knowledge is taken from many resources and assimilated into one kind of data, which appears to be a unified view of data. The data can be in any form such as sets, documents and tables and so on which can be used for personal or commercial access. This is all about the Data integration process [4].
- ***Data reduction:*** In this step, The training of the model, or the prediction may take a longer amount of time when considering huge volumes of data or training sets. So, for the present application, the data set is optimized in such a way that there is no chance of losing the accuracy in predicting the disease for the user with given symptoms [5].

Classifying algorithms that are used will be, Decision trees, Random forest, Naive Bayes classifiers. These algorithms work on the given symptoms by the user along with analyzing the data set which the model already trained.

### 2.3 Output

After finishing up the above preprocessing steps along with training, a set of rules are formed. The output will be occurred based on the user inputs and the algorithm that is chosen by the user for predicting the disease with reference to the given symptoms as inputs.Now, Methodology deals more about this machine learning model.

## 3. Methodology

Basically, there are three algorithms that are used in this particular application viz. Decision tree classifier,Random forest classifier, Naïve Bayes Classifier. Each of them has its own significance in prediction or mining.

### 3.1 Decision Tree

As the name suggests, a Decision tree classifier works just like a tree data structure. It works like a tree data structure in computing so that the classification of the symptoms(here) is performed. Any given volume of data will be bifurcated into smaller modules or clusters, sometimes these smaller divisions are also called smaller subsets that straight away target disease in the given data set. As this sounds like a tree data structure, similarly decision tree classifiers also have a leaf node and a decision node [6].

The formal definition for decision tree classifier will be, A decision tree is a flow chart-like tree structure, with each internal node representing the test of the attribute, each branch representing the result of the test, and the class label represented by each leaf node [7].

The formal definition for decision tree classifier is A flowchart-like tree structure, in which an internal node targets the output (disease here), label for the class is described by each terminal node or simply a lead node. The decision tree classifier algorithm is said to be quicker and more accurate when compared with any other classifier algorithm. This mining algorithm has a decision tree that behaves like a predictive model. This model tracks all the observations about an item or an input, in order to conclude the target value for the item or that given input. These algorithms have a special kind of feature that makes it more sensitive and functional by building SQL statements from the tree that makes the decision tree get access over a database. ID3 is said to be

the ramp-up version for the Decision tree classifier algorithm based on Hunt's algorithm and introduced by Quinlan Ross in the year 1986. This ID3 algorithm stands for Iterative Dichotomiser which is an easy decision tree learning algorithm. There are many more algorithms like C4.5, CART algorithm, and many more[8].

The decision tree has a wide scope, and it has too many applications. It is been successfully implemented in sectors like business, intrusion detection, energy modeling, E-Commerce, Image processing, Medicine, Industry, intelligent or automatic vehicles, remote sensing related appliances, web applications, software applications e.t.c,

### 3.2 Random forest

Random forest. Leo Bremen is the person who found this algorithm. In general, a Forest means the group of trees in real-life. In the same way, the random forest also consists of a group of trees or rather the random forest algorithm constructs a lot of decision trees to get an accurate and steady prediction value [9]. This brings more randomness to a project or the model which we are constructing [10]. A very special and strategical feature that random forests have is while doing any bifurcation of nodes, the split nodes or subsets seek for the best subset of random articles or features which at last brings a more sure-shot result(output of disease). In short, the random forest algorithm synthesizes more number decision trees and mixes them together for a next-level prediction outcome. This algorithm comes under supervised classification and an ensemble learning type of classifier that has a wide range of applications in every sector(varieties of applications). Random forest is very capable of dealing with any large volume of data given to it for the analysis part. It shows up on the off chance that the tree has memorized the information.

Generally, a random forest classifier chooses a subset randomly from training samples and variables, which brings more randomness to the model and also the classifications made more accurate due to this reason. Random forest is very effective and capable in terms of handling high data. Not only the data which is dimensionality but also multicollinearity [11] . Random forest classifier is a classical example for ensemble learning method, which builds many decision trees that will be classified a new instance that has a majority vote[12] .Random forest classifiers with many of the above-described features gained a good scope and served in many sectors(fields) which result produced a wide range of applications in daily life for human beings. The fields which make use of random forests are medicine, banking, e-commerce, stock markets, software industry, remote sensing geography, and many more [13].

### 3.3 Naïve Bayes Classifier

A typical definition of Naïve Bayes Classifier is that A classifier which is quite probabilistic in nature and this algorithm is established upon applying Bayes theorem with strong independent assumption [14]. Naïve Bayes Classifier is a powerful form of Bayes' theorem that can be compared to any other classifiers in terms of accuracy. Naïve Bayes Classifier is actually mixed up with the computational productivity as well as with many promising features. Some of the key and highlight features of the Naïve Bayes Classifier are computational efficiency, Low variance, Incremental learning, prediction of posterior probabilities, Robustness in the face of noise as well as in the face of missing values [15]. With many of these features and flexibility in Naïve Bayes Classifier has successfully proven its simplicity and efficiency [16]. The performance of the Naïve Bayes Classifier is seen clearly when the features are completely independent and functionally dependent and also that the worst performance can be observed in between the above-described conditions [17]. Naïve Bayes can be considered as a highly practical classifier as it was involved in a sector having more applications.

Applications of the Naïve Bayes classifier will be spam filters, diagnosis for cancer, face recognition, Sentimental analysis, and many more examples and applications. And also, the Naïve Bayes requires very fewer amounts of computational energy or power[18].There are proper evidence and research which took place stating that in a comparing classification algorithm, both classifiers viz. Naïve Bayes and decision tree classifiers are comparable to each other performance perspective, and also Naïve Bayes classifier manifested high accuracy as well as speed when the training data set or variables of huge volumes[19].Thus, the Naïve Bayes classifier is one of the classifiers which works effectively.

## 4. TOOLS USED FOR BUIDLING APPLICATION

Tools here refer to modules or libraries which are the fundamental and working agents of any machine learning algorithm implementation. These are the roots for every single model which makes the model to compile without any interruption. The in-built data structures in python programming language are capable of establishing up to certain speeds in computing, but these tools makes the model to achieve another level of compilation power and speed without loosing a trace of accuracy. Accuracy never gets effected by using these modules. Data scientist are aware that python has abundant number of modules for programming.

### 4.1 NumPy

**Numpy** is a very renowned open-source library in python which got awestruck features for computation and building high-end software applications. NumPy has very revealing syntax and it is very significant to have a simple and meaningful syntax. Numpy is treated as a very important and mandatory requirement for any software application construction. Along with these features, NumPy has a very simply made syntax for accessing and also for manipulating, operating the data available in vectors, n-dimensional arrays as well as in matrices. NumPy Array data structure from NumPy module which is very effective in storing and also accessing multidimensional arrays which are also known as tensors, which acts as a gateway for scientific computation and complex problem-solving [20].The data in the model can be of any numerical stored in any data structural format, by using NumPy method called ravel is used to transform any given multidimensional array or n-dimensional array into a typical array or 1-d array.

### 4.2. Pandas

**Pandas** is a productive open-source library that is very compelling used in python programming. It is a promising

library for data analysis and data processing. Pandas is a must and inevitable who works in the field of data science, as the data scientist frequently works on large data for data analysis and data training. To analyze and process, pandas can actually deal with any file format Here in the application development, the pandas library is used in training the data set for the models [21]. For an instance, the training data set is a .CSV file which read_csv() function is used to read the training data set later on manipulates on the data which is read as part of the training of model with the given data set.

### 4.3 Sklearn

**Sklearn** also known as scikit-learn is an open-source python library which is developed by David Cournapeau, which well known to every data scientist and programmer related to data science. This sklearn or scikit-learn is a library that is a sub-unit of SciPy also known as Scientific Python[22]. As this is a part of SciPy, sklearn can be utilized in scientific computing, and also the main use of this library is to integrate any algorithms in machine learning, data analysis. Here in this application, the fit(x,y) function is used, which supports training for the supervised estimators[23]. For an instance, As the decision tree comes under supervised estimator or supervised learning algorithm, fit(x,y) function is used in the application. Not only all these described features but also, sklearn has rangy pre-cleaned datasets available within the Sklearn library, which is extremely efficacious for any developer or data scientist that are building any kind of machine learning algorithms, machine learning prediction models, and so on.

### 4.4 Flask

Python is high-level easy-to-use programming. Python is known for its compatibility, readability of source code, even it supports web browsers as well, which makes programmers run their application on web browsers. For making the application execute on the web browser, there are two popular libraries for this task. One among them is flask which is used to make the application run on web servers or browsers. Here, pickle is used along with flask. Flask has that inbuilt functions which make the programmer more ease of writing a script in this particular library. Flask is a light weight framework that supports a project to run on web browsers and this library is built based on WSGI and jinja2 template for this flask framework[24].Flask is used here in this application and a basic HTML page is being made that successfully runs on a web browser.The output of that web application will be displayed in future sections along with other locally run User Interfaces.

## 5. RESULT

Accuracy score in a table format, about GUI and how the input takes (this will be changed accordingly and if flask works, it will be attached along with python Tkinter GUI), how to operate ( a brief ), I/O and O/P things in the whole model. In this particular section of the report, the Efficiency measure of each of the algorithms used in the model, All the information about GUI, the final output of the application will be described herein detailed.

### 5.1 Efficiency measure of the application

The application is trained based on the training data set which is containing about 40 diseases and symptoms of about 130, these symptoms cause the diseases. Using a few

methods and upon printing the accuracy of each algorithm in numbers. The fact is that all the three considered mining algorithms have done the job in a splendid way in terms of performance and efficiency. And these mining algorithms producing a better accuracy supports a finer prediction of illness or a disease for user of application. Here is a table that displays the accuracy for each algorithm:

**Table -1: Accuracy Scores**

| Mining algorithm used | Accuracy score |
|---|---|
| Decision Tree mining algorithm | 0.95121 |
| Naïve Bayes mining algorithm | 0.95121 |
| Random Decision algorithm | 0.95121 |

### 5.2 GUI Preview

GUI plays a predominate role in fetching inputs from the user, GUI acts as a bridge in between the trained model and user. Initially the User Interface looks like this:

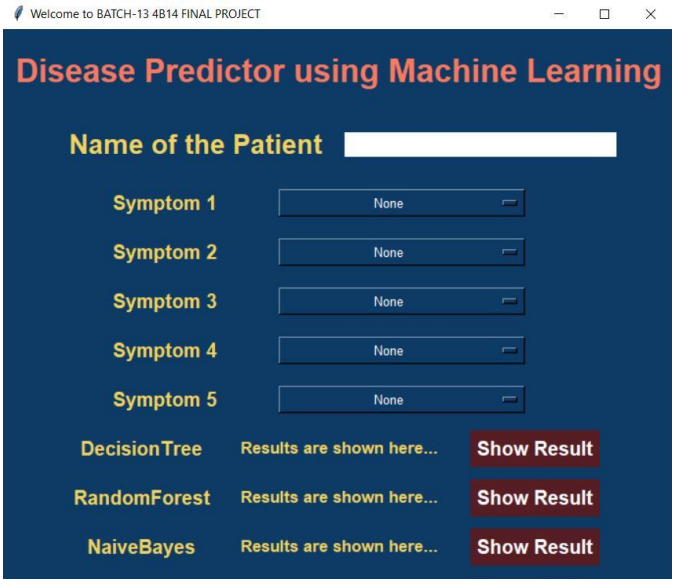


**Fig.3 Initial preview of UI**

As the picture shows the user end Interface, this consists of 5 sections. These 5 sections are supposed to be symptoms that the user is suffering from. After hitting the button, there will be a list shown containing symptoms. At most the user can enter 5 symptoms. After that, the algorithm should be chosen. The final prediction of each algorithm will be displayed down in the text field. Here is the real-time example below,

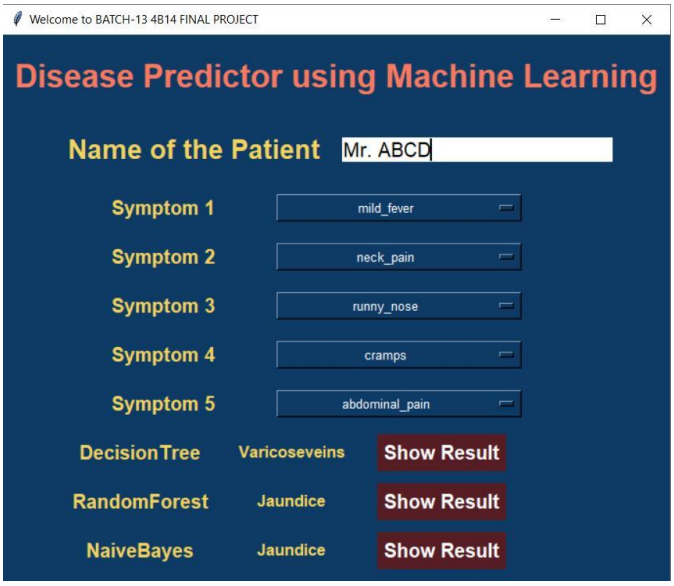


**Fig.4 End Result of the application**

In this case user has entered a few symptoms like, “mild_fever”, “neck_pain”, “runny_nose”, “cramps, “abdominal_pain”, and the predictions are “Varicoseveins” , “Jaundice”, “Jaundice” for the mining algorithms Decision Tree classifier, Random Forest classifier, Naive Bayes classifier respectively. The main observation is, “Jaundice” is the disease that is more predominant as it appears as the prediction disease output by 2 out of 3 algorithms. Therefore, the user must probably be suffering from “Jaundice” as per the model and assuming the given symptoms as inputs.

## 6. PROPOSED SYSTEM

As the implementation of the above proposal is successful in a local computing environment, one can actually level up the same project with many booming technologies and resources. Developers can make this application into a mobile application with new cloud technologies or web technologies as well. The proposed system should actually work on a cloud platform by deploying the whole application in Heroku via GitHub and getting the global host address with Heroku deployment. Once after finishing up the hosting on Heroku servers, with the website URL, we can make a Progressive Web App.

Progressive Web Application can be a kind of an app, which acts like a native application but, internally it runs Chrome mobile services and run the given URL with few native android bits in it and gives an extraordinary feel just like a native android mobile application, eventually, this PWA can be available on iPhones as well. The advantage of making PWA is that these progressive web applications are extremely light in weight unlike a hybrid android application or native Android/iOS application that occupies a large amount of storage space inside a mobile phone’s internal memory. Secondly, An edge over making this PWA is that it can be very easy to develop with basic web technologies such as HTML, CSS, Vanilla JavaScript, JSON. The fact that cannot be denied is, this application needs no automatic or manual app updates.

Because the progressive web app has that URL running inside chrome of mobile and website never get updates like mobile applications. Whenever there is a change in website, the web browser anyhow refresh automatically and reflect the change or update that have to be happened. Which the maintenance is pretty easy. For, making this machine learning application on web browser, we need to use flask or pandas. Flask is very easy to write the code and not very complex, so it is better to use. Here is a preview of the attempt towards web hosting.

The practical implementation, which is successful in local server hosting, can be linked to any kind of front-end document and make it more user friendly. We have run this application locally on a PC web browser with a very basic web User Interface composed with HTML:


**Fig.5 Initial Web UI**


**Fig.6 Entering Symptoms**

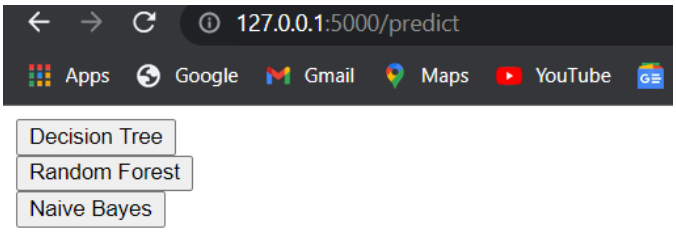


**Fig.7 UI for Choosing Algorithm**

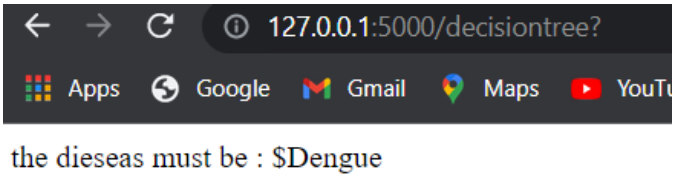


**Fig .8 Result For Decision tree**

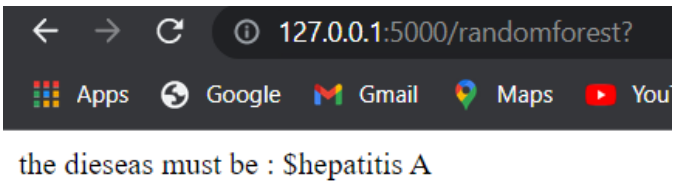


**Fig .9 Result for Random Forest**

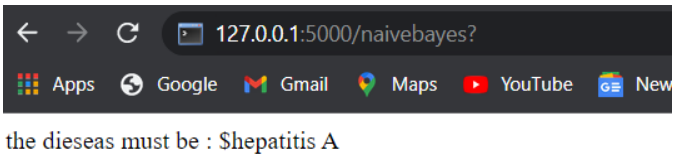


**Fig .10 Result For Naïve Bayes**

## 7. CONCLUSION

Machine Learning is obnoxiously powerful, futuristic, unimaginable , out of the box but yet, it is everywhere and linking each and every aspect of human life style which is making livelihood for humans in a way different from the past. Machine learning makes the human living more sophisticated wit many more enlightening ideas. These ideas are being used in every field these days and health care sector is no exception in it. A sweet and simple UI with better accuracy will lead to the prevention of many diseases by predicting it before it turns out to be life-threatening for somebody who is suffering with few health issues. This may or may not be used in emergency cases, but can be considered as an application which warns before the disease gain more severity and cost a life in few cases. This model is yielding a stunning 95% accuracy rate, which is best at this level. These are few things which hints that, Artificial Intelligence, Machine Learning, Deep Learning technologies will never fade away and inevitable for mankind in coming decades, this will rule all the fields and can create wonder that will lasts forever. These are few things that still make technology to be mandatory in daily life and also the future generations should make use of these advanced technologies to invent and reproduce more advanced and impeccable power generating technologies that will last forever.

## REFERENCES

[1] Aldahiri, Amani, Bashair Alrashed, and Walayat Hussain. "Trends in using iot with machine learning in health prediction system." Forecasting 3.1 (2021): 181-206.

[2] Ahmad, Muhammad Aurangzeb, Carly Eckert, and Ankur Teredesai. "Interpretable machine learning in healthcare." Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. 2018.

[3] Chu, Xu, et al. "Data cleaning: Overview and emerging challenges." Proceedings of the 2016 international conference on management of data. 2016.

[4] Czarnowski, I., & Jędrzejowicz, P. (n.d.). Data Reduction Algorithm for Machine Learning and Data Mining. Lecture Notes in Computer Science, 276–285. doi:10.1007/978-3-540-69052-8_29

[5] Czarnowski, I., & Jędrzejowicz, P. (n.d.). Data Reduction Algorithm for Machine Learning and Data Mining. Lecture Notes in Computer Science, 276–285. doi:10.1007/978-3-540-69052-8_29

[6] Radhika, S., et al. “Symptoms Based Disease Prediction Using Decision Tree and Electronic Health Record Analysis.” European Journal of Molecular & Clinical Medicine, vol. 7, no. 4, 2020, pp. 2060-2067.

[7] Tu, P-L., and J-Y. Chung. "A new decision-tree classification algorithm for machine learning." TAI'92-Proceedings Fourth International Conference on Tools with Artificial Intelligence. IEEE Computer Society,1992.

[8] Sharma, Himani, and Sunil Kumar. "A survey on decision tree algorithms of classification in data mining." International Journal of Science and Research (IJSR)” 5.4 (2016): 2094-2097.

[9] Liu, Yanli, Yourong Wang, and Jian Zhang. "New machine learning algorithm: Random forest." International Conference on Information Computing and Applications. Springer, Berlin, Heidelberg, 2012.

[10] Keerthan Kumar, T. G., C. Shubha, and S. A. Sushma. "Random forest algorithm for soil fertility prediction and grading using machine learning." Int J Innov Technol Explor Eng (IJITEE) 9.1 (2019).

[11] Belgiu, Mariana, and Lucian Drăguţ. "Random forest in remote sensing: A review of applications and future directions." ISPRS journal of photogrammetry and remote sensing 114 (2016): 24-31.

[12] Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas. "How many trees in a random forest?." International workshop on machine learning and data mining in pattern recognition. Springer, Berlin, Heidelberg, 2012.

[13] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[14] Garg, Bandana. "Design and Development of Naive Bayes Classifier." (2013).

[15] Webb, Geoffrey I., Eamonn Keogh, and Risto Miikkulainen. "Naïve Bayes." Encyclopedia of machine learning 15 (2010): 713-714.

[16] Karthika, S., and N. Sairam. "A Naïve Bayesian classifier for educational qualification." Indian Journal of Science and Technology 8.16 (2015): 1-5.

[17] Wickramasinghe, Indika, and Harsha Kalutarage. "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation." Soft Computing 25.3 (2021): 2277-2293

[18] Grampurohit, Sneha, and Chetan Sagarnal. "Disease prediction using machine learning algorithms." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.

[19] Katkar, Vijay D., and Siddhant Vijay Kulkarni. "A novel parallel implementation of Naive Bayesian classifier for Big Data." 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE). IEEE, 2013.

[20] Harris, C. R., et al. "Smith 474 nj." Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'ıo JF, Wiebe M, Peterson P, G'erard-475 Marchant P, et al. Array programming with NumPy. Nature 585.7825 (2020): 357-362.

[21] Hagedorn, Stefan, Steffen Kläbe, and Kai-Uwe Sattler. "Putting Pandas in a Box." CIDR. 2021.

[22] Kramer, Oliver. "Scikit-learn." Machine learning for evolution strategies. Springer, Cham, 2016. 45-53.

[23] Nelli, Fabio. "Machine Learning with scikit-learn." Python Data Analytics. Apress, Berkeley, CA, 2018. 313-347.

[24] Vyshnavi, Vangala Rama, and Amit Malik. "Efficient Way of Web Development Using Python and Flask." Int. J. Recent Res. Asp 6.2 (2019):

## BIOGRAPHIES

**A V V Sai Pranav*** is an undergraduate student, pursuing Bachelor of Technology in Computer Science and engineering department from GITAM UNIVERSITY, Andhra Pradesh, India. His research was on Machine Learning domain and working on few papers.

**N Sai Susmitha Naidu** has completed her M.Tech in Artificial Intelligence & Robotics from Andhra University College of Engineering(A). She has completed her B.Tech (CSE) in 2016 from Sir C.R.Reddy College of Engineering. She is currently working as Assistant Professor in the Department of CSE, GITAM(Deemed) University, Visakhapatnam. Her research interests include Network Security and Machine Learning.

**N Manishanker** currently pursuing his bachelor of technology in field of computer science at GITAM UNIVERSITY, Andhra Pradesh, India. His research interest are in the field of machine learning

**P Dolavanya** is an undergraduate student, pursuing Bachelor of Technology in Computer Science and Engineering department from GITAM UNIVERSITY, Andhra Pradesh, India. Her research was on Machine Learning domain

**T Gowtham Srinivas** is currently pursuing his Bachelor of Technology in Computer Science and Engineering from GITAM UNIVERSITY, Andhra Pradesh, India.