"DISEASE PREDICTION SYSTEM USING MACHINE LEARNING"

A PROJECT REPORT

Submitted by

Sahil Ranjan Mund (1701210232)
Manraj Singh (1701210161)
Akash Kumar Sahu (1701210272)

in the partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY,8th Semester

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

Under the esteemed guidance of

Mr. T Appa Rao

(Associate Professor, Department of ECE)

At

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
GANDHI INSTITUTE OF ENGINEERING AND TECHNOLOGY MAIN CAMPUS
GUNUPUR – 765022
2020-21

DECLARATION

I hereby declare that the project entitled **"Disease Prediction System Using Machine Learning"** submitted for the B.Tech. Degree is my original work and the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles.

Date: 05/06/2021

Place: GIETU, Gunupur

Name of the Student:

Sahil Ranjan Mund (1701210232)

Manraj Singh (1701210161)

Akash Kumar Sahu (1701210272)

.

Gandhi Institute of Engineering & Technology Main Campus

GUNUPUR – 765 022, Dist: Rayagada (Odisha), India

*(Approved **by AICTE, Govt. of Orissa** and Affiliated **to Biju Patnaik University of Technology**)*

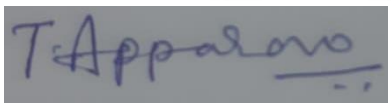
☎: 06857 – 250172(Office), 251156(Principal), 250232(Fax), e-mail: gandhi_giet@yahoo.comvisit us at www.giet.org

ISO 9001:2000 Certified

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

*This is to certify that the project entitled **"Disease Prediction System using Machine learning"** Is the bonafide work carried out by **Sahil Ranjan Mund, Manraj Singh, Akash Kumar Sahu** having registration number **1701210232 , 1701210161 and 1701210272** student of **B.Tech(ECE),Gandhi Institute of Engineering and Technology Main Campus, Gunupur** during the academic year 2020-21 in partial fulfillment of the requirements for the award of the Degree of **BACHELOR of Technology in Electronics and Communication Engineering**.*



[MR. T APPA RAO]
Project Guide
Associate.Prof. in DeptofElectronics,
G.I.E.T.,Gunupur

[DR. SUBHRAJIT PRADHAN]
HOD, B. Tech.
Dept of Electronics,
G.I.E.T., Gunupur

[EXTERNAL EXAMINER]

ACKNOWLEDGEMENT

Any work has to have contributions and supports of a few at least and a bunch at best. It gives me immense pleasure to have an opportunity to express my gratitude to various people directly or indirectly related to my present project work.

I sincerely believe that I was fortunate enough to have come across a group of human beings whose mere mode of living can inspire someone to work wonders. My project guide Associate Prof. Mr T Appa Rao is paradigm of that suit. He not only encouraged me to pursue my studies but also instilled the self-belief and confidence, which helps me to take on various trivial, serious and inevitable hassles of life.

I am very much thankful to the Head, Department of Electronics and Instrumentation Engineering, GIET Gunupur, Prof. Dr Subhrajit Pradhan for his valuable suggestions and supports during the project periods.

I would also grab the opportunity to thank entire faculty group for their moral support and cooperation to make this project success.

My family has been the axis of my mental strength and their support leaves me with irreparable debt and honest gratitude. Sincere thanks to them for willingly accepting my decision of opting for higher studies.

I would like to express my heartiest gratitude to the central library staff members for their kind co-operation during my work.

I express my thanks to all my friends for their timely suggestions and encouragements.

I might have missed a many more names and they all deserve the best of my feeling for their selfless helps and supports in their own way.

ABSTRACT

Nowadays, people face various diseases due to the environmental condition and their living habits and because of this disease prediction at an earlier stage has become very important. But the accurate prediction based on symptoms becomes too difficult for doctors. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays a vital role in predicting diseases. Every year medical science provides a large amount of data increasing exponentially. The development and exploitation of several prominent Data mining techniques in numerous real-world application areas (e.g.Industry, Healthcare, and Bioscience) has led to the utilization of such techniques in machine learning environments, to extract useful pieces of information of the specified data in healthcare communities, biomedical fields, etc. The accurate analysis of medical database benefits in early disease prediction, patient care, and community services. The techniques of machine learning have been successfully employed in assorted applications including Disease prediction. With the help of disease data, data mining finds hidden pattern information in the huge amount of medical data. in this project, we proposed general disease prediction based on the symptoms of the patient. For the disease prediction, we use Logistic Regression, Support Vector Machine (Linear Classifier), Support Vector Machine (RBF Classifier), Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbour (KNN) prediction of disease. In this general disease prediction the living habits of a person and check-up information consider for the accurate prediction, the person fills in the check-up details over a user-friendly web interface which processes the details over various machine learning algorithms as mentioned above and then shows the result and also the accuracy for all the algorithms used for processing. Anyone who wants to check for breast cancer, diabetes, heart disease, kidney disease, lung cancer can easily know it just by filling in the details.

Methodologies used:-

- Machine learning
- Django

Conclusion:-

The project aims to immensely help to solve health-related issues by assisting physicians to predict and diagnose diseases at an early stage. This project presents a survey of various models based on such algorithms and techniques and analyse their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and logistic regression are used here.

TABLE OFCONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION TO MACHINE LEARNING

1.1 What is Machine Learning?

Machine learning is a branch of artificial intelligence (AI) focused on building applications that learn from data and improve their accuracy over time without being programmed to do so.

In data science, an algorithm is a sequence of statistical processing steps. In machine learning, algorithms are 'trained' to find patterns and features in massive amounts of data in order to make decisions and predictions based on new data. The better the algorithm, the more accurate the decisions and predictions will become as it processes more data.

Today, examples of machine learning are all around us. Digital assistants search the web and play music in response to our voice commands. Websites recommend products and movies and songs based on what we bought, watched, or listened to before. Robots vacuum our floors while we do . . . something better with our time. Spam detectors stop unwanted emails from reaching our inboxes. Medical image analysis systems help doctors spot tumors they might have missed. And the first self-driving cars are hitting the road.

We can expect more. As big data keeps getting bigger, as computing becomes more powerful and affordable, and as data scientists keep developing more capable algorithms, machine learning will drive greater and greater efficiency in our personal and work lives.

1.2 Machine Learning Methods

Machine learning methods (also called machine learning styles) fall into three primary categories:-

Supervised machine learning :-

Supervised machine learning trains itself on a labeled data set. That is, the data is labeled with information that the machine learning model is being built to determine and that may even be classified in ways the model is supposed to classify data. For example, a computer vision model designed to identify purebred German Shepherd dogs might be trained on a data set of various labeled dog images.

Supervised machine learning requires less training data than other machine learning methods and makes training easier because the results of the model can be compared to actual labeled results. But, properly labeled data is expensive to prepare, and there's the danger of overfitting, or creating a model so closely tied and biased to the training data that it doesn't handle variations in new data accurately.

Unsupervised machine learning :-

Unsupervised machine learning ingests unlabeled data—lots and lots of it—and uses algorithms to extract meaningful features needed to label, sort, and classify the data in real-time, without human intervention. Unsupervised learning is less about automating decisions and predictions, and more about identifying patterns and relationships in data that humans would miss. Take spam detection, for example—people generate more email than a team of data scientists could ever hope to label or classify in their lifetimes. An unsupervised learning algorithm can analyze huge volumes of emails and uncover the features and patterns that indicate spam (and keep getting better at flagging spam over time).

Semi-supervised learning :-

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of having not enough labeled data (or not being able to afford to label enough data) to train a supervised learning algorithm. It is also called as Reinforcement learning.

1.3 Real World Machine Learning Use Cases

Machine learning is everywhere.Here are just a few examples of machine learning we might have encounter every day:

- **Digital Assistants**: Apple Siri, Amazon Alexa, Google Assistant, and other digital assistants are powered by natural language processing (NLP), a machine learning application that enables computers to process text and voice data and 'understand' human language the way people do. Natural language processing also drives voice-driven applications like GPS and speech recognition (speech-to-text) software.

- **Recommendations**: Deep learning models drive 'people also liked' and 'just for you' recommendations offered by Amazon, Netflix, Spotify, and other retail, entertainment, travel, job search, and news services.

- **Contextual online advertising**: Machine learning and deep learning models can evaluate the content of a web page—not only the topic, but nuances like the author's opinion or attitude—and serve up advertisements tailored to the visitor's interests.

- **Chatbots**: Chatbots can use a combination of pattern recognition, natural language processing, and deep neural networks to interpret input text and provide suitable responses.

- **Fraud detection**: Machine learning regression and classification models have replaced rules-based fraud detection systems, which have a high number of false positives when flagging stolen credit card use and are rarely successful at detecting criminal use of stolen or compromised financial data.

- **Cybersecurity**: Machine learning can extract intelligence from incident reports, alerts, blog posts, and more to identify potential threats, advise security analysts, and accelerate response.

- **Medical image analysis**: The types and volume of digital medical imaging data have exploded, leading to more available information for supporting diagnoses but also more opportunity for human error in reading the data. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and other deep learning models have proven increasingly successful at extracting features and information from medical images to help support accurate diagnoses.

1.4 How Does Machine Learning Work?

There are four basic steps for building a machine learning application (or model). These are typically performed by data scientists working closely with the business professionals for whom the model is being developed.

Step 1: Select and prepare a training data set

Training data is a data set representative of the data the machine learning model will ingest to solve the problem it's designed to solve. In some cases, the training data is labeled data tagged' to call out features and classifications the model will need to identify. Other data is unlabeled, and the model will need to extract those features and assign classifications on its own.

In either case, the training data needs to be properly prepared randomized, de-duped, and checked for imbalances or biases that could impact the training. It should also be divided into two subsets: the training subset, which will be used to train the application, and the evaluation subset, used to test and refine it.

Step 2: Choose an algorithm to run on the training data set

Again, an algorithm is a set of statistical processing steps. The type of algorithm depends on the type (labeled or unlabeled) and amount of data in the training data set and on the type of problem to be solved.

Common types of machine learning algorithms for use with labeled data include the following:

- Regression algorithms: Linear and logistic regression are examples of regression algorithms used to understand relationships in data. Linear regression is used to predict the value of a dependent variable based on the value of an independent variable. Logistic regression can be used when the dependent variable is binary in nature: A or B.

For example, a linear regression algorithm could be trained to predict a salesperson's annual sales (the dependent variable) based on its relationship to the salesperson's education or years of experience (the independent variables.) Another type of regression algorithm called a support vector machine is useful when dependent variables are more difficult to classify.

- Decision trees: Decision trees use classified data to make recommendations based on a set of decision rules. For example, a decision tree that recommends betting on a particular horse to win, place, or show could use data about the horse (e.g., age, winning percentage, pedigree) and apply rules to those factors to recommend an action or decision.
- Instance-based algorithms: A good example of an instance-based algorithm is K-Nearest Neighbor or knn. It uses classification to estimate how likely a data point is to be a member of one group or another based on its proximity to other data points.

Algorithms for use with unlabeled data include the following:

- Clustering algorithms: Think of clusters as groups. Clustering focuses on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without prior knowledge about the groups and their characteristics. Types of clustering algorithms include the K-means,TwoStep, and Kohonen clustering.
- Association algorithms: Association algorithms find patterns and relationships in data and identify frequent 'if-then' relationships called association rules. These are similar to the rules used in data mining.
- Neural networks: A neural network is an algorithm that defines a layered network of calculations featuring an input layer, where data is ingested; at least one hidden layer, where calculations are performed make different conclusions about input; and an output layer. where each conclusion is assigned a probability. A deep neural network defines a network with multiple hidden layers, each of which successively refines the results of the previous layer. (For more, see the "Deep learning" section below.)

Step 3: Training the algorithm to create the model

Training the algorithm is an iterative process–it involves running variables through the algorithm, comparing the output with the results it should have produced, adjusting weights and biases within the algorithm that might yield a more accurate result, and running the variables again until the algorithm returns the correct result most of the time. The resulting trained, accurate algorithm is the machine learning model an important distinction to note, because 'algorithm' and 'model' are incorrectly used interchangeably, even by machine learning mavens.

Step 4: Using and improving the model

The final step is to use the model with new data and, in the best case, for it to improve in accuracy and effectiveness over time. Where the new data comes from will depend on the problem being solved.

CHAPTER 2

Machine Learning Using Python

2.1 What is Python?

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures,combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.it is used in various domains like :

- Web development
- Software development
- Mathematics
- System scripting
- Computations and Analysis

2.2 How does python work?

This image illustrates how python runs on our machines:

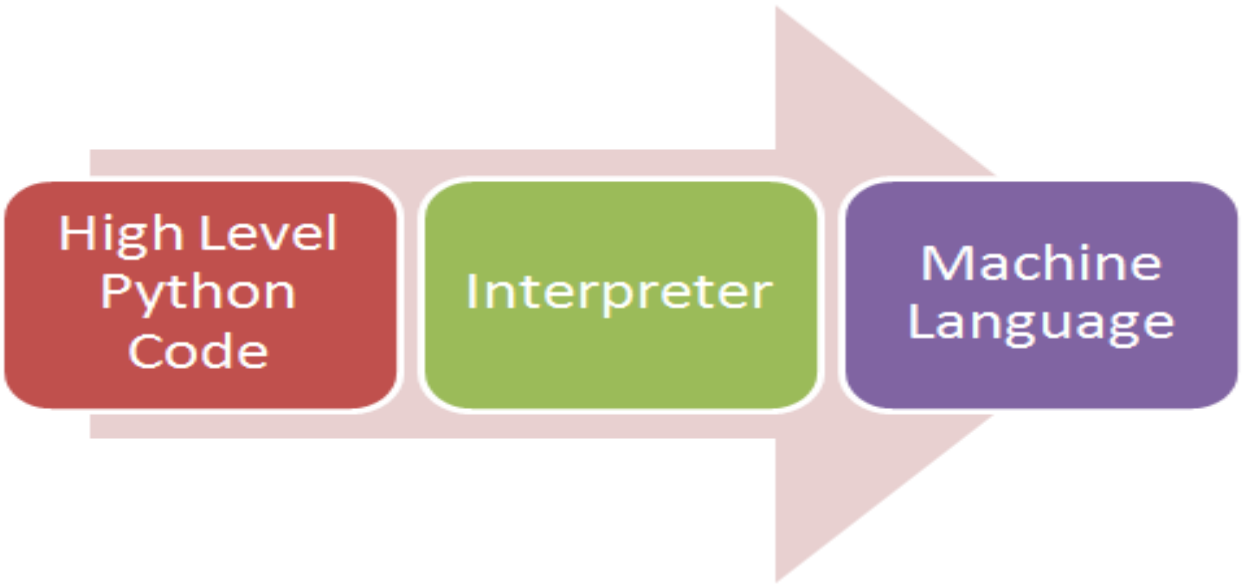


Figure 1: How python works

The key here is the Interpreter that is responsible for translating high-level Python language to low-level machine language.

The way Python works is as follows:

1. A Python virtual machine is created where the packages (libraries) are installed. Think of a virtual machine as a container.
2. The python code is then written in .py files
3. Python compiler compiles the Python code to bytecode. This bytecode is for the Python virtual machine.

Now, this virtual machine is machine-dependent but the Python code isn't.

4. When you want to execute the bytecode then the code will be interpreted at runtime. The code will then be translated from the bytecode into the machine code. The bytecode is not dependent on the machine on which you are running the code. This makes Python machine-independent.

Python byte code is Python virtual machine-dependent and this makes Python machine-independent.

The point to note is that we can write Python code in one OS, copy it to another OS and simply run it.

2.3 Why we use python for machine learning?

AI projects differ from traditional software projects. The differences lie in the technology stack, the skills required for an AI-based project, and the necessity of deep research. To implement your AI aspirations,

you should use a programming language that is stable, flexible, and has tools available. Python offers all of this, which is why we see lots of Python AI projects today.

From development to deployment and maintenance, Python helps developers be productive and confident about the software they're building. Benefits that make Python the best fit for machine learning and AI-based projects include simplicity and consistency, access to great libraries and frameworks for AI and machine learning (ML), flexibility, platform independence, and a wide community. These add to the overall popularity of the language.

Simple and consistent

Python offers concise and readable code. While complex algorithms and versatile workflows stand behind machine learning and AI, Python's simplicity allows developers to write reliable systems. Developers get to put all their effort into solving an ML problem instead of focusing on the technical nuances of the language.

Additionally, Python is appealing to many developers as it's easy to learn. Python code is understandable by humans, which makes it easier to build models for machine learning.

Many programmers say that Python is more intuitive than other programming languages. Others point out the many frameworks, libraries, and extensions that simplify the implementation of different functionalities. It's generally accepted that Python is suitable for collaborative implementation when multiple developers are involved. Since Python is a general-purpose language, it can do a set of complex machine learning tasks and enable you to build prototypes quickly that allow you to test your product for machine learning purposes.

Extensive selection of libraries and frameworks

Implementing AI and ML algorithms can be tricky and requires a lot of time. It's vital to have a well-structured and well-tested environment to enable developers to come up with the best coding solutions.

To reduce development time, programmers turn to a number of Python frameworks and libraries. A software library is pre-written code that developers use to solve common programming tasks. Python, with its rich technology stack, has an extensive set of libraries for artificial intelligence and machine learning. Here are some of them:

- Keras, TensorFlow, and Scikit-learn for machine learning
- NumPy for high-performance scientific computing and data analysis
- SciPy for advanced computing
- Pandas for general-purpose data analysis
- Seaborn for data visualization

Scikit-learn features various classification, regression, and clustering algorithms, including support vector machines, random forests, gradient boosting, k-means, and DBSCAN, and is designed to work with the

Python numerical and scientific libraries NumPy and SciPy.With these solutions, you can develop your product faster. Your development team won't have to reinvent the wheel and can use an existing library to implement necessary features.Here's a table of common AI use cases and technologies that are best suited for them:

DATA ANALYSIS AND VISUALIZATION	NUMPY, SCIPY, PANDAS, SEABORN
Machine learning	TensorFlow, Keras, Scikit-learn
Computer vision	OpenCV
Natural language processing	NLTK, spaCy

Platform independence

Platform independence refers to a programming language or framework allowing developers to implement things on one machine and use them on another machine without any (or with only minimal) changes. One key to Python's popularity is that it's a platform independent language. Python is supported by many platforms including Linux, Windows, and macOS. Python code can be used to create standalone executable programs for most common operating systems, which means that Python software can be easily distributed and used on those operating systems without a Python interpreter.

What's more, developers usually use services such as Google or Amazon for their computing needs. However, you can often find companies and data scientists who use their own machines with powerful Graphics Processing Units (GPUs) to train their ML models. And the fact that Python is platform independent makes this training a lot cheaper and easier.

Great community and popularity

In the Developer Survey 2018 by Stack Overflow, Python was among the top 10 most popular programming languages, which ultimately means that you can find and hire a development company with the necessary skill set to build your AI-based project.If you look closely at the image below, you'll see that Python is the language that people Google more than any other.

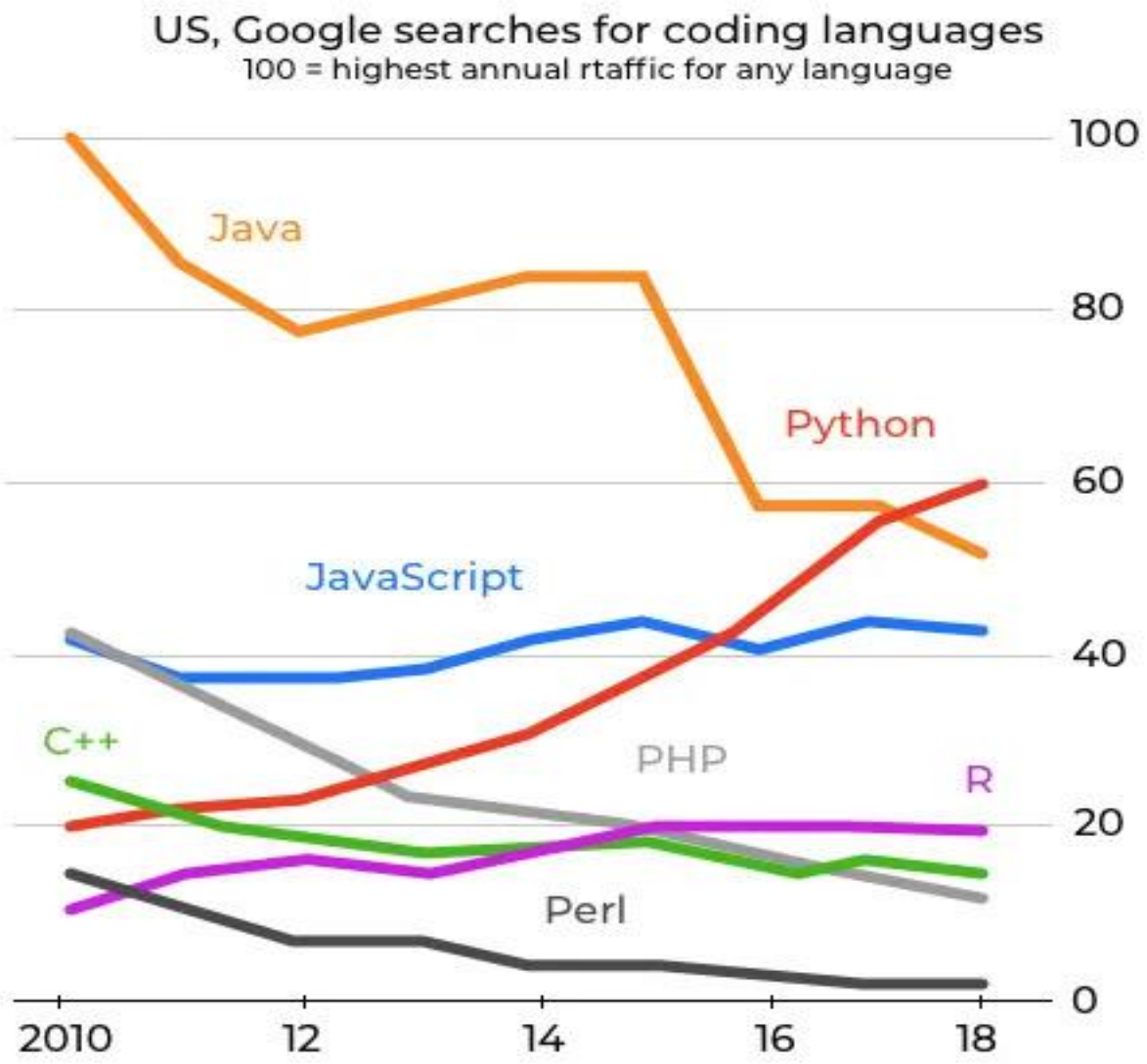


Figure 2: Google search statistics

In the Python Developers Survey 2017, we observe that Python is commonly used for web development. At first glance, web development prevails, accounting for over 26% of the use cases shown in the image below. However, if you combine data science and machine learning, they make up a stunning 27%.

2.4 Python vs other machine learning languages.

The other machine learning programming alternatives are:-

R

R is generally applied when you need to analyze and manipulate data for statistical purposes. R has packages such as Gmodels, Class, Tm, and RODBC that are commonly used for building machine

learning projects. These packages allow developers to implement machine learning algorithms without extra hassle and let them quickly implement business logic.R was created by statisticians to meet their needs. This language can give you in-depth statistical analysis whether you're handling data from an IoT device or analyzing financial models.

What's more, if your task requires high-quality graphs and charts, you may want to use R. With ggplot2, ggvis,googleVis, Shiny, rCharts, and other packages, R's capabilities are greatly extended, helping you turn visuals into interactive web apps.Compared to Python, R has a reputation for being slow and lagging when it comes to large-scale data products. It's better to use Python or Java, with its flexibility, for actual product development.

Scala

Scala is invaluable when it comes to big data. It offers data scientists an array of tools such as Saddle, Scalalab, and Breeze. Scala has great concurrency support, which helps with processing large amounts of data. Since Scala runs on the JVM, it goes beyond all limits hand in hand with Hadoop, an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems. Despite fewer machine learning tools compared to Python and R, Scala is highly maintainable.

Julia

If you need to build a solution for high-performance computing and analysis, you might want to consider Julia. Julia has a similar syntax to Python and was designed to handle numerical computing tasks. Julia provides support for deep learning via the TensorFlow.jl wrapper and the Mocha framework.

However, the language is not supported by many libraries and doesn't yet have a strong community like Python because it's relatively new.

Java

Another language worth mentioning is Java. Java is object-oriented, portable, maintainable, and transparent. It's supported by numerous libraries such as WEKA and Rapidminer.

Java is widespread when it comes to natural language processing, search algorithms, and neural networks. It allows you to quickly build large-scale systems with excellent performance.

But if you want to perform statistical modeling and visualization, then Java is the last language you want to use. Even though there are some Java packages that support statistical modeling and visualization, they aren't sufficient. Python, on the other hand, has advanced tools that are well supported by the community.

2.5 Python Libraries Used In Machine learning

1) NumPy

NumPy is a well known general-purpose array-processing package. An extensive collection of high complexity mathematical functions make NumPy powerful to process large multi-dimensional arrays and matrices. NumPy is very useful for handling linear algebra, Fourier transforms, and random numbers. Other libraries like TensorFlow uses NumPy at the backend for manipulating tensors.

2) Scikit-learn

Scikit-learn has a wide range of supervised and unsupervised learning algorithms that works on a consistent interface in Python. The library can also be used for data-mining and data analysis. The main machine learning functions that the Scikit-learn library can handle are classification, regression, clustering, dimensionality reduction, model selection, and preprocessing**.**

3) Pandas

Pandas are turning up to be the most popular Python library that is used for data analysis with support for fast, flexible, and expressive data structures designed to work on both “relational” or “labeled” data. Pandas today is an inevitable library for solving practical, real-world data analysis in Python. Pandas is highly stable, providing highly optimized performance. The backend code is purely written in C or Python.

4) Matplotlib

Matplotlib is a data visualization library that is used for 2D plotting to produce publication-quality image plots and figures in a variety of formats. The library helps to generate histograms, plots, error charts, scatter plots, bar charts with just a few lines of code.

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

5)Seaborn

Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

CHAPTER – 3

CREATING UI USING DJANGO

3.1 Introduction

Django is a high-level Python Web framework encouraging rapid development and pragmatic, clean design. A web application framework is a toolkit of components all web applications need. The goal here is to allow developers to instead of implementing the same solutions over and over again, focus on the parts of their application that are new and unique to their project. In fact, Django is much more fully-featured than many other frameworks out there. It takes care of a lot of the hassle of Web development, letting you focus on writing your application without any need to reinvent the wheel. It's free and open source. Additionally, the Django framework enables you to model your domain and code classes, and before you know it, you already have an ORM. The Django community, like the Python community, contributes a whole lot of useful packages and utilities for use by the wider world. Django is of course, great for getting started and surprisingly enough it's great when it comes to scaling, too. Django, at its heart, is a series of components, wired up and ready-to-go by default. Now, since these components are decoupled, that is not dependent on each other, they can be unplugged and replaced as and when we require more specific solutions.

3.2 Why Django?

1. **Easy to Use**

Django uses Python programming language which is a popular language in 2015 and now most choosing language by programmers who are learning to code and applications of the Django framework is widely used as it is free and open-source, developed and maintained by a large community of developers. It means we can find answers to the problems easily using Google.

2. **It's fast and simple**

One of Django's main goals is to simplify work for developers. To do that, the Django framework uses:

- The principles of rapid development, which means developers can do more than one iteration at a time without starting the whole schedule from scratch
- DRY philosophy — Don't Repeat Yourself — which means developers can reuse existing code and focus on the unique one.

3. **Excellent Documentation for real-world application**

Applications of Django have one of the best documentation for its framework to develop different kinds of real-world applications whereas many other frameworks used an alphabetical list of modules, attributes, and methods. This is very useful for quick reference for developers when we had confused between two methods or modules but not for freshers who are learning for the first time. It's a difficult task for Django developers to maintain the documentation quality as it is one of the best open-source documentation for any framework.

4. **It's secure**

Security is also a high priority for Django. It has one of the best out-of-the-box security systems out there, and it helps developers avoid common security issues, including

- clickjacking,
- cross-site scripting
- SQL injection.

Django promptly releases new security patches. It's usually the first one to respond to vulnerabilities and alert other frameworks.

5. **It suits any web application project**

With Django, you can tackle projects of any size and capacity, whether it's a simple website or a high-load web application. Why use Django for your project? Because:

- It's fully loaded with extras and scalable, so you can make applications that handle heavy traffic and large volumes of information
- It is cross-platform, meaning that your project can be based on Mac, Linux or PC
- It works with most major databases and allows using a database that is more suitable in a particular project, or even multiple databases at the same time

3.3 Setting up project environment.

Firstly download and install python,to ensure it's fully downloaded, open up your terminal and type in python

An interactive shell shows up:

Python 2.7.12 (default, Nov 19 2016, 06:48:10)

[GCC 5.4.0 20160609] on linux2

Type "help", "copyright", "credits" or "license" for more information.

>>>

To exit, type in ctrl + z.

To have a more neater arrangement, it's always advisable to create a directory for your projects

mkdir folder_name

Then cd into the project with: cd folder_name (every other step will be carried out while inside this folder)

Set Up Your Virtual Environment:

Next thing we need to do, is set up our virtual environment, a virtual environment helps you run several versions of python/django right on your machine (e.g you could have two different python/django projects running on different versions, to avoid them clashing and to give you room to run them both without errors, the virtual environment comes to your rescue. One virtual environment = one python/django version).It's strongly advised to always use a virtual environment.

To set up our virtual environment, we'll be using python's package manager pip to do the installation, type in:

pip install virtualenv

After installation,it's time to create a virtual environment that would enable us use a preferred django version of our choice:

virtualenv env_name

Note: env_name should be replaced with the preferred name of your environment. (I like to name my environments with the django version installed in it for easier recognition).

Activating Virtual Environment:

To activate our virtual environment for linux/Mac OS:

source env_name/bin/activate

For windows:

env_name/script
activate

Install Django:

Now it's time to install django on to our machine:

pip install django==1.8

Using ==1.8 only gives a direction to django about the particular version you want to install, in this case

version 1.8. To just go ahead and download the latest version, input pip install django .

Starting A project:

Now we have django up and running, it’s time to start up our . in our command line, type in :

django-admin.py startproject project_name

Note: project_name = name of your project . In this case, we’ll work with mask_off as our project name.

This creates a sub-folder with the name mask_off and a skeleton structure of

mask_off
├─mask_off
│ ├── __init__.py
│ ├── settings.py
│ ├── urls.py
│ └── wsgi.py
└── manage.py

Django gives us more ease by creating the above files:

1) the __init__.py helps python treat the directories as containing packages; so as to prevent directories with a common name, from unintentionally hiding valid modules that occur later (deeper) on the module search path.In most cases, it’s usually an empty file.

2) The settings.py file contains all settings your project requires, as we progress, we’ll visit this file often.

3) The WSGI (Web Server Gateway Interface) acts as the interface our web server uses to interact with our web application. Read more about it here.

Run Server:

There’s no fun thing as that of visiting your own webpage, so let’s run our server which also generate a link for us to view our webpage

python manage.py runserver

This shows up…

python manage.py runserver
Performing system checks...System check identified no issues (0 silenced).You have unapplied migrations; your app may not work properly until they are applied.
Run 'python manage.py migrate' to apply them.July 12, 2017 - 15:19:01
Django version 1.8, using settings 'mask_off.settings'
Starting development server at http://127.0.0.1:8000/

Quit the server with CONTROL-C.

Notice the warning message about having unapplied migrations? Now let’s do a small but very important talk about migrations;

Sqlite:

SQLite is a file-based SQL database system.it is the default for Django. It should not be used in production since it is usually slow.

```
#myapp/settings/settings.py

DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.sqlite3',
        'NAME': 'db/development.sqlite3',
        'USER': '',
        'PASSWORD': '',
        'HOST': '',
        'PORT': '',
    },
}
```

Making Migrations:

Migrations helps us make changes to our database schema without losing any data, each time we create a new model or make changes to a current one and run migrations, it helps update our database tables with the schemas without having to go through all the stress of dragging and recreating the database ourselves.

To make our migration:

python manage.py migrate

An output of this sort should show up:

Operations to perform:

Synchronize unmigrated apps: staticfiles, messages

Apply all migrations: admin, contenttypes, auth, sessions

Synchronizing apps without migrations:

Creating tables...

Running deferred SQL...

Installing custom SQL...

Running migrations:

Rendering model states... DONE

Applying contenttypes.0001_initial... OK

Applying auth.0001_initial... OK

Applying admin.0001_initial... OK

Applying contenttypes.0002_remove_content_type_name... OK

Applying auth.0002_alter_permission_name_max_length... OK

Applying auth.0003_alter_user_email_max_length... OK

Applying auth.0004_alter_user_username_opts... OK

Applying auth.0005_alter_user_last_login_null... OK

Applying auth.0006_require_contenttypes_0002... OK

Applying sessions.0001_initial... OK

This implies a successful migration, Now we can successfully run our server with no issues python manage.py runserver . We get a success message on our webpage like the one below

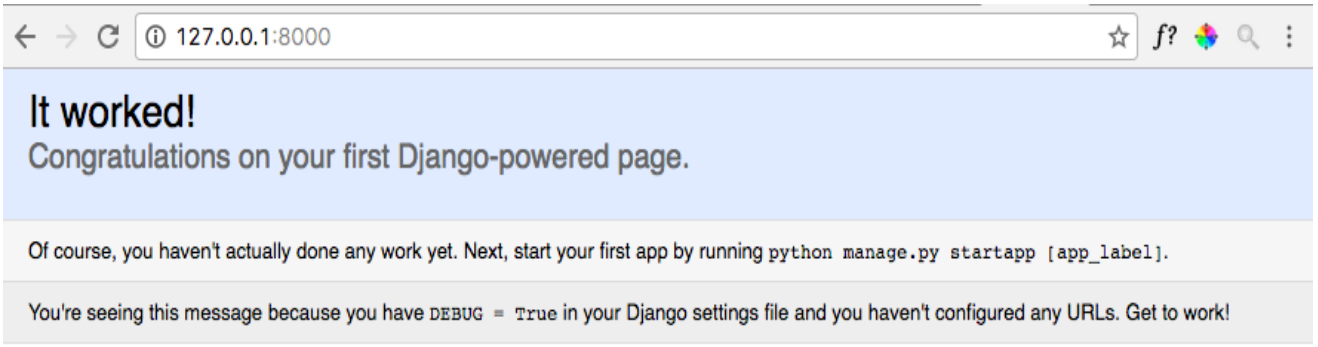


Figure 3:Django webpage

Now we have our server running in port 8000.

3.4 Django Architecture

Django is based on MVT (Model-View-Template) architecture. MVT is a software design pattern for developing a web application.

MVT Structure has the following three parts –

Model: Model is going to act as the interface of your data. It is responsible for maintaining data. It is the logical data structure behind the entire application and is represented by a database (generally relational databases such as MySql, Postgres).

View: The View is the user interface — what you see in your browser when you render a website. It is represented by HTML/CSS/Javascript and Jinja files.

Template: A template consists of static parts of the desired HTML output as well as some special syntax describing how dynamic content will be inserted.

Benefits of Django Architecture:-

The Django Framework is based on this architecture and it actually communicates between all these three components without needing to write complex code. That's why Django is gaining popularity.

This architecture in Django has various advantages like:

1. Rapid Development

Actually, this Django architecture that separates in different components makes it easy for multiple developers to work on different aspects of the same application simultaneously. That is also one of the features of Django.

2. Loosely Coupled

This architecture of Django has different components which require each other at certain parts of the application, at every instant, that increases the security of the overall website. As the model file will now only save on our server rather than saving on the webpage.

3. Ease of Modification

This is an important aspect of development as there are different components in Django architecture. If there is a change in different components, we don't have to change it in other components.

This is actually one of the special features of Django, as here it provides us with much more adaptability of our website than other frameworks.

CHAPTER – 4

PROJECT DETAILS

4.1 Introduction:-

Nowadays, people face various diseases due to the environmental condition and their living habits and due to this early prediction of diseases very important. But, the accurate prediction of disease based on symptoms, becomes too much difficult for a doctor to predict and diagnose the disease. Therefore, the correct prediction of disease is almost a challenging task. To overcome this problem, machine learning computation plays a vital role in disease prediction. Every year medical science provides voluminous amount of data which is increasing exponentially. The development and exploitation of several prominent machine learning techniques has led to the utilization of such medical data in better way by extracting useful information from specified data to help healthcare communities and biomedical fields in an effective way. Machine learning finds hidden pattern of information from the huge amount of medical data. Anyone, who wants to check breast cancer, diabetes, heart disease, kidney disease, lung cancer can easily use the algorithms and predict the severity of diseases effectively. We have stretched our work on this aspect and use of different machine learning classifier for disease prediction. We have classified the some of the widely spread diseases perfectly using machine learning using python tool. Our study primarily begins from breast cancer and lung cancer dataset of UCI machine learning repository. In a world the mortality rate of breast cancer, lung cancer is quite high and death rate due to heart disease is also high in number. In addition to it, kidney diseases and diabetes are also life threating diseases for human beings. Therefore, effective measures and proper data computation is indeed essentially to analyse the disease dataset accurately and promptly. We perform our data analysis on diseases such as :

1. Lung Cancer,
2. Breast cancer,
3. Diabetes,
4. Heart Disease,
5. Kidney Disease.

Cancer is a fatal illness often caused by genetic disorder aggregation and a variety of pathological changes. Cancerous cells are abnormal areas often growing in any part of human body that are life-threatening. Cancer is also known as tumor. Cancer must be detected quickly and correctly in the initial stage such that better diagnosis can be possible. Even though modality has different considerations, such as complicated history, improper diagnostics and treatment which are primary causes of deaths. Lung cancer and breast cancer are now embracing as danger for human kind. Lung cancer occurs when cells divide in the lungs uncontrollably. This causes tumor to grow. These can

reduce a person's ability to breathe and spread to other parts of the body. Lung cancer is now placed in third position due to high mortality. There are several reason and factors that cause lung cancer but some of the study says that smoking, toxin absorption and inhaled chemicals are primary cause of lung cancer. Lung cancer can be fatal, but effective diagnoses and treatments are improving the patient health condition. The two main types of lung cancer are small cell lung cancer and non-small cell lung cancer. Non-small cell lung cancer is more common than small cell lung cancer. Similarly, breast cancers are cause of highest death rate in female in the world. Breast cancer occurs in breast cells. Typically, the cancer formation happens either in the lobules or the ducts of the breast. Lobules are the glands that produce milk, and ducts are the pathways that bring the milk from the glands to the nipple. Cancer can also occur in the fatty tissue or the fibrous connective tissue within your breast. The uncontrolled cancer cells often invade other healthy breast tissue and can travel to the lymph nodes under the arms. The lymph nodes are a primary pathway that helps the cancer cells to move to other parts of the body. The most common kinds of breast cancer are Invasive ductal carcinoma: - The cancer cells grow outside the ducts into other parts of the breast tissue. Invasive cancer cells can also spread, or metastasize, to other parts of the body. Invasive lobular carcinoma: - Cancer cells spread from the lobules to the breast tissues that are close by. These invasive cancer cells can also spread to other parts of the body. Peoples in all races are mostly affected by diabetes which is well known as silent poison. Diabetes is a condition that impairs the body's ability to process blood glucose, otherwise known as blood sugar. Three major diabetes types can develop: Type 1, type 2, and gestational diabetes.Type I diabetes: Also known as juvenile diabetes, this type occurs when the body fails to produce insulin. People with type 1 diabetes are insulin-dependent, which means they must take artificial insulin daily to stay alive. Type 2 diabetes: It affects the way the body uses insulin. While the body still makes insulin, unlike in type I, the cells in the body do not respond to it as effectively as they once did. This is the most common type of diabetes, according to the National Institute of Diabetes and Digestive and Kidney Diseases, and it has close link with obesity. Gestational diabetes: This type occurs in women during pregnancy when the body can become less sensitive to insulin. Gestational diabetes does not occur in all women and usually resolves after giving birth. Heartdisease:- Heart disease, such as coronary heart disease, heart attack, congestive heart failure, and congenital heart disease, are the leading cause of death for men and women in world. Prevention includes quitting smoking, lowering cholesterol, controlling high blood pressure, maintaining a healthy weight, and exercising.Unlike cardiovascular disease, which includes problems with the entire circulatory system, heart disease affects only the heart. Kidney disease:- Kidney disease can affect your body's ability to clean your blood, filter extra water out of your blood, and help control your blood pressure. It can also affect red blood cell production and vitamin D metabolism.When your kidneys are damaged, waste products and fluid can build up in your body. That can cause swelling in your ankles, nausea weakness, poor sleep, and shortness of breath.

Without treatment, the damage can get worse and your kidneys may eventually stop working. That's serious, and it can be life-threatening.

Our study analyses patient's dataset, addresses effective computation and predict the disease using supervised machine learning techniques. The study highlights how cancer diagnosis, cure process is assisted using supervised machine learning techniques. Our state of art focuses a detail analysis of benchmark datasetsin terms of accuracy, sensitivity, specificity, false-positive metrics comparison with high precision. Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Even though, it is evident that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice. In this work, we present a review of recent ML approaches employed in modelling of cancer progression. The predictive models discussed here are based on various supervised ML techniques as well as on different input features and data samples. Given the growing trend on the application of ML methods in cancer research, we present here the most recent publications that employ these techniques as an aim to model cancer risk or patient outcome.

4.2 Diseases Description.

The diseases that are covered :-

1. Lung Cancer
2. Breast cancer
3. Diabetes
4. Heart Disease
5. Kidney Disease

- **Lung Cancer :-**

Lung cancer occurs when cells divide in the lungs uncontrollably. This causes tumors to grow. These

can reduce a person's ability to breathe and spread to other parts of the body.Lung cancer is the third most common cancer and the main cause of cancer-related death in the United States. It is most common in males, and in the U.S., Black males are around 15% more likely to develop it than white males. Smoking is a major risk factor, though not everyone who develops lung cancer has a history of smoking. Lung cancer can be fatal, but effective diagnoses and treatments are improving the outlook. The two main types of lung cancer are small cell lung cancer and non-small cell lung cancer, depending on how they appear under a microscope. Non-small cell lung cancer is more common than small cell lung cancer. Anyone can develop lung cancer, but cigarette smoking and having exposure to smoke, inhaled chemicals, or other toxins can increase the risk.

- **Breast cancer :-**

Breast cancer is cancer that develops in breast cells. Typically, the cancer forms in either the lobules or the ducts of the breast. Lobules are the glands that produce milk, and ducts are the pathways that bring the milk from the glands to the nipple. Cancer can also occur in the fatty tissue or the fibrous connective tissue within your breast. The uncontrolled cancer cells often invade other healthy breast tissue and can travel to the lymph nodes under the arms. The lymph nodes are a primary pathway that help the cancer cells move to other parts of the body.

The most common kinds of breast cancer are—

- Invasive ductal carcinoma :- The cancer cells grow outside the ducts into other parts of the breast tissue. Invasive cancer cells can also spread, or metastasize, to other parts of the body.
- Invasive lobular carcinoma:- Cancer cells spread from the lobules to the breast tissues that are close by. These invasive cancer cells can also spread to other parts of the body.

- **Diabetes:-**

Diabetes is a condition that impairs the body's ability to process blood glucose, otherwise known as blood sugar.

Three major diabetes types can develop: Type 1, type 2, and gestational diabetes.

Type I diabetes: Also known as juvenile diabetes, this type occurs when the body fails to produce insulin. People with type I diabetes are insulin-dependent, which means they must take artificial insulin daily to stay alive.

Type 2 diabetes: Type 2 diabetes affects the way the body uses insulin. While the body still makes insulin, unlike in type I, the cells in the body do not respond to it as effectively as they once did. This is the most common type of diabetes, according to the National Institute of Diabetes and Digestive and Kidney Diseases, and it has strong links with obesity.

Gestational diabetes: This type occurs in women during pregnancy when the body can become less

sensitive to insulin. Gestational diabetes does not occur in all women and usually resolves after giving birth

- **Heart disease:-**

Heart disease, such as coronary heart disease, heart attack, congestive heart failure, and congenital heart disease, is the leading cause of death for men and women in the U.S. Prevention includes quitting smoking, lowering cholesterol, controlling high blood pressure, maintaining a healthy weight, and exercising. Unlike cardiovascular disease, which includes problems with the entire circulatory system, heart disease affects only the heart.

The following symptoms may indicate a heart problem:

- angina, or chest pain
- difficulty breathing
- fatigue and light headedness
- swelling due to fluid retention, or edema

- **Kidney disease:-**

Kidney disease can affect your body's ability to clean your blood, filter extra water out of your blood, and help control your blood pressure. It can also affect red blood cell production and vitamin D metabolism needed for bone health. You're born with two kidneys. They're on either side of your spine, just above your waist.When your kidneys are damaged, waste products and fluid can build up in your body. That can cause swelling in your ankles, nausea, weakness, poor sleep, and shortness of breath. Without treatment, the damage can get worse and your kidneys may eventually stop working. That's serious, and it can be life-threatening.

4.3 Data flow diagram

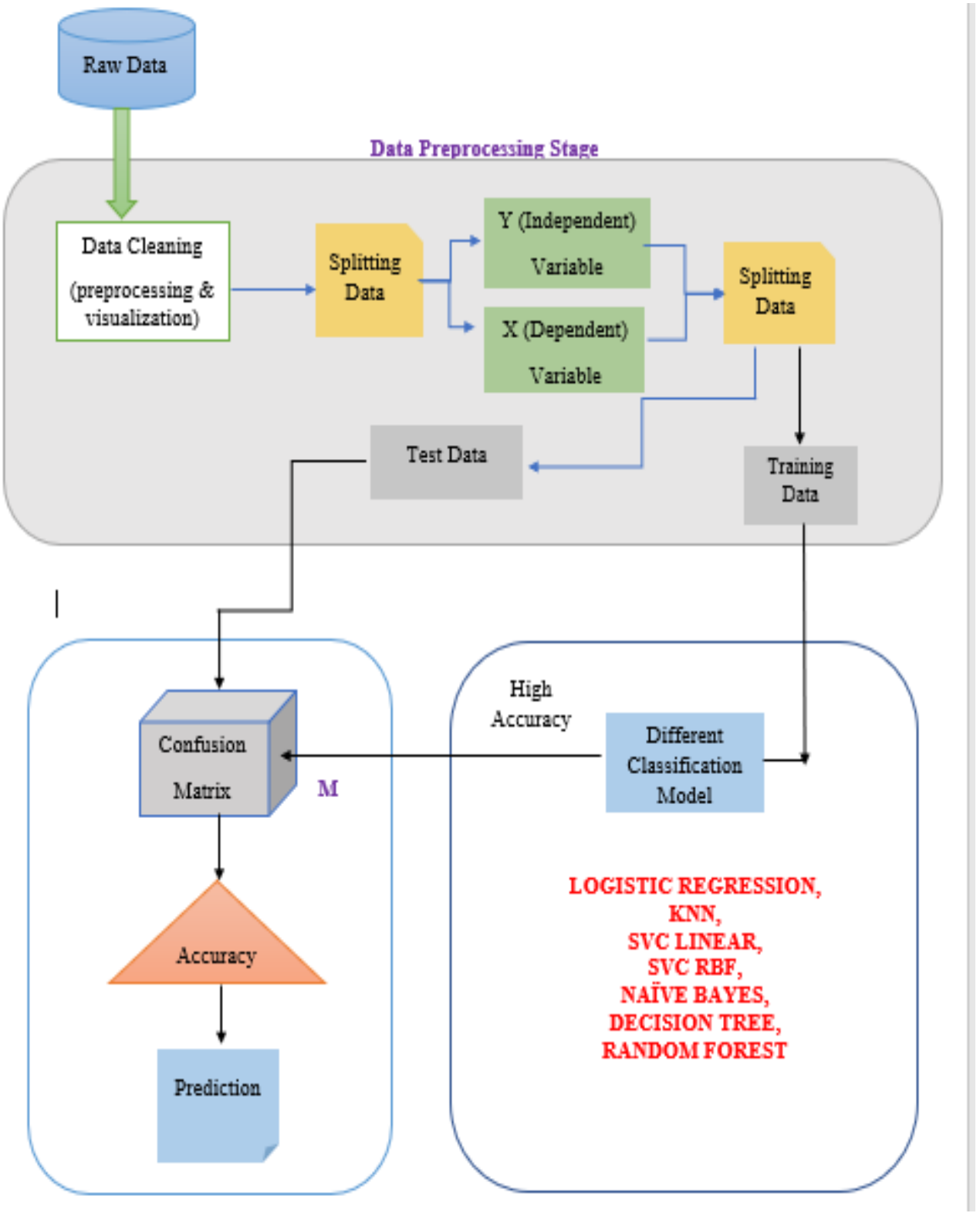


Figure 4: Data flow diagram

4.4 Dataset description

Datasets are downloaded from Kaggle and UCI machine learning repository. Data set comprises of some missing values which degrades the algorithm performances so data cleaning is required. Datasets used here are: -

Lung cancer dataset: -

This dataset contains of 60 rows and 7 columns namely ‘Name’, ’Surname’, ’Age’,’Smokes’, ’AreaQ’, ’Alcohol’ and ‘Result’. ‘Result’ column is the output in the form of either one or zero which means either the person has lung cancer or not. first two attribute represents patient name and other four attributes are numeric data types and last column is class label. Class labels are described as ‘0’ for absence of lung cancer and ‘1’ for presence of lung cancer.

Table1: Demonstration of Lung Cancer Dataset

Sl. no	Name of the attribute	Datatype	Min range	Max range
1	Age	Numeric	18	77
2	Smokes	Numeric	0	34
3	AreaQ	Numeric	1	10
4	Alcohol	Numeric	0	8
5	Result	Numeric	0	1

Breast cancer dataset: -

Data Set Characteristics:

:Number of Instances: 569

:Number of Attributes: 30 numeric, predictive attributes and the class

:Attribute Information:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

- class:
  - WDBC-Malignant
  - WDBC-Benign

:Class Distribution: 212 - Malignant, 357 - Benign

Summary Statistics:

Table 2: Demonstration of Breast cancer Dataset

Sl. no	Name of the attribute	Datatype	Min range	Max range
1	Radius	Numeric	6.981	28.11
2	Texture	Numeric	9.71	39.28
3	Perimeter	Numeric	43.79	188.5
4	Area	Numeric	143.5	2501.0
5	Smoothness	Numeric	0.053	0.163
6	Compactness	Numeric	0.019	0.345
7	Concavity	Numeric	0.0	0.427
8	Symmetry	Numeric	0.106	0.304
9	Fractal dimension	Numeric	0.05	0.097

Pima Indians Diabetes database (PIDD): -

Pima Indian Diabetes dataset consists of 768 rows and 9 columns. Predictor variables are described as 'No of pregnancies', 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin', 'BMI', 'Diabetes pedigree function', 'Age', 'Outcome'. 'Outcome' represents class label. Class value '1' is interpreted as 'testing positive for diabetes'. Diabetes patients in total is 268 and non-diabetic patentsare500 in number.

Table 3: Demonstration of PIDD Dataset

Sl. no	Name of the attribute	Datatype	Min range	Max range
1	Pregnancies	Numeric	0	17
2	Glucose	Numeric	0	199
3	BloodPressure	Numeric	0	122
4	SkinThickness	Numeric	0	99
5	Insulin	Numeric	0	846
6	BMI	Numeric	0.0	67.1
7	Diabetes Pedigree Function	Numeric	0.078	2.42
8	Age	Numeric	21	81
9	Outcome	Numeric	0	1

Heart disease dataset: -

This dataset contains of 303 rows and 14 columns namely 'Age','Sex', 'Cp','Trtbps', 'Chol', 'Fbs', 'Restecg','Thalachh ','Exang ','Oldpeak ','Slope ','Ca ' ,'Thal ','Target' . where 'Target' column is the

output in the form of either one or zero which means either the person has heart disease or not. There are six records which has missing attributes. Target has class label 0 (164 in number)signifies absence of heart disease and class label 1,2,3,4 represents presence of heart disease.

Table 4: Demonstration of heart disease Dataset

Sl. No	Name of the attribute	Datatype	Min range	Max range
1	Age	Numeric	29	77
2	Sex	Numeric	0	1
3	Cp	Numeric	0	3
4	Trtbps	Numeric	94	200
5	Chol	Numeric	126	564
6	Fbs	Numeric	0	1
7	Restecg	Numeric	0	2
8	Thalachh	Numeric	71	202
9	Exang	Numeric	0	1
10	Oldpeak	Numeric	0	6.2
11	Slope	Numeric	1	3
12	Ca	Numeric	0	3
13	Thal	Numeric	3	7
14	Target	Numeric	0	4

Kidney disease: -

This dataset contains of 401 rows and 25columns namely ‘id’, ‘Age’, ‘Blood pressure(Bp)’, ‘Specific Gravity (Sg)’, ‘Albumin(Al)’, ‘Sugar(Su)’, ‘Red blood cells (Rbc)’, ’ Puscell (Pc)’,’ Puscell clumps (Pcc)’,’ Bacteria (Ba)’,’ Blood glucose random (Bgr)’,’ Blood urea (Bu)’,’ Serum creatinine (Sc)’,’ Sodium(Sod)’,’ Potassium(Pot)’,’ Haemoglobin (Hemo)’,’ Packed cell volume (PCV)’,’ White blood cell count(Wc)’,’ Red blood cell count (Rc)’, ‘Hypertension (Htn)’,’ Diabetes mellitus(Dm)’, ‘Coronary artery disease (Cad)’, ‘Appetite (Appet)’, Pedal edema (Pe)’, ‘Anaemia (Ane)’ and ‘class’. ‘Class’ column is the output in the form of ‘ckd’ and ‘notckd’ where ‘ckd’(250 in numbers) refers to chronic kidney disease and ‘notckd’ (150 in numbers) refers to nonchronic kidney disease. Here 24 numbers of missing data.

Table 5: Demonstration of kidney disease Dataset

S.l. no	Name of the attribute	Datatype	Min range	Max range
1	Age	Numeric	2	90
2	Blood pressure	Numeric	50.0	180.0
3	Specific Gravity	Nominal	1.005	1.025
4	Albumin	Nominal	0.0	5.0
5	sugar	Nominal	0.0	5.0
6	Red blood cells	Nominal	0.0	1.0
7	Puscell	Nominal	0.0	1.0
8	Puscell clumps	Nominal	0.0	1.0
9	Bacteria	Nominal	0.0	1.0
10	Blood glucose random	Numeric	22.0	490.0
11	Blood urea	Numeric	1.5	391.0

12	Serum creatinine	Numeric	0.4	76
13	Sodium	Numeric	4.5	163.0
14	Potassium	Numeric	2.5	47.0
15	Haemoglobin	Numeric	3.1	17.8
16	Packed cell volume	Numeric	9	54
17	White blood cell count	Numeric	2200	26400
18	Red blood cell count	Numeric	2.1	8.0
19	Hypertension	Nominal	0.0	1.0
20	Diabetes mellitus	Nominal	0.0	1.0
21	Coronary artery disease	Nominal	0.0	1.0
22	Appetite	Nominal	0.0	1.0
23	Pedal edema	Nominal	0.0	1.0
24	Anaemia	Nominal	0.0	1.0
25	Class	Nominal	0	1

4.5 Algorithms And Techniques Used:-

The steps involved are:-

[1] **Data Collection** :-

As we are in the era of big data,data can be accessed from anywhere in the world. Big data is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. The first approach is to gather data because without it machine learning is nothing.

[2] **Data Cleaning** :-

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. So the next approach should be data cleaning. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Data cleaning includes Remove duplicate or irrelevant observations, Fix structural errors, Filter unwanted outliers, Handle missing data etc.

[3] **Data Preparation** :-

Now, preparing of data should be done for which we plot correlogram(graph of correlation matrix),histogram plot and count plot. Correlation matrix is a table showing coefficients between variables. Which helps us to slice our data into dependant and independent variables.Determining cause and effect is one of the most important parts of scientific research. It's essential to know which is the cause – the independent variable – and which is the effect – the dependent variable. Here in ML, when the dependent (or output variable-y) and independent variable(or input variable- x) gives to a Machine Learning model it gives a mapping function y=f (x).

[4] **Splitting into training and test data** :-

Scikit learn library helps to split the dataset into training and test datasets. The main reason behind the

splitting is that when the dataset is split into train and test sets, there will not be enough data in the training dataset for the model to learn an effective mapping of inputs to outputs. There will also not be enough data in the test set to effectively evaluate the model performance. Also this splitting of data prevent our model from overfitting and helps to improve the accuracy of the model. While training the model, data is usually split in the ratio of 80:20 i.e. 80% as training data and rest as testing data. In training data, we feed input as well as output for 80% data. The model learns from training data only. We use different machine learning algorithms(which we will discuss in detail in the next articles) to build our model. By learning, it means that the model will build some logic of its own. Once the model is ready then it is good to be tested. At the time of testing, the input is fed from the remaining 20% data which the model has never seen before, the model will predict some value and we will compare it with actual output and calculate the accuracy.

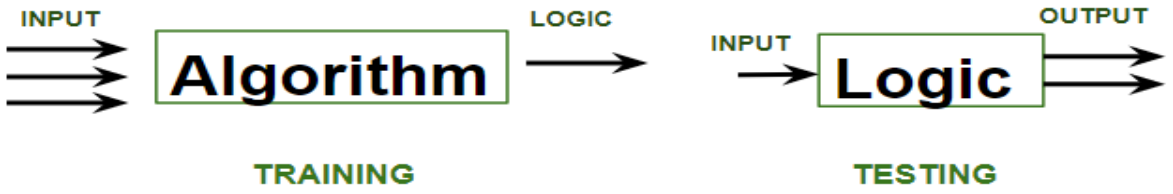


Figure 5: Splitting data

28

[5] **Model Selection** :-

First of all, Machine Learning is divided into 4 types.

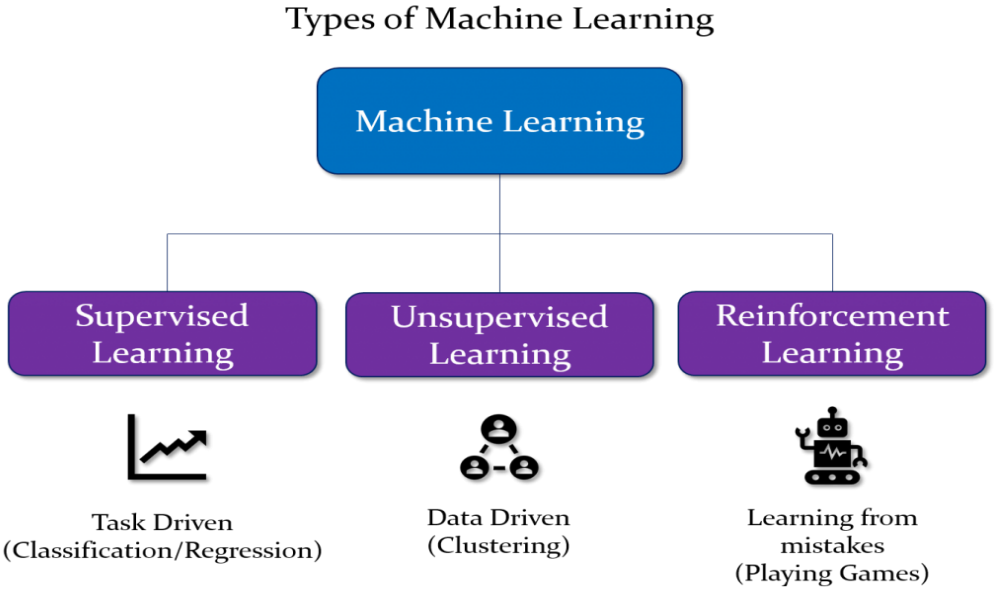


Figure 6: Types of learning

In this project we are working on labelled data, so we consider supervised machine learning approach. In supervised learning, the system tries to learn from the previous data that are given to the system previously and according to it tries to predict the future data.

29

Supervised Learning is of two types as shown below

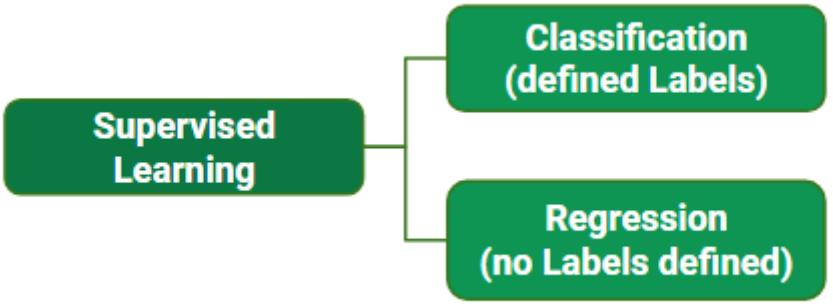


Figure 7: Supervised learning

In our project we used classification rather than regression as here dependent variables are having defined labels(discrete value). For example in above Figure A, Output – Purchased has defined labels i.e. 0 or 1 ; 1 means the customer will purchase and 0 means that customer won't purchase. The goal here is to predict discrete values belonging to a particular class and evaluate on the basis of accuracy.

In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes. Now we are going to train our models and verify which algorithm gives maximum accuracy

The Algorithms used in the project :-

1. Logistic Regression
2. Support Vector Machine (Linear Classifier &RBF Classifier)
3. Naive Bayes
4. Decision Tree
5. Random Forest
6. K-Nearest Neighbour (KNN)

1. **Logistic Regression** :-

It is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome. The outcome is measured with a dichotomous variable meaning it will have only two possible outcomes. Logistic regression is named for the function used at the core of the method, the logistic function.

The name ‘Regression’ here implies that a linear model is fit into the feature space. This algorithm applies a logistic function to a linear combination of features to predict the outcome of a categorical dependent variable based on predictor variables.

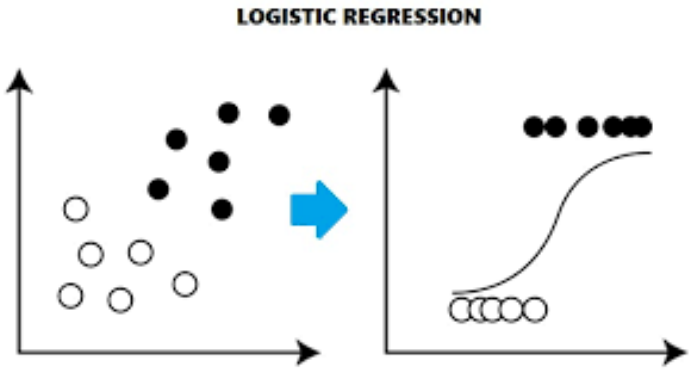


Figure 8:Logistic regression

Mathematically, a logistic regression model predicts P(y=1) as a function of x. Logistic regression can be expressed as: $\log(p(X)/(1-p(X)))=\beta_0+\beta_1X$

Where, the lefthand side is called he logiestor log odds function,andp(x)/(1-p(x))
is called odds.The odds signifies the ratio of probability of success to probability of failure. Therefore in logistic Regression,linear combination of inputs are mapped to the log(odds) -the output being adequate to1.

Advantages : -

Logistic regression is specifically meant for classification, it is useful in understanding how a set of independent variables affect the outcome of the dependent variable.

2. **Support Vector Machine** :-

The support vector machine is a classifier that represents the training data as points in space separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to.

Support Vector Machine is a supervised machine learning algorithm for classification or regression problems where the dataset teaches SVM about the classes so that SVM can classify any new data. It works by classifying the data into different classes by finding a line(hyperplane) which separates the training data set into classes. As there are many such linear hyperplanes, SVM algorithm tries to maximize the distance between the various classes that are involved and this is referred as margin maximization. If

the line that maximizes the distance between the classes is identified, the probability to generalize well to unseen data is increase.

SVM's are classified into two categories:-

- Linear SVM's – In linear SVM's the training data i.e. classifiers are separated by a hyperplane.

Non-Linear SVM's- In non-linear SVM's it is not possible to separate the training data using a hyperplane. For example, the training data for Face detection consists of group of images that are faces and another group of images that are not faces (in other words all other images in the world except faces). Under such conditions, the training data is too complex that it is impossible to find

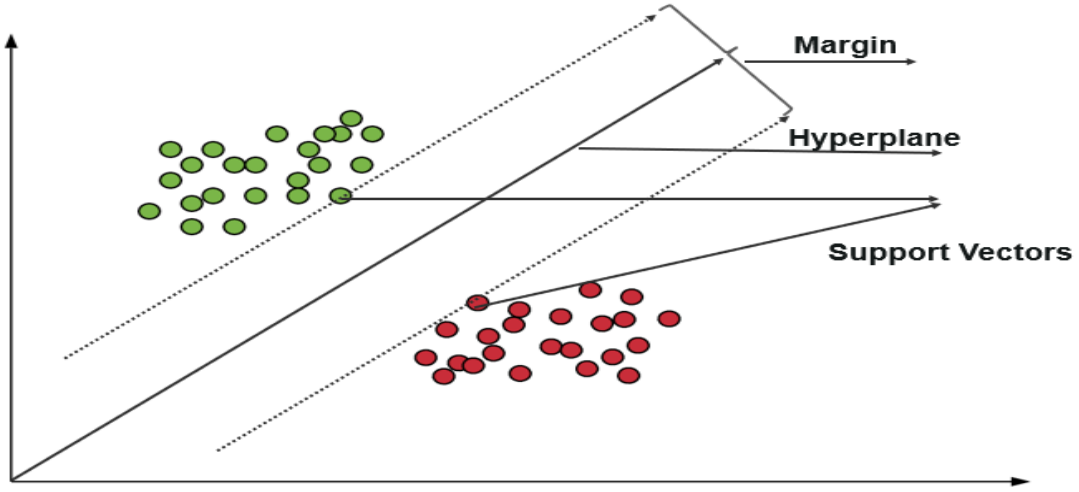


Figure 9:Support vector machine

are presentation for every feature vector. Separating the set of faces linearly from the set of non-face is a complex task.

Advantages:-

- SVM offers best classification performance (accuracy) on the training data
- SVM renders more efficiency for correct classification of the future data
- The best thing about SVM is that it does not make any strong assumptions on data
- It does not over-fit the data.

3. **Naive Bayes** :-

It is a classification algorithm based on Bayes theorem which gives an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Even if the features depend on each other, all of these properties contribute to the probability independently. Naive Bayes model is easy to make and is particularly useful for comparatively large data sets. Even with a simplistic approach, Naive Bayes is known to outperform most of the classification

methods in machine learning. Following is the Bayes theorem to implement the Naive Bayes Theorem.

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \frac{P(x_1, x_2 \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2 \ldots, x_n)} \; for\; 1 < i < k$$

Figure 10:Bayes theorem

Advantages :-

The Naive Bayes classifier requires a small amount of training data to estimate the necessary parameters to get the results. They are extremely fast in nature compared to other classifiers.

4. **Decision Tree** :-

A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision based on certain conditions.

The decision tree algorithm builds the classification model in the form of a tree structure. It utilizes the if-then rules which are equally exhaustive and mutually exclusive in classification. The process goes on with breaking down the data into smaller structures and eventually associating it with an incremental decision tree. The final structure looks like a tree with nodes and leaves. The rules are learned sequentially using

the training data one at a time. Each time a rule is learned, the tuples covering the rules are removed. The process continues on the training set until the termination point is met.

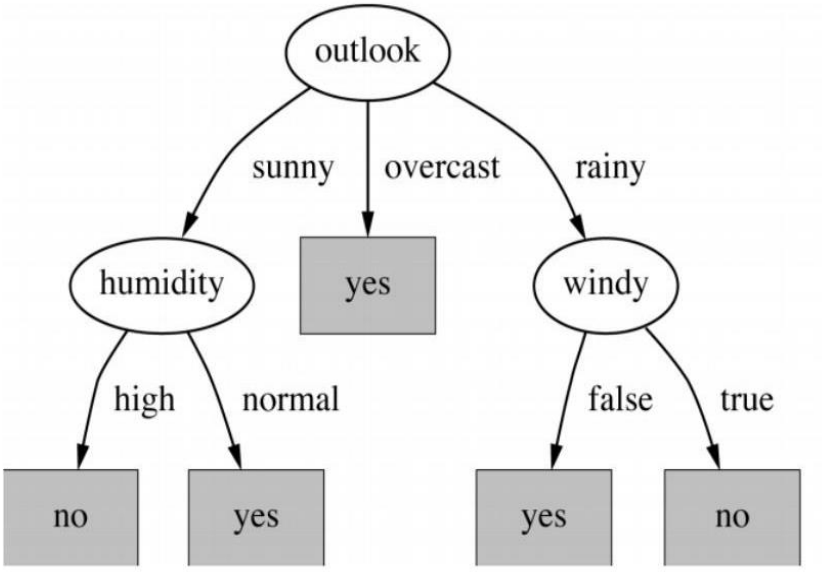


Figure 11: Decision tree example

The tree is constructed in a top-down recursive divide and conquer approach. A decision node will have two or more branches and a leaf represents a classification or decision. The topmost node in the decision tree that corresponds to the best predictor is called the root node, and the best thing about a decision tree is that it can handle both categorical and numerical data.

5. **Random Forest** :-

Random decision trees or random forest are an ensemble learning method for classification, regression, etc. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction(regression) of the individual trees.

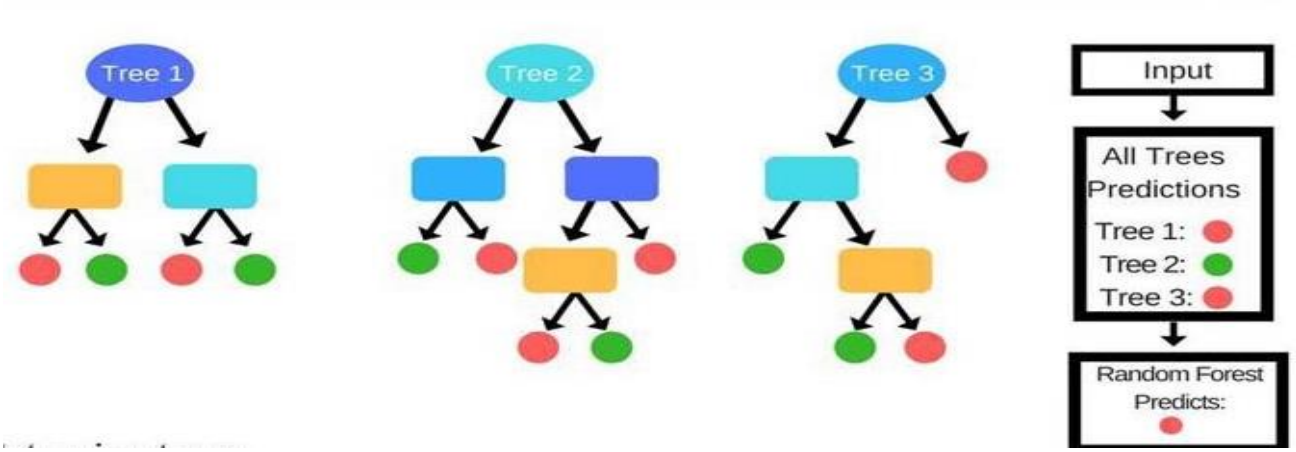


Figure 12: Random forest example

How it Works ?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

A random forest is a meta-estimator that fits a number of trees on various subsamples of data sets and then uses an average to improve the accuracy in the model's predictive nature. The sub-sample size is always the same as that of the original input size but the samples are often drawn with replacements.

Advantages:-
it is more accurate than the decision trees due to the reduction in the over-fitting.

6. **K-Nearest Neighbour (KNN)** :-

K Nearest Neighbor (KNN) could be a terribly easy, simple to grasp, versatile and one amongst the uppermost machine learning algorithms. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

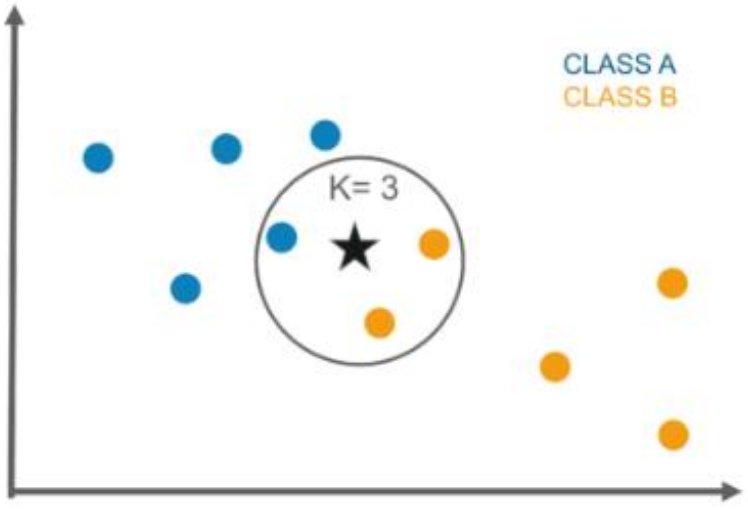


Figure 13: KNN example

The K-NN working can be explained on the basis of the below algorithm:

- Step-1: Select the number K of the neighbours.
- Step-2: Calculate the Euclidean distance of K number of neighbours

- Step-3: Take the K nearest neighbours as per the calculated Euclidean distance.
- Step-4: Among these k neighbours, count the number of the data points in each category.
- Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.
- Step-6: Our model is ready.

Advantages:-

This algorithm is quite simple in its implementation and is robust to noisy training data. Even if the training data is large, it is quite efficient.

The model is now trained using training dataset that we got in the process of data preparation using different machine learning classification algorithms.

4.6 Performance of models.

Training Accuracy: -

Diseases / Algorithms	**Diabetes**	**Lung Cancer**	**Breast Cancer**	**Kidney disease**	**Heart Disease**
Logistic Regression	77.65 %	97.73 %	99.06 %	100%	87.13%
SVM (Linear Classifier)	77.84 %	97.73 %	98.59 %	100%	85.96 %
SVM (RBF Classifier)	83.14%	97.73%	98.59 %	100%	92.98 %
Naive Bayes	77.08 %	97.73%	94.84 %	95.33 %	84.21 %
Decision Tree	100.0 %	100.0 %	100.0 %	100.0 %	100 %
KNN	80.49%	97.73%	97.42 %	98.00%	85.38 %
Random Forest	97.92 %	100.0 %	99.77 %	100.0 %	99.42 %

Table 6: Demonstration of Accuracy for training dataset values

Testing Accuracy: -

After training the data set, our model is tested using test dataset to predict the accuracy and various classification parameter.

Diseases / Algorithms	Diabetes	Lung Cancer	Breast Cancer	Kidney disease	Heart Disease
Logistic Regression	78.98 %	100.0 %	95.8%	100.0%	89.47 %
SVM (Linear Classifier)	79.55 %	100.0 %	97.2%	99.0 %	85.96 %
SVM (RBF Classifier)	75.57%	100.0%	96.5 %	98.0 %	85.96 %
Naive Bayes	76.14%	100.0 %	96.61 %	95.0 %	91.23 %
Decision Tree	62.5%	100.0 %	95.8 %	97.0 %	91.23 %
KNN	75.57 %	100.0 %	95.1%	97.0 %	87.72 %
Random Forest	72.14%	100.0 %	98.6 %	100.0 %	87.72 %

Table 7: Demonstration of Accuracy for testing dataset values

All the highlighted values represent maximum accuracy that we have obtained. So, we have considered SVM (linear classifier) model for diabetes prediction, logistic regression model for lung cancer detection, random forest for breast cancer, logistic regression model for kidney disease prediction and naïve bayes model for heart disease prediction as all these model gives highest accuracy as shown in the above figure. Generally, in confusion matrix Accuracy, Recall, Precision and F-Measure are the key process parameter for classification. Classification accuracy is the measure of number of correct predictions made out from total number of predictions. These parameters depends on some specific outcome. Those are 'TP (True Positive) which is the correctly predicted event values and 'TN (True Negative) is correctly predicted no event values. Similarly False Positive is incorrectly predicted event values and 'FN (False Negative) for incorrectly predicted no event values. Now we are going to calculate the recall,precision,f1-score etc using scikit learn library.

The class labelled 1 is the positive class in our example. The class labelled as 0 is the negative class here. As we can see, the Positive and Negative Actual Values are represented as columns, while the Predicted Values are shown as the rows.

- TP = True Positive – The model predicted the positive class correctly, to be a positive class.
- FP = False Positive – The model predicted the negative class incorrectly, to be a positive class.

- FN = False Negative – The model predicted the positive class incorrectly, to be the negative class.
- TN = True Negative – The model predicted the negative class correctly, to be the negative class.

The testing accuracy of the model can be calculated as:

accuracy = (TP + TN) / (TP + TN + FN + FP)

Recall: Out of all the positive classes, how many instances were identified correctly.

Recall = TP / (TP + FN)

Precision: Out of all the predicted positive instances, how many were predicted correctly.

Precision = TP / (TP + FP)

F-Score: From Precision and Recall, F-Measure is computed and used as metrics sometimes. F – Measure is nothing but the harmonic means of Precision and Recall.

AUC-ROC

AUC–ROC curve is the model selection metric for bi–multi class classification problem. ROC is a probability curve for different classes. ROC tells us how good the model is for distinguishing the given classes, in terms of the predicted probability. A typical ROC curve has False Positive Rate (FPR) on the X-axis and True Positive Rate (TPR) on the Y-axis**.**

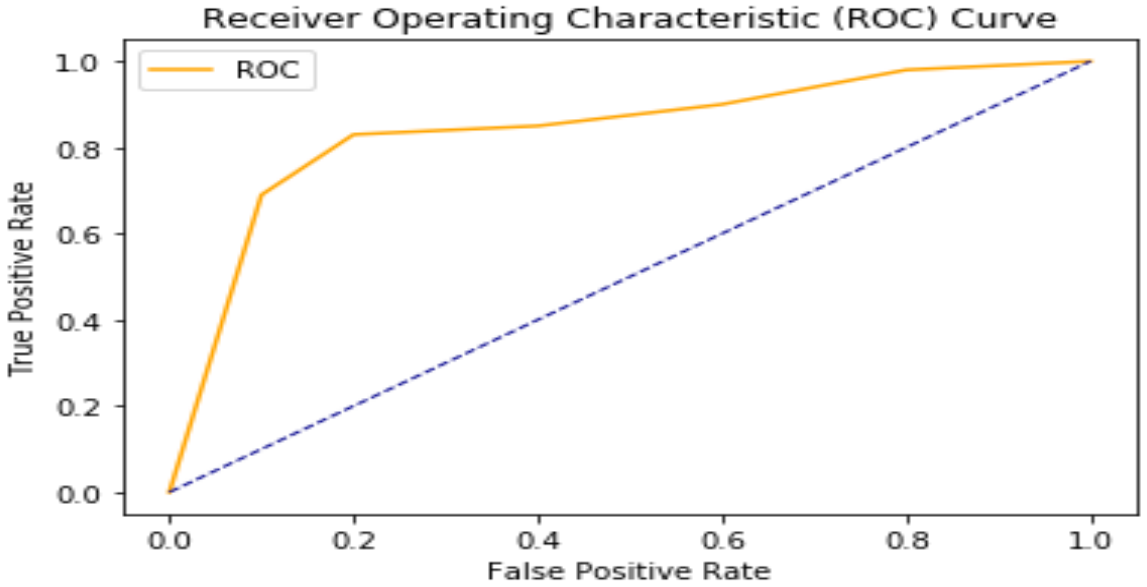


Figure 14:AUC ROC curve Representation

The area covered by the curve is the area between the orange line (ROC) and the axis. This area covered is AUC. The bigger the area covered, the better the machine learning models is at distinguishing the given classes. Ideal value for AUC is 1.

Algorithms	0/1	Precision	Recall	F1-score	ROC AUC Score (%)
Logistic Regression	0 - 1 -	0.98 1.00	1.00 0.98	0.99 0.99	99.10
KNN	0 - 1 -	0.94 1.00	1.00 0.95	0.97 0.97	97.32
SVM (Linear Classifier)	0 - 1 -	0.98 1.00	1.00 0.98	0.99 0.99	99.10
SVM (RBF Classifier)	0 - 1 -	0.96 1.00	1.00 0.96	0.98 0.98	98.21
Naive Bayes	0 - 1 -	0.91 0.98	0.98 0.93	0.95 0.95	95.29
Decision Tree	0 - 1 -	0.98 0.96	0.95 0.98	0.97 0.97	96.83
Random Forest	0 - 1 -	1.00 1.00	1.00 1.00	1.00 1.00	100

Table 8: Representation of various performance parameters for Kidney Disease

Logistic Regression model considered here as it gives maximum accuracy i.e., 100%, so ROC curve is displayed as below.

The ROC Curve: -

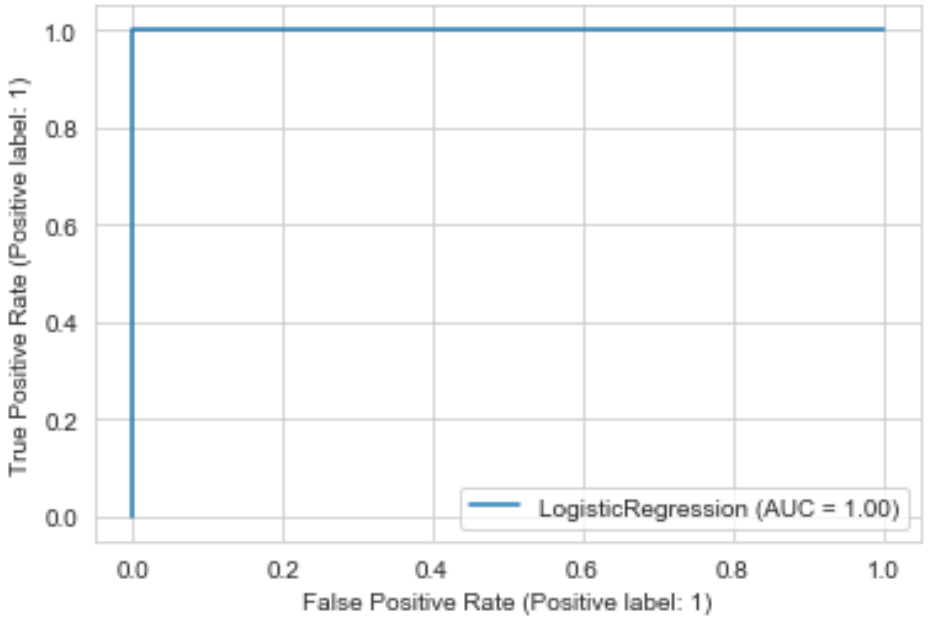


Figure 15: ROC curve representation of kidney diseases

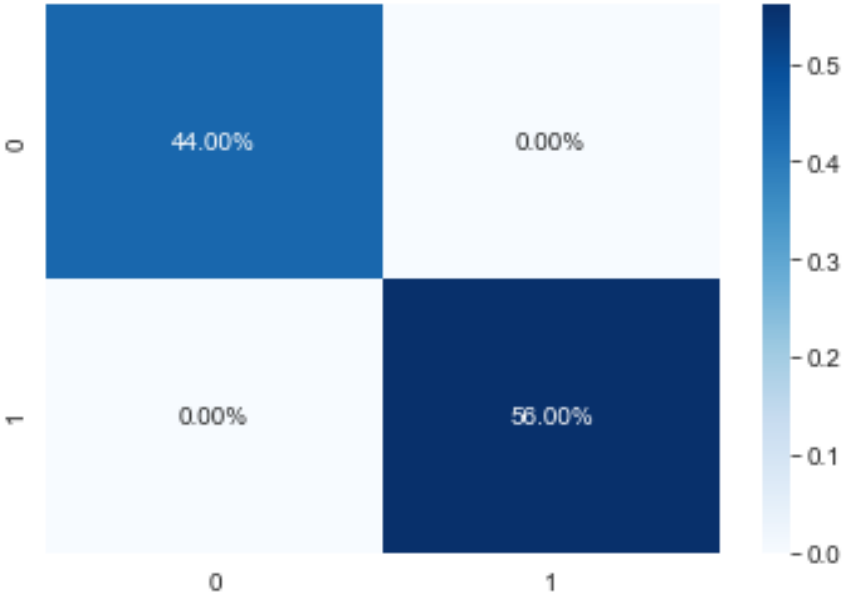


Figure 16: Confusion matrix for kidney diseases.

Algorithms	2/4	Precision	Recall	F1-score	ROC AUC Score (%)
Logistic Regression	2- 4-	0.97 0.94	0.97 0.94	0.97 0.94	95.50
KNN	2- 4-	0.94 0.98	0.99 0.89	0.96 0.93	93.78
SVM (Linear Classifier)	2- 4-	0.98 0.96	0.98 0.98	0.98 0.96	97.00
SVM (RBF Classifier)	2- 4-	0.97 0.96	0.98 0.94	0.97 0.95	96.05
Naive Bayes	2- 4-	0.93 0.89	0.93 0.89	0.93 0.89	91.00
Decision Tree	2- 4-	0.98 0.93	0.96 0.96	0.97 0.94	95.89
Random Forest	2- 4-	0.99 0.98	0.94 0.98	0.99 0.98	98.50

Table9: Representation of various performance parameters for breast cancer diseases

The ROC Curve: -

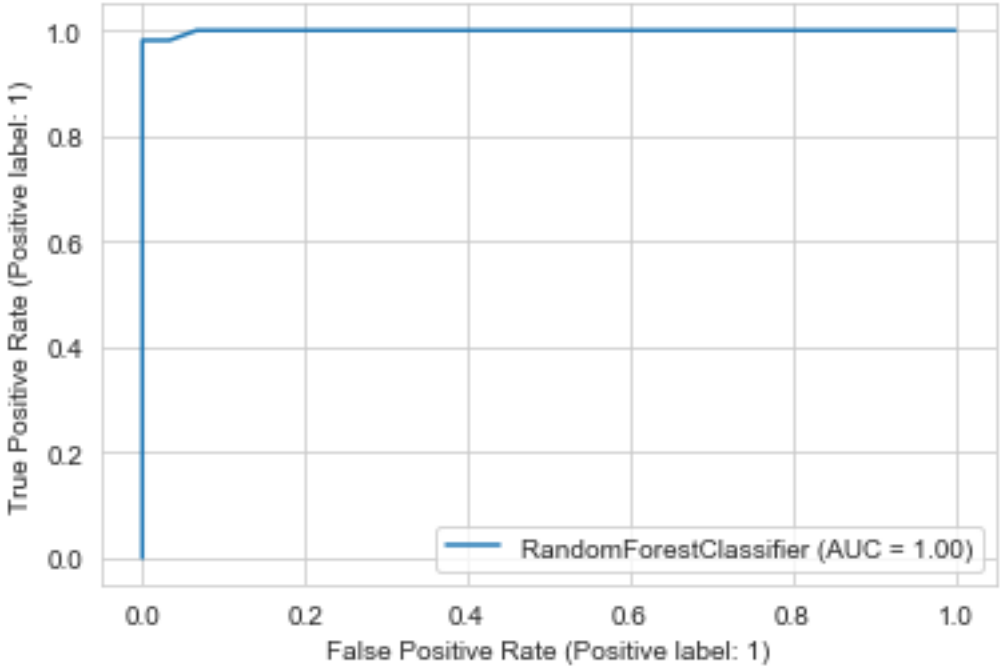


Figure 17: ROC curve representation of breast cancer

Here Random Forest model is selected as it gives 98.6 % accuracy.

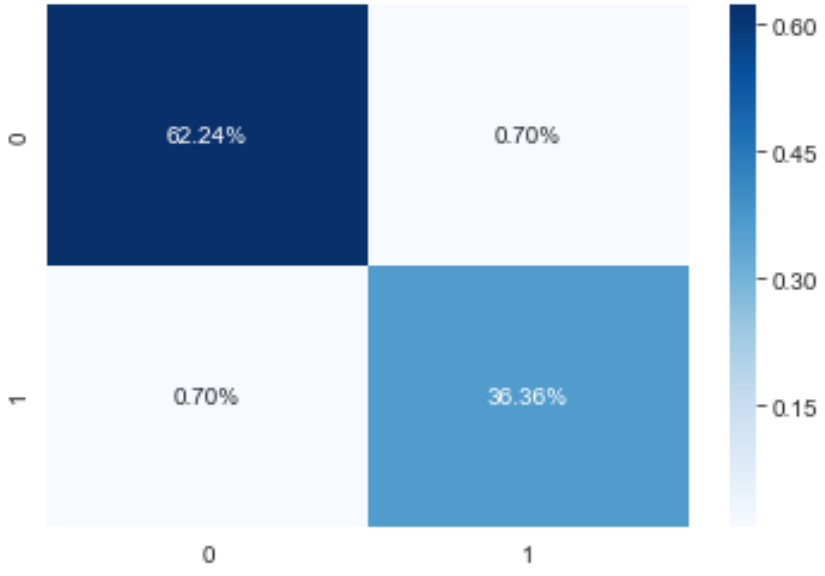


Figure 18: Confusion matrix for breast cancer

Algorithms	0/1	Precision	Recall	F1-score	ROC AUC Score (%)
Logistic Regression	0 - 1 -	0.80 0.74	0.91 0.55	0.85 0.63	72.92
KNN	0 - 1 -	0.79 0.66	0.86 0.53	0.83 0.59	69.94
SVM (Linear Classifier)	0 - 1 -	0.80 0.78	0.92 0.53	0.86 0.63	72.91
SVM (RBF Classifier)	0 - 1 -	0.79 0.67	0.87 0.52	0.83 0.58	69.50
Naive Bayes	0 - 1 -	0.80 0.66	0.86 0.57	0.83 0.61	71.24
Decision Tree	0 - 1 -	0.72 0.43	0.72 0.43	0.72 0.43	57.56
Random Forest	0 - 1 -	0.78 0.70	0.90 0.48	0.83 0.57	69.05

Table 10: Representation of various performance parameters for diabetes Prediction

ROC Curve: -

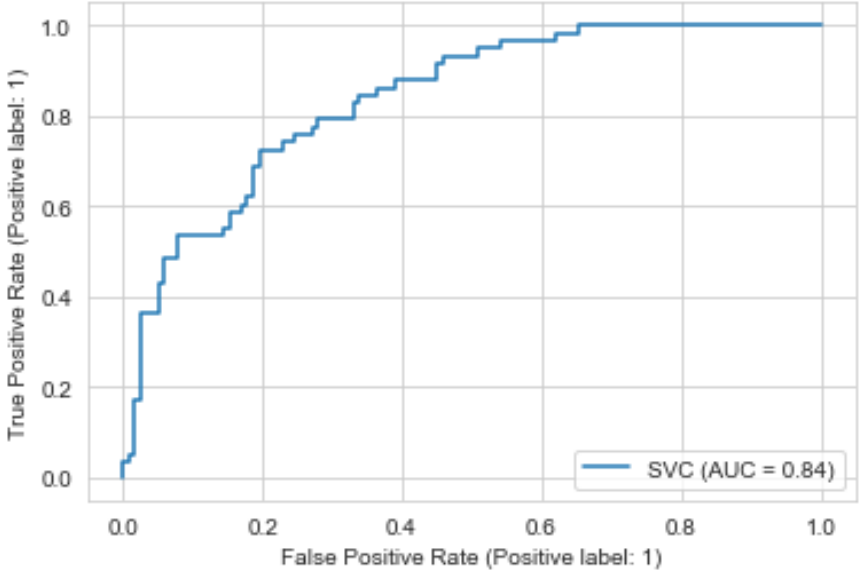


Figure 19: ROC curve representation for PIDD dataset

SVM(Linear Classifier) model is selected as it gives maximum accuracy of 80.72. The confusion matrix :-

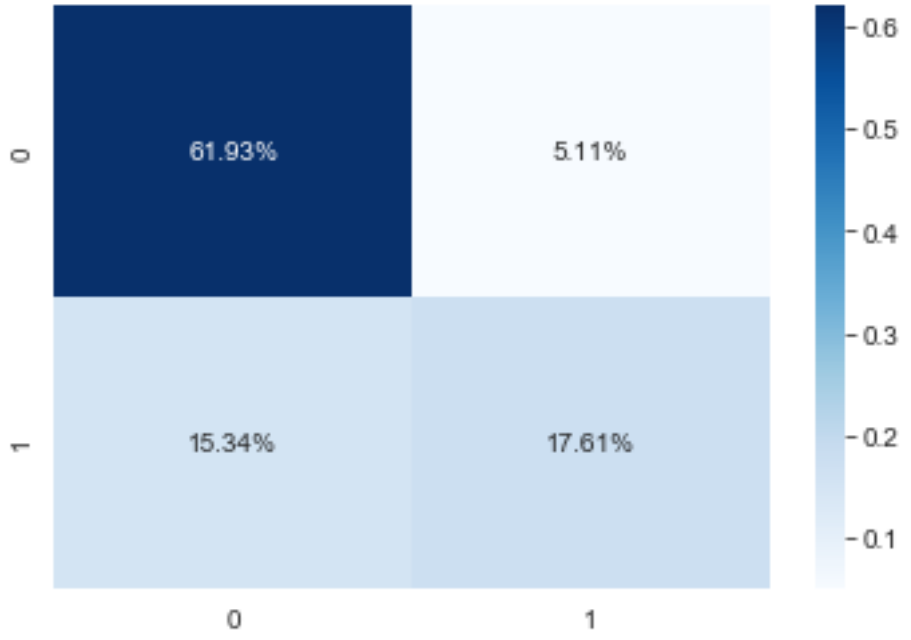


Figure 20:Confusion matrix for PIDD dataset.

Algorithms	0/1	Precision	Recall	F1-score	ROC AUC Score (%)
Logistic Regression	0 - 1 -	0.84 0.94	0.91 0.88	0.87 0.91	89.76
KNN	0 - 1 -	0.81 0.94	0.91 0.85	0.86 0.89	88.29
SVM (Linear Classifier)	0 - 1 -	0.78 0.93	0.91 0.82	0.84 0.87	86.82
SVM (RBF Classifier)	0 - 1 -	0.78 0.93	0.91 0.82	0.84 0.87	86.82
Naive Bayes	0 - 1 -	0.88 0.94	0.91 0.91	0.89 0.93	91.22
Decision Tree	0 - 1 -	0.85 0.97	0.96 0.88	0.90 0.92	91.94
Random Forest	0 - 1 -	0.83 0.91	0.87 0.88	0.85 0.90	87.59

Table 11: Representation of various performance parameters for Heart Disease

The ROC Curve: -

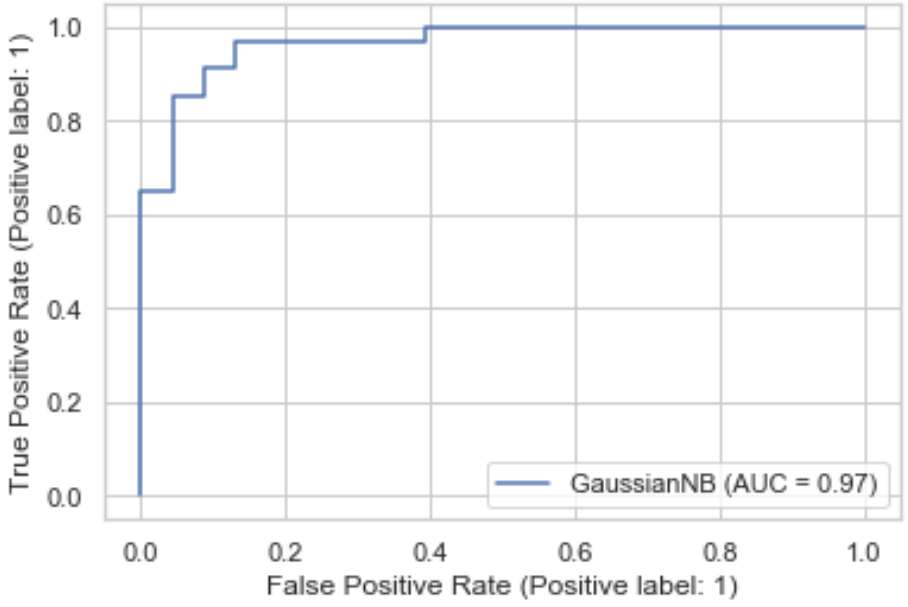


Figure 21:: AUC curve for heart dataset

Naïve Bayes model is selected as it gives maximum accuracy of 91.23 %.

The confusion matrix :-

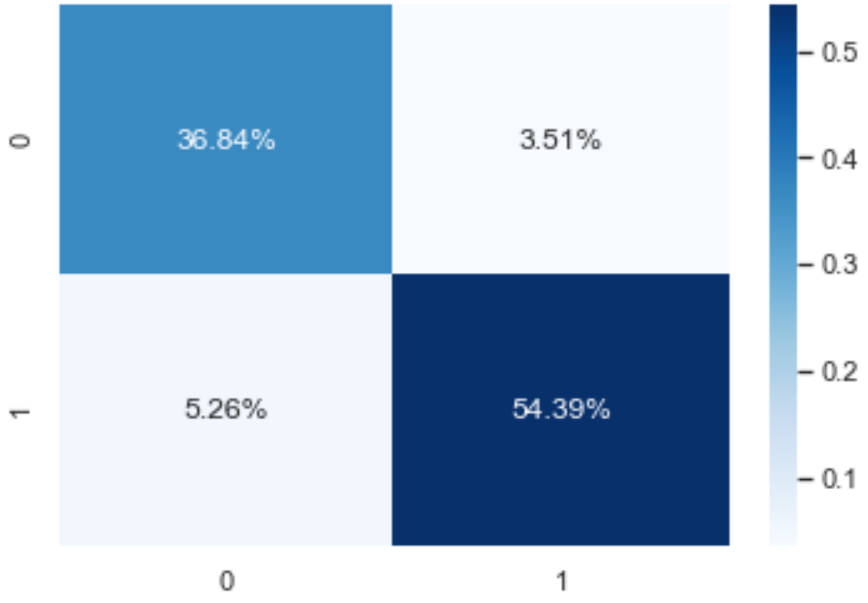


Figure 22:Confusion matrix for Heart dataset

Table 12: Representation of various performance parameters for Lung Cancer

Algorithms	0/1	Precision	Recall	F1-score	ROC AUC Score (%)
Logistic Regression	0 - 1 -	1.00 1.00	1.00 1.00	1.00 1.00	100.0
KNN	0 - 1 -	1.00 1.00	1.00 1.00	1.00 1.00	100.0
SVM (Linear Classifier)	0 - 1 -	1.00 1.00	1.00 1.00	1.00 1.00	100.0
SVM (RBF Classifier)	0 - 1 -	1.00 1.00	1.00 1.00	1.00 1.00	100.0
Naive Bayes	0 - 1 -	1.00 1.00	1.00 1.00	1.00 1.00	100.0
Decision Tree	0 - 1 -	1.00 1.00	1.00 1.00	1.00 1.00	100.0
Random Forest	0 - 1 -	1.00 1.00	1.00 1.00	1.00 1.00	100.0

Logistic Regression gives the accuracy of 100 %,so this model is selected.

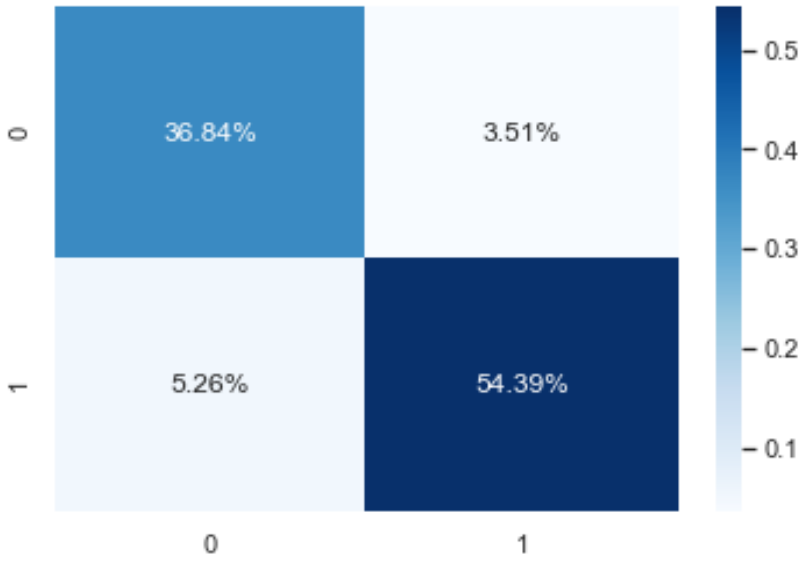


Figure 23:Confusion matrix for lung cancer dataset

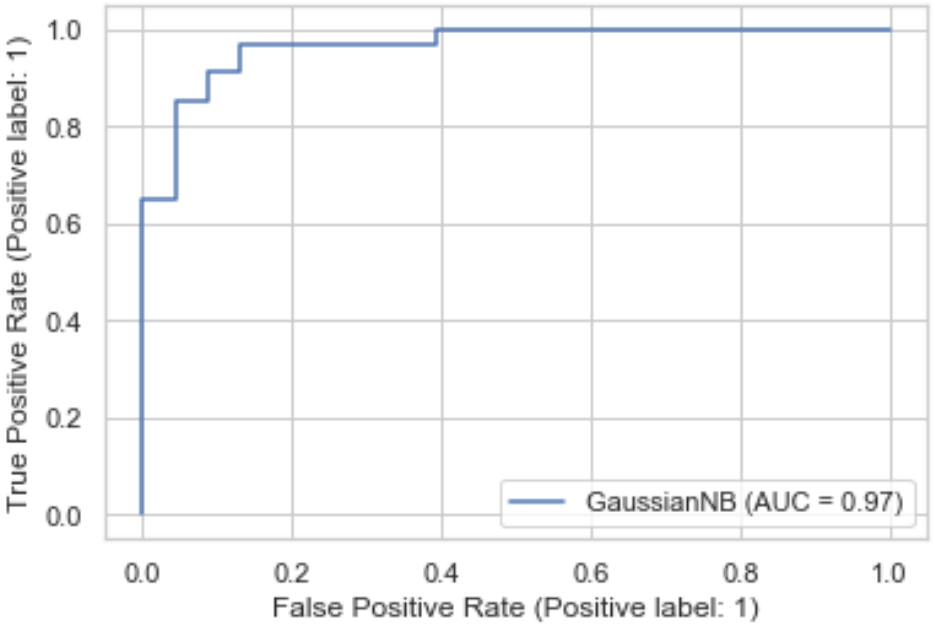


Figure 24: AUC curve for heart dataset

CHAPTER 5

LITERATURE STUDY

Lahoura et al. (2021) have modelled a frame work using extreme machine learning classifier with cloud-based computing techniques to improve classification accuracy and speed. Gain ration method is used on WBCD dataset to extract best features and use of extreme machine learning technique in cloud environment provides 98.8% classification accuracy. Dhahri et al. (2020) have performed analysis on WBCD (Wisconsin diagnostics breast cancer dataset) using genetic programming and ensemble machining learning classifier to classify benign and malignant.Ananalysis graphof machine learning technique and classifier log loss is depicted to plot the performance and accuracy.Ada-boostclassifier has shown aclassification accuracy of 98.24% with the above-mentioned approach.Christo et al. (2020)have proposed a novel approach for breast cancer and hepatitis dataset classification. They have introduced wrapper approach by combing three bio inspiredalgorithm and AdaBoost-SVM as classifier. Theiranalysis hasresulted a classification accuracy of 98.47% for WBCD dataset and 95.51% for hepatitis dataset.Osman et al. (2020) have proposed ensemble boosting classifier with RBFneural network to classify breast cancer original(WBC), breast cancer diagnosis (BCD), breast cancer prognostics (BCP) and Wisconsin dataset for diagnostics breast cancer (WBCD) datasets with atest accuracy of 97.45%,98.4%,97.7%,97.0%respectively. Muhammet faith Ak (2020) hasapplied various machine learning classifier to classify breast cancer dataset of UCI machine learning repository and finally gave a prediction that logistic regression inconsideration with all features resulted a classification accuracy of 98.1%. Ed-daoudy et al. (2020) haveused association rule to extract best features from WBCD dataset and applied SVM classifier to classify benign and malignant with a classification accuracy of 98% for eight attributes of dataset.Sakri et al. (2018) enhanced the classification performance by integrating feature selection algorithm that is particle swarm optimization (PSO) along with three machine learning classifier which are k nearest neighbour, Naive Bayes (NB) and reduced error pruning (REP) tree. Authors used particle swarm optimization technique to select features from Wisconsin breast cancer dataset of UCI repository to give a classification accuracy of 81.3%,80% and 75% using naive bayes, REP tree and k-nearest neighbour algorithm respectively.Houssainy et al. (2019) used probabilistic neural network (PNN) and radial basis function (RBF) to classify stages of UCI learning repository chronic kidney diseases with an overall accuracy of 96.7%. Amansouret al. (2019)have performed analysis on chronic kidney diseases of UCI learning repository with the optimized parameter for designing Artificial neural network that predicted a classification accuracy of 99.75%. Kauar et al. (2020) have introduced various supervised machine learning classifier such as linear kernel support vector machine, radial basic function, k nearest neighbour, kernel support vector machine, artificial neural network and multifactor dimensionality

reduction technique on pima Indian diabetes dataset (PIDD). Authors have suggested linear kernel SVM performs better and gives an accuracy of 89% and area under curve 0.90.Atik Mahabub (2019) implemented a novel ensemble voting classifier which is combination of three classifier.Author used support vector classifier, multiplayer perceptron and k-nearest neighbour to classify Pima Indian diabetes dataset with a classification accuracy of 86%.Sisodia et al. (2018) implemented naive bayes, support vector machine and decision tree on pima Indian diabetes dataset and concluded that naive bayes performs better and gives an accuracy of 76.3%. Escamillaet al. (2020) have proposed use of chi with principal component data reduction technique analysis to improve the machine learning classifier.So, chi with principal component analysis with random forest classifier is used to predict Cleveland heart dataset with classification accuracy of 98.7%, Hungarian dataset with a classification accuracy of 99% and 99.4% classification accuracy for mixture of Cleveland and Hungarian dataset.Tama et al. (2020) proposed two tier ensemble architecture. In first phase correlation-based feature selection and particle swarm optimization is applied to select best features and in second phase classification model comprises of random forest, gradient boosting technique and XGBoost are framed to predict an accuracy of 93.55 for Stat Logheart dataset, F1 score of 86.49% for Cleveland heart dataset, 91.9% classification accuracy for Hungarian heart dataset and 98.3% for Alizadeh sani dataset.Mohan et al. (2020) proposed hybrid random forest and linear model to classify Cleveland dataset with an accuracy of 88.4%. Shakeel et al. (2019) used wolf prey searching process for feature selection and discrete AdaBoost optimized ensemble learning generalized neural network, a novel hybrid approach of ANN to predict lung cancer data available from ELVIRA an accuracy of 99.6%. Pian Li et al. (2020) discussed feature selection algorithm such as RELIEF, minimal redundancy and maximum relevance(MRMR), least absolute shrinkage selection operator(LASSO), local learning based feature selection(LLBFS) algorithms and a newly proposed conditional mutual information feature selection (FCMIM) algorithm with machine learning classifier such linear regression, K- nearest neighbour, Artificial neural network, support vector machine, Naïve Bayes and decision tree classifiersto classify Cleveland heart dataset. The proposed conditional mutual information feature selection algorithm (FCMIM) technique with SVM resulted a classification accuracy of 92.37%.

CHAPTER 6

CONCLUSION

Based on the above review, it can be concluded that there is a huge scope for machine learning algorithms in predicting different diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases. SVM(linear classifier),when used for predicting diabetes gave best accuracy. For lung cancer all the algorithms gave best accuracy of 100pct but we have selected Logistic regression model. For breast cancer Random Forest gave best accuracy of 98.6pct. For kidney disease Logistic regression and Random Forest gave best accuracy of 100pct but we have selected Logistic regression model. And for heart disease naïve bayes and decision tree performed best with accuracy of 91.23pct but we have selected naïve bayes model. Systems based on machine learning algorithms and techniques have been very accurate in predicting the diseases.

REFERENCES

[1] Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated breast cancer diagnosis based on machine learning algorithms. Journal of healthcare engineering, 2019.

[2] Elgin Christo, V. R., Khanna Nehemiah, H., Minu, B., & Kannan, A. (2019). Correlation-Based Ensemble Feature Selection Using Bioinspired Algorithms and Classification Using Backpropagation Neural Network. Computational and mathematical methods in medicine, 2019, 7398307. https://doi.org/10.1155/2019/7398307

[3] Varadharajan, R., Priyan, M.K., Panchatcharam, P. et al. A new approach for prediction of lung carcinoma using back propogation neural network with decision tree classifiers. J Ambient Intell Human Comput (2018). https://doi.org/10.1007/s12652-018-1066-y

[4] Ak M. F. (2020). A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare (Basel, Switzerland), 8(2), 111. https://doi.org/10.3390/healthcare8020111

[5] Ed-daoudy, A., Maalmi, K. Breast cancer classification with reduced feature set using association rules and support vector machine. Netw Model Anal Health Inform Bioinforma 9, 34 (2020). https://doi.org/10.1007/s13721-020-00237-8

[6] Assiri, A. S., Nazir, S., & Velastin, S. A. (2020). Breast tumor classification using an ensemble machine learning method. Journal of Imaging, 6(6), 39.

[7] Rady, E. H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked, 15, 100178.

[8] Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., ... & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in biology and medicine, 109, 101-111.

[9] Jongbo, O. A., Adetunmbi, A. O., Ogunrinde, R. B., & Badeji-Ajisafe, B. (2020). Development of an ensemble approach to chronic kidney disease diagnosis. Scientific African, 8, e00456.

[10] Varadharajan, R., Priyan, M.K., Panchatcharam, P. et al. A new approach for prediction of lung carcinoma using back propogation neural network with decision tree classifiers. J Ambient Intell Human Comput (2018). https://doi.org/10.1007/s12652-018-1066-y

[11] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied computing and informatics.

[12] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

[13] Mahabub, A. (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Applied Sciences, 1(12), 1667.

[14] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/

[15] https://towardsdatascience.com/understanding-cancer-using-machine-learning-84087258ee18#:~:text=Researchers%20are%20now%20using%20ML,able%20to%20suppress%20its%20expression.

[16] https://www.geeksforgeeks.org/ml-types-learning-supervised-learning/

[17] https://www.edureka.co/blog/classification-in-machine-learning/#:~:text=In%20machine%20learning%2C%20classification%20is,recognition%2C%20document%20classification%2C%20etc.

[18] https://machinelearningmastery.com/logistic-regression-for-machine-learning/

[19] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[20] https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674

[21] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[22] Python Machine Learning by Sebastian Raschka and Vahid Mirjalili