

Chronic Kidney Disease Classification Using Machine Learning
at
Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award of Bachelor
of Engineering Degree in Computer Science and Engineering

By

Pavan Sai
REG. NO. 38110394



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU

MARCH 2022



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade “A” by NAAC
(Established under Section 3 of UGC Act, 1956)
JEPPIAAR NAGAR, RAJIV GANDHI SALAI
CHENNAI– 600119
www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of Musunuru Pavan sai(38110394) who carried out the project entitled “ Chronic Kidney Disease Classification using Machine Learning under my supervision from June 2021 to November 2021.

Internal Guide

Ms.Jeyanthi .s

Head of the Department

Dr.S.Vigneshwari ,ME, Ph.D, and Dr.L.Lakshmanan ME,PH.D

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, Musunuru Pavan Sai hereby declare that the project report entitled “**Chronic Kidney Diseases Classification Using Machine Learning**” done by me under the guidance of **MS.Jeyanthi.S** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

DATE:10-04-2022

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGE,. MENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D, Dean**, School of Computing, **Dr. S. Vigneshwari, M.E., Ph.D. and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms.Jeyanthi .S** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Chronic kidney disease is a general term for heterogeneous disorders affecting kidney structure and function. It is recognized now that even mild abnormalities in measures of kidney structure and function are associated with increased risk for developing complications in other organ systems which lead to mortality, all of which occur more frequently than kidney failure. Machine Learning has been a current trend for attaining diagnostic results. Huge amount of un mined data is collected by the healthcare industry in order to discover hidden information for effective diagnosis and decision making. Machine Learning involves extracting hidden information from massive dataset, categorizing valid and unique patterns in data. In this research we are going to use machine learning algorithms SVM, Decision Tree and Random Forest to improve the classification of chronic kidney disease.

TABLE OF CONTENTS

Chapter no	Title	Page no
	ABSTRACT	i
	LIST OF FIGURES	ii
	LIST OF TABLES	
	LIST OF ABBREVIATIONS	iii
1	INTRODUCTION	1
	1.1 INTRODUCTION	1
2	LITERATURE SURVEY	3
	2.1 LITERATURE SURVEY	3
3	METHODOLOGY	5
	3.1 EXISTING SYSTEM	5
	3.2 PROPOSED SYSTEM	5
	3.3 FEASIBILITY STUDY	6
	3.4 SYSTEM REQUIREMENT	8
	3.5 SYSTEM ARCHITECTURE	29
	3.6 USE CASE DIAGRAM	32

	3.7 MODULE	37
4	CONCLUSION	45
	REFERENCES	46
	APPENDIX	47
	A. SOURCE CODE	48
	B.SCREENSHOTS	50

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO
1	SYSTEM ARCHITETURE	26
2	FLOW DIAGRAM	27
3	USE CASE DIAGRAM	30
4	SEQUENCE DIAGRAM	31
5	ACITIVITY DIAGRAM	32

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
ML	MACHINE LEARNING
FR	FOREST RONDON
DT	DECISION TREE
SVM	SUPER VECTOR MACHINE

CHAPTER 1

INTRODUCTION

1.1 Introduction

Chronic Kidney Disease (CKD) is considered as an important threat for the society with respect to the health in the present era. Chronic kidney disease can be detected with regular laboratory tests, and some treatments are present which can prevent development, slow disease progression, reduce complications of decreased Glomerular Filtration Rate(GFR) and

risk of cardiovascular disease, and improve survival and quality of life. CKD can be caused due to lack of water consumption, smoking, improper diet, loss of sleep and many other factors. This disease affected 753 million people globally in 2016 in which 417 million are females and 336 million are males. Majority of the time the disease is detected in its final stage and which sometimes leads to kidney failure. The existing system of diagnosis is based on the examination of urine with the help of serum creatinine level. Many medical methods are used for this purpose such as screening, ultrasound method. In screening, the patients with hypertension, history of cardiovascular disease, disease in the past, and the patients who have relatives who had kidney disease are screened. This technique includes the calculation of the estimated GFR from the serum creatinine level, and measurement of urine albumin-to-creatinine ratio (ACR) in a first morning urine specimen. This paper focuses on machine

learning techniques like Decision tree, Random Forest and SVM algorithms by minimizing the features and selecting best features to improve the accuracy of prediction.

1.2 Objective

First we will train a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using better algorithm depending upon the accuracy. Visualization of dataset is done to analyze the Chronic Kidney Disease (CKD) . This work helps to identify Chronic Kidney Disease using machine learning techniques like Decision tree, Random Forest and SVM algorithms.

CHAPTER 2

LITERATURE SURVEY

2.1 Literature Survey

1. Chronic Kidney Disease Prediction using Data Mining, J. Snegha, International Conference on Emerging Trends, 2020.

The proposed a system that uses various data mining techniques like Random Forest algorithm and Back propagation neural Network. Here they compare both of the algorithm and found that Back Propagation algorithm gives the best result as it uses the supervised learning network called feedforward neural network.

2. Multiple Pheromone table based on Ant Colony Optimization for Clustering, Hindawi, Research article, Kai-Cheng Hu, 2015.

They proposed a system which uses a multiple pheromone table based on ACO for clustering. Here they divided the problem into a set of several different patterns based on their features. Two pheromone tables are used here one for keeping the track of the promising information and the other to hold the details of unpromising information which in turn increases the probability of searching directions.

3. Chronic Kidney Disease Prediction Using Back Propagation Neural Network Algorithm, Advances in Intelligent Systems and Computing, Nilesch Borisagar, 2017

This system which uses Back Propagation Neural Network algorithm for prediction. Here Levenberg, Bayesian regularization, Scaled Conjugate and resilient back propagation algorithm are discussed. Matlab R2013a is used for

the implementation purpose. Based on the training time, scaled conjugate gradient and resilient back propagation are found more efficient than Levenberg and Bayesian regularization.

4. Chronic Kidney Disease Prediction using Machine Learning, Reshma S , Salma Shaji , S R Ajina , Vishnu Priya S R, Janisha A, 2020, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY

They proposed a system that uses extreme learning machine and ACO for CKD prediction. Classification is done using MATLAB tool and ELM has few constraints in the optimization. This technique is an improvement under the Sigmoid additive type of SLFNs.

5. Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Arif-Ul-Islam, Ant-Miner and J48 Decision Tree, 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE).

They proposed a system in which prediction of disease is done using Boosting Classifiers, Ant-Miner and J48 Decision Tree. The aim of this paper is two fold that is, analyzing the performance of boosting algorithms for detecting CKD and deriving rules illustrating relationships among the attributes of CKD. Experimental results prove that the performance of AdaBoost was less than of LogitBoost by a fraction.

CHAPTER 3

METHODOLOGY

3.1 Existing System

Chronic Kidney Disease (CKD) is considered as an important threat for the society with respect to the health in the present era. Chronic kidney disease can be detected with regular laboratory tests, and some treatments are present which can prevent the disease. In previous system they have implemented ACO for CKD prediction. Classification is done using MATLAB tool and ELM has few constraints in the optimization.

Disadvantages

- Sequences of random decisions (not independent)
- Probability distribution changes by iteration

3.2 Proposed System

Our proposed system uses machine learning techniques to get high prediction which uses Decision tree, Random Forest and SVM techniques. While using machine learning techniques prediction process is less time consuming so that

doctors can analyze the patients within a less time period.

3.3 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are

Economic feasibility

Technical feasibility

Social feasibility

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical

6

requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands being placed on the client. A feasibility study evaluates the project's potential for success.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The

The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some

constructive criticism, which is welcomed, as he is the final user of the system.

3.4 SYSTEM REQUIREMENTS

3.4.1 HARDWARE REQUIREMENTS

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware. The minimal hardware requirements are as follows,

1. Processor : Pentium IV
2. RAM : 8 GB
3. Processor : 2.4 GHz
- 7
4. Main Memory : 8GB RAM
5. Hard Disk Drive : 1tb
6. Keyboard : 104 Keys

3.4.2 SOFTWARE REQUIREMENTS

Software requirements deals with defining resource requirements and prerequisites that needs to be installed on a computer to provide functioning of an application. The minimal software requirements are as follows,

1. Front end : python
2. Dataset : csv
3. IDE : anaconda
4. Operating System : Windows 10

Python Language

Python is an object-oriented programming language created by Guido Rossum in 1989. It is ideally designed for rapid prototyping of complex applications. It has interfaces to many OS system calls and libraries and is extensible to C or C++. Many large companies use the Python programming language include NASA, Google, YouTube, BitTorrent, etc. Python programming is widely used in Artificial Intelligence, Natural Language Generation, Neural Networks and other advanced fields of Computer Science. Python had deep focus on code readability & this class will teach you python from basics.

Python Programming Characteristics

It provides rich data types and easier to read syntax than any other programming languages

8

It is a platform independent scripted language with full access to operating system API's

Compared to other programming languages, it allows more run-time flexibility

It includes the basic text manipulation facilities of Perl and Awk

A module in Python may have one or more classes and free functions

Libraries in Python are cross-platform compatible with Linux, Macintosh, and Windows

For building large applications, Python can be compiled to byte-code

Python supports functional and structured programming as well as OOP

It supports interactive mode that allows interacting Testing and debugging of snippets of code

In Python, since there is no compilation step, editing, debugging and testing is fast.

Applications of Python Programming

Web Applications

You can create scalable Web Apps using frameworks and CMS (Content Management System) that are built on Python. Some of the popular platforms for creating Web Apps are: Django, Flask, Pyramid, Plone, Django CMS. Sites like Mozilla, Reddit, Instagram and PBS are written in Python.

Scientific and Numeric Computing

There are numerous libraries available in Python for scientific and numeric computing. There are libraries like: SciPy and NumPy that are used in general purpose computing. And, there are specific libraries like: EarthPy for earth science, AstroPy for Astronomy and so on. Also, the language is heavily used in machine learning, data mining and deep learning.

Creating software Prototypes

Python is slow compared to compiled languages like C++ and Java. It might not

9

be a good choice if resources are limited and efficiency is a must. However, Python is a great language for creating prototypes. For example: You can use

Pygame (library for creating games) to create your game's prototype first. If you like the prototype, you can use language like C++ to create the actual game.

Good Language to Teach Programming

Python is used by many companies to teach programming to kids and newbies. It is a good language with a lot of features and capabilities. Yet, it's one of the easiest language to learn because of its simple easy-to-use syntax.

About Opencv Package

Python is a general purpose programming language started by Guido van Rossum, which became very popular in short time mainly because of its simplicity and code readability. It enables the programmer to express his ideas in fewer lines of code without reducing any readability.

Compared to other languages like C/C++, Python is slower. But another important feature of Python is that it can be easily extended with C/C++. This feature helps us to write computationally intensive codes in C/C++ and create a Python wrapper for it so that we can use these wrappers as Python modules. This gives us two advantages: first, our code is as fast as original C/C++ code (since it is the actual C++ code working in background) and second, it is very easy to code in Python. This is how OpenCV-Python works, it is a Python wrapper around original C++ implementation.

And the support of Numpy makes the task more easier. Numpy is a highly optimized library for numerical operations. It gives a MATLAB-style syntax. All the OpenCV array structures are converted to-and-from Numpy arrays. So whatever operations you can do in Numpy, you can combine it with OpenCV, which increases number of weapons in your arsenal. Besides that, several other libraries like SciPy, Matplotlib which supports Numpy can be used with this.

So OpenCV-Python is an appropriate tool for fast prototyping of computer

vision problems.

1.1 DOMAIN DESCRIPTION

DATA MINING

Data Mining?

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data. It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology. The insights derived via Data Mining can be used for marketing, fraud detection, and scientific discovery, etc. Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, etc. Data mining is a big area of data sciences, which aims to discover patterns and features in data, often large data sets. It includes regression, classification, clustering, detection of anomaly, and others. It also includes preprocessing, validation, summarization, and ultimately the making sense of the data sets. The evolution of the Internet and the social media has resulted in the huge explosion of the data volumes and complexity, the so-called big data nowadays. Consequently, data mining has also expanded beyond the traditional data modeling such as regression and statistical models. For example, the aim of clustering is dividing n observations into some different clusters, based on certain clustering measures or objectives. Classification is dividing the data set into different classes with different labels such as normal or abnormal, yes or no. There are a vast array of data mining methods. In this book, we will introduce the most widely used techniques.

Types of Data

Data mining can be performed on following types of data

Relational databases

Data warehouses

11

Advanced DB and information repositories

Object-oriented and object-relational databases

Transactional and Spatial databases

Heterogeneous and legacy databases

Multimedia and streaming database

Text databases

Text mining and Web mining

Data Mining Techniques

1.Classification:

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

2. Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. Regression:

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. Outer detection:

This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud

12

or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

6. Sequential Patterns:

This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

7. Prediction:

Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

Benefits of Data Mining:

Data mining technique helps companies to get knowledge-based information.

Data mining helps organizations to make the profitable adjustments in operation and production.

The data mining is a cost-effective and efficient solution compared to other statistical data applications.

Data mining helps with the decision-making process.

Facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.

It can be implemented in new systems as well as existing platforms

It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

Types of Data Mining

Data mining has two primary processes: supervised and unsupervised learning.

Supervised Learning

The goal of supervised learning is prediction or classification. The easiest way to conceptualize this process is to look for a single output variable. A process is considered supervised learning if the goal of the model is to predict the value of

13

an observation. One example is spam filters, which use supervised learning to classify incoming emails as unwanted content and automatically remove these messages from your inbox.

Common analytical models used in supervised data mining approaches are:

Linear Regressions

Linear regressions predict the value of a continuous variable using one or more independent inputs. Realtors use linear regressions to predict the value of a house based on square footage, bed-to-bath ratio, year built, and zip code.

Logistic Regressions

Logistic regressions predict the probability of a categorical variable using one or more independent inputs. Banks use logistic regressions to predict the probability that a loan applicant will default based on credit score, household income, age, and other personal factors.

Time Series

Time series models are forecasting tools which use time as the primary independent variable. Retailers, such as Macy's, deploy time series models to predict the demand for products as a function of time and use the forecast to accurately plan and stock stores with the required level of inventory.

Classification or Regression Trees

Classification Trees are a predictive modeling technique that can be used to predict the value of both categorical and continuous target variables. Based on the data, the model will create sets of binary rules to split and group the highest proportion of similar target variables together. Following those rules, the group

14

that a new observation falls into will become its predicted value.

Neural Networks

A neural network is an analytical model inspired by the structure of the brain, its neurons, and their connections. These models were originally created in

1940s but have just recently gained popularity with statisticians and data scientists. Neural networks use inputs and, based on their magnitude, will “fire” or “not fire” its node based on its threshold requirement. This signal, or lack thereof, is then combined with the other “fired” signals in the hidden layers of the network, where the process repeats itself until an output is created. Since one of the benefits of neural networks is a near-instant output, self-driving cars are deploying these models to accurately and efficiently process data to autonomously make critical decisions.

K-Nearest Neighbor

The K-nearest neighbor method is used to categorize a new observation based on past observations. Unlike the previous methods, k-nearest neighbor is data-driven, not model-driven. This method makes no underlying assumptions about the data nor does it employ complex processes to interpret its inputs. The basic idea of the k-nearest neighbor model is that it classifies new observations by identifying its closest K neighbors and assigning it the majority’s value. Many recommender systems nest this method to identify and classify similar content which will later be pulled by the greater algorithm.

Summary:

1. Data Mining is all about explaining the past and predicting the future for analysis.

15

2. Data mining helps to extract information from huge sets of data. It is the procedure of mining knowledge from data.

3. Data mining process includes business understanding, Data Understanding, Data Preparation, Modelling, Evolution, Deployment.
4. Important Data mining techniques are Classification, clustering, Regression, Association rules, Outlier detection, Sequential Patterns, and prediction
5. R-language and Oracle Data mining are prominent data mining tools.
6. Data mining technique helps companies to get knowledge-based information.
7. The main drawback of data mining is that many analytics software is difficult to operate and requires advance training to work on.
8. Data mining is used in diverse industries such as Communications, Insurance, Education, Manufacturing, Banking, Retail, Service providers, eCommerce, Supermarkets Bioinformatics.

Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine Learning Methods

Machine learning algorithms are often categorized as supervised or unsupervised.

Supervised machine learning algorithms can apply what has been learned in

the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is

required for the agent to learn which action is best; this is known as the reinforcement signal.

17

Advantages of Machine Learning

Machine Learning undoubtedly helps people to work more creatively and efficiently. Basically, you too can delegate quite complex or monotonous work to the computer through Machine Learning - starting with scanning, saving and filing paper documents such as invoices up to organizing and editing images.

In addition to these rather simple tasks, self-learning machines can also perform complex tasks. These include, for example, the recognition of error patterns. This is a major advantage, especially in areas such as the manufacturing industry: the industry relies on continuous and error-free production. While even experts often cannot be sure where and by which correlation a production error in a plant fleet arises, Machine Learning offers the possibility to identify the error early this saves down times and money. Self-learning programs are now also used in the medical field. In the future, after "consuming" huge amounts of data (medical publications, studies, etc.), apps will be able to warn a in case his doctor wants to prescribe a drug that he cannot tolerate. This "knowledge" also means that the app can propose alternative options which for example also take into account the genetic requirements of the respective patient.

Applications of Machine Learning

1. Virtual Personal Assistants

Siri, Alexa, Google Now are some of the popular examples of virtual personal assistants. As the name suggests, they assist in finding information, when asked over voice. All you need to do is activate them and ask “What is my schedule for today?”, “What are the flights from Germany to London”, or similar questions. For answering, your personal assistant looks out for the information,

recalls your related queries, or send a command to other resources (like phone apps) to collect info. You can even instruct assistants for certain tasks like “Set an alarm for 6 AM next morning”, “Remind me to visit Visa Office day after tomorrow”. Machine learning is an important part of these personal assistants as

18

they collect and refine the information on the basis of your previous involvement with them. Later, this set of data is utilized to render results that are tailored to your preferences. Virtual Assistants are integrated to a variety of platforms. For example: Smart Speakers: Amazon Echo and Google Home Smartphones: Samsung Bixby on Samsung S8 Mobile Apps: Google Allow

2. Predictions while Commuting

Traffic Predictions: We all have been using GPS navigation services. While we do that, our current locations and velocities are being saved at a central server for managing traffic. This data is then used to build a map of current traffic. While this helps in preventing the traffic and does congestion analysis, the underlying problem is that there are less number of cars that are equipped with GPS. Machine learning in such scenarios helps to estimate the regions where congestion can be found on the basis of daily experiences.

Online Transportation Networks: When booking a cab, the app estimates the price of the ride. When sharing these services, how do they minimize the detours? The answer is machine learning. Jeff Schneider, the engineering lead at Uber ATC reveals in a an interview that they use ML to define price surge hours by predicting the rider demand. In the entire cycle of the services, ML is playing a major role.

3. Videos Surveillance

Imagine a single person monitoring multiple video cameras! Certainly, a difficult job to do and boring as well. This is why the idea of training computers to do this job makes sense. The video surveillance system nowadays are powered by AI that makes it possible to detect crime before they happen. They track unusual behaviour of people like standing motionless for a long time, stumbling, or napping on benches etc. The system can thus give an alert to

19

human attendants, which can ultimately help to avoid mishaps. And when such activities are reported and counted to be true, they help to improve the surveillance services. This happens with machine learning doing its job at the backend.

4. Social Media Services

From personalizing your news feed to better ads targeting, social media platforms are utilizing machine learning for their own and user benefits. Here are a few examples that you must be noticing, using, and loving in your social media accounts, without realizing that these wonderful features are nothing but the applications of ML. People You May Know: Machine learning works on a simple concept: understanding with experiences. Facebook continuously notices the friends that you connect with, the profiles that you visit very often, your interests, workplace, or a group that you share with someone etc. On the basis of continuous learning, a list of Facebook users are suggested that you can become friends with. Face Recognition: You upload a picture of you with a friend and Facebook instantly recognizes that friend. Facebook checks the poses and projections in the picture, notice the unique features, and then match them with the people in your friend list. The entire process at the backend is complicated and takes care of the precision factor but seems to be a simple application of ML at the front end. Similar Pins: Machine learning is the core element of Computer

Vision, which is a technique to extract useful information from images and videos. Pinterest uses computer vision to identify the objects (or pins) in the images and recommend similar pins accordingly.

5. Email Spam and Malware Filtering

There are a number of spam filtering approaches that email clients use. To ascertain that these spam filters are continuously updated, they are powered by machine learning. When rule-based spam filtering is done, it fails to track the latest tricks adopted by spammers. Multi Layer Perceptron, C 4.5 Decision Tree

20

Induction are some of the spam filtering techniques that are powered by ML. Over 325, 000 malwares are detected everyday and each piece of code is 90–98% similar to its previous versions. The system security programs that are powered by machine learning understand the coding pattern. Therefore, they detects new malware with 2–10% variation easily and offer protection against them.

6. Online Customer Support

A number of websites nowadays offer the option to chat with customer support representative while they are navigating within the site. However, not every website has a live executive to answer your queries. In most of the cases, you talk to a chatbot. These bots tend to extract information from the website and present it to the customers. Meanwhile, the chatbots advances with time. They tend to understand the user queries better and serve them with better answers, which is possible due to its machine learning algorithms.

7. Search Engine Result Refining

Google and other search engines use machine learning to improve the search results for you. Every time you execute a search, the algorithms at the backend keep a watch at how you respond to the results. If you open the top results and

stay on the web page for long, the search engine assumes that the the results it displayed were in accordance to the query. Similarly, if you reach the second or third page of the search results but do not open any of the results, the search engine estimates that the results served did not match requirement. This way, the algorithms working at the backend improve the search results.

8. Product Recommendations

You shopped for a product online few days back and then you keep receiving emails for shopping suggestions. If not this, then you might have noticed that the shopping website or the app recommends you some items that somehow matches with your taste. Certainly, this refines the shopping experience but did

21

you know that it's machine learning doing the magic for you? On the basis of your behaviour with the website/app, past purchases, items liked or added to cart, brand preferences etc., the product recommendations are made.

9. Online Fraud Detection

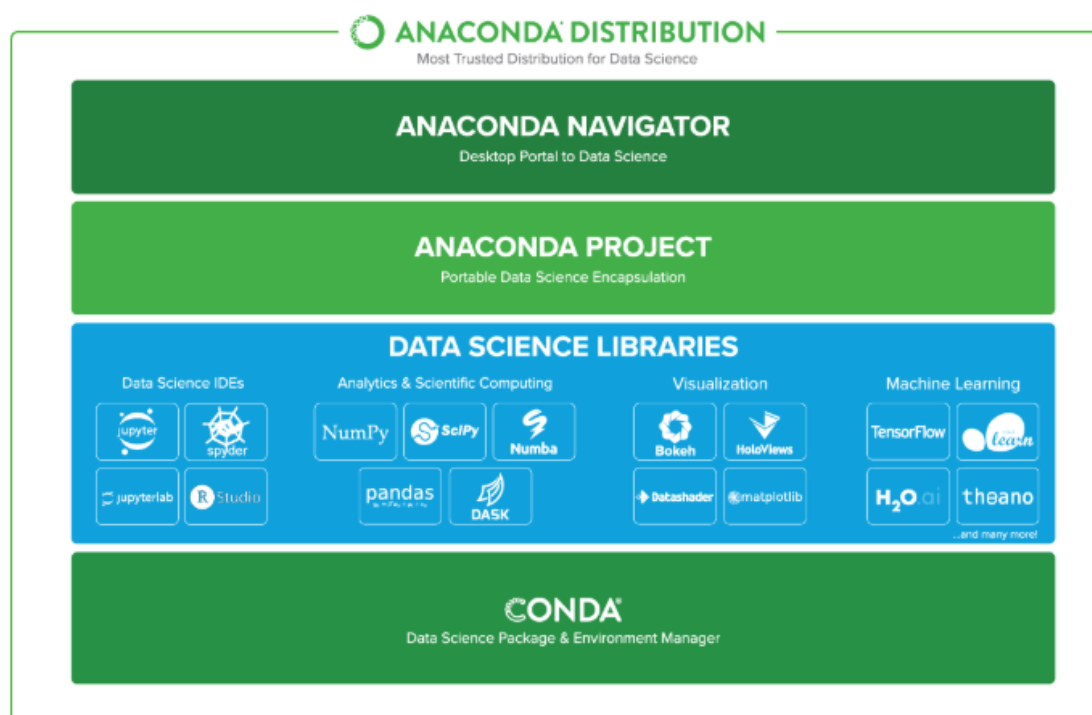
Machine learning is proving its potential to make cyberspace a secure place and tracking monetary frauds online is one of its examples. For example: Paypal is using ML for protection against money laundering. The company uses a set of tools that helps them to compare millions of transactions taking place and distinguish between legitimate or illegitimate transactions taking place between the buyers and sellers.

FEATURES OF ANACONDA NAVIGATOR

Anaconda is a free and open source, easy to install distribution of Python and R programming languages. Anaconda provides a working environment which is used for scientific computing, data science, statistical analysis and machine learning.

The latest distribution of Anaconda is Anaconda 5.3 and is released in October, 2018. It has the conda package, environment manager and a collection of 1000+ open source packages long with free community support.

22

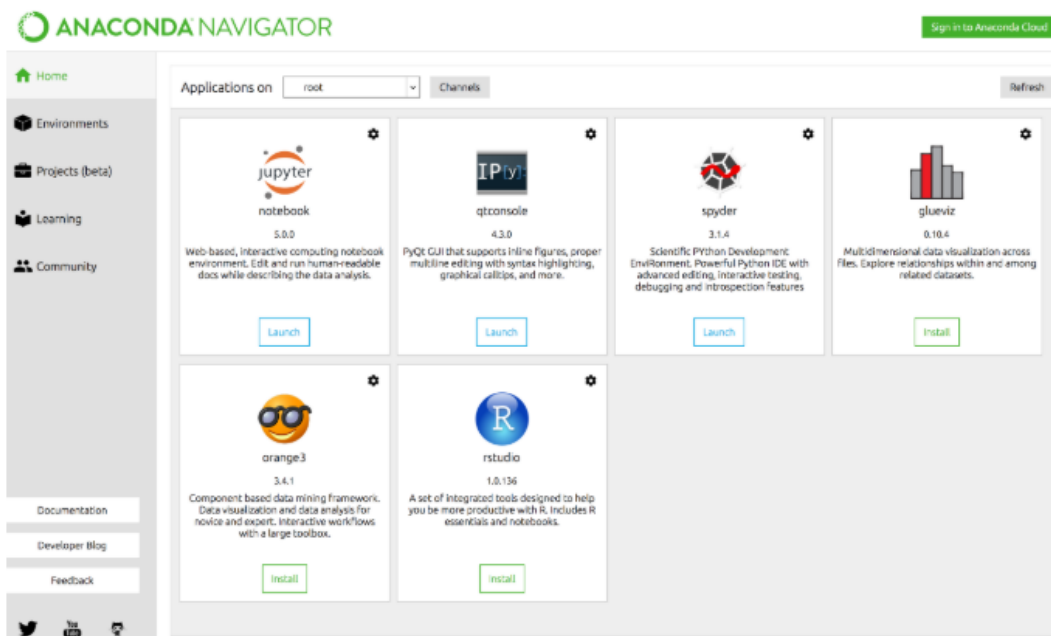


What is Anaconda Navigator?

25

Anaconda Navigator is a desktop graphical user interface (GUI) included in the Anaconda distribution. It allows us to launch applications provided in the Anaconda distribution and easily manage conda packages, environments and channels without the use of command-line commands. It is available for Windows, macOS and Linux.

23



Anaconda Navigator

Applications Provided In Anaconda Distribution

26

The Anaconda distribution comes with the following applications along with Anaconda Navigator.

1. JupyterLab
2. Jupyter Notebook
3. Qt Console
4. Spyder
5. Glueviz
6. Orange

24

7. RStudio
8. Visual Studio Code

> JupyterLab: This is an extensible working environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

> Jupyter Notebook: This is a web-based, interactive computing notebook environment. We can edit and run human-readable docs while describing the data analysis.

> Qt Console: It is the PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips and more.

> Spyder: Spyder is a scientific Python Development Environment. It is a powerful Python IDE with advanced editing, interactive testing, debugging and introspection features.

> VS Code: It is a streamlined code editor with support for development operations like debugging, task running and version control.

> Glueviz: This is used for multidimensional data visualization across files. It explores relationships within and among related datasets.

> Orange 3: It is a component-based data mining framework. This can be used for data visualization and data analysis. The workflows in Orange 3 are very interactive and provide a large toolbox.

25

> Rstudio: It is a set of integrated tools designed to help you be more productive with R. It includes R essentials and notebooks.

New Features of Anaconda 5.3



Compiled with Latest Python release: Anaconda 5.3 is compiled with Python 3.7, taking advantage of Python's speed and feature improvements.

- **Better Reliability:** The reliability of Anaconda has been improved in the latest release by capturing and storing the package metadata for installed packages.

- **Enhanced CPU Performance:** The Intel Math Kernel Library 2019 for Deep Neural Networks(MKL 2019) has been introduced in Anaconda 5.3 distribution. Users deploying Tensorflow can make use of MKL 2019 for Deep Neural Networks. These Python binary packages are provided to achieve high CPU performance.

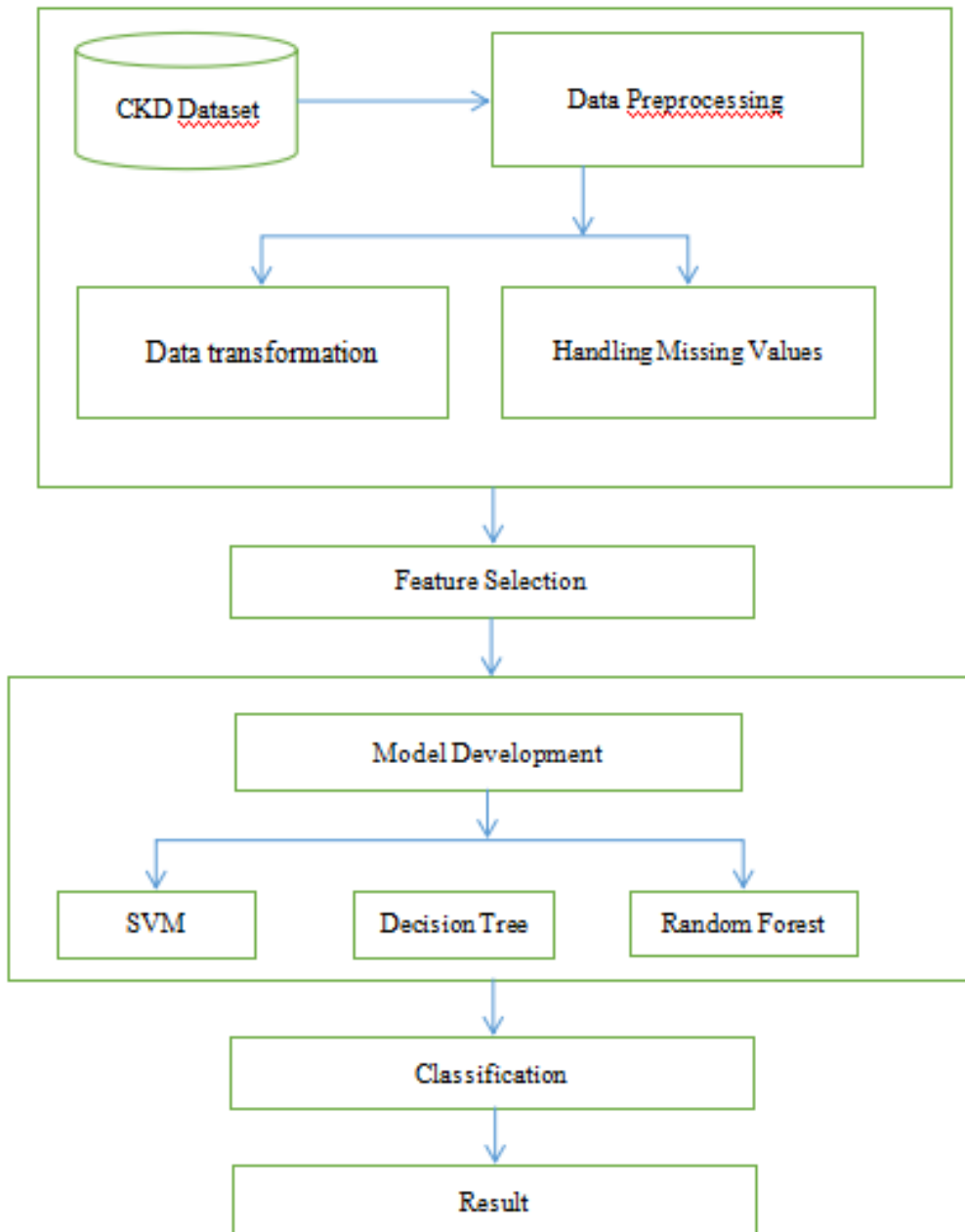
26

- **New packages are added:** There are over 230 packages which has been updated and added in the new release.

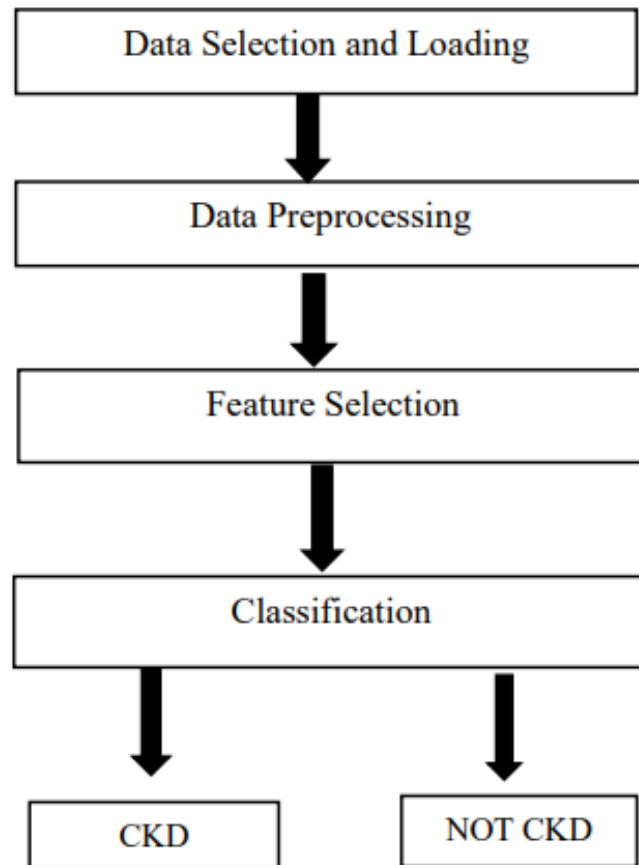
29

- **Work in Progress:** There is a casting bug in Numpy with Python 3.7 but the team is currently working on patching it until Numpy is updated

3.5 SYSTEM ARCHITECTURE



FLOW DIAGRAM



UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

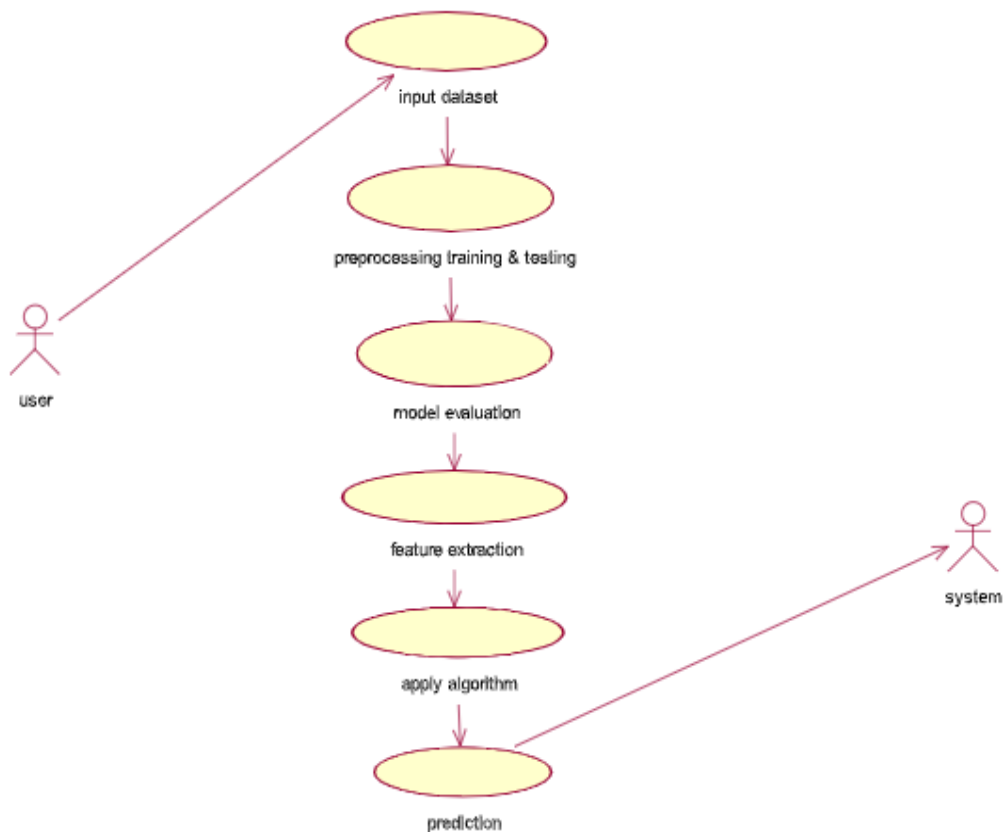
3.6 USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies

between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the

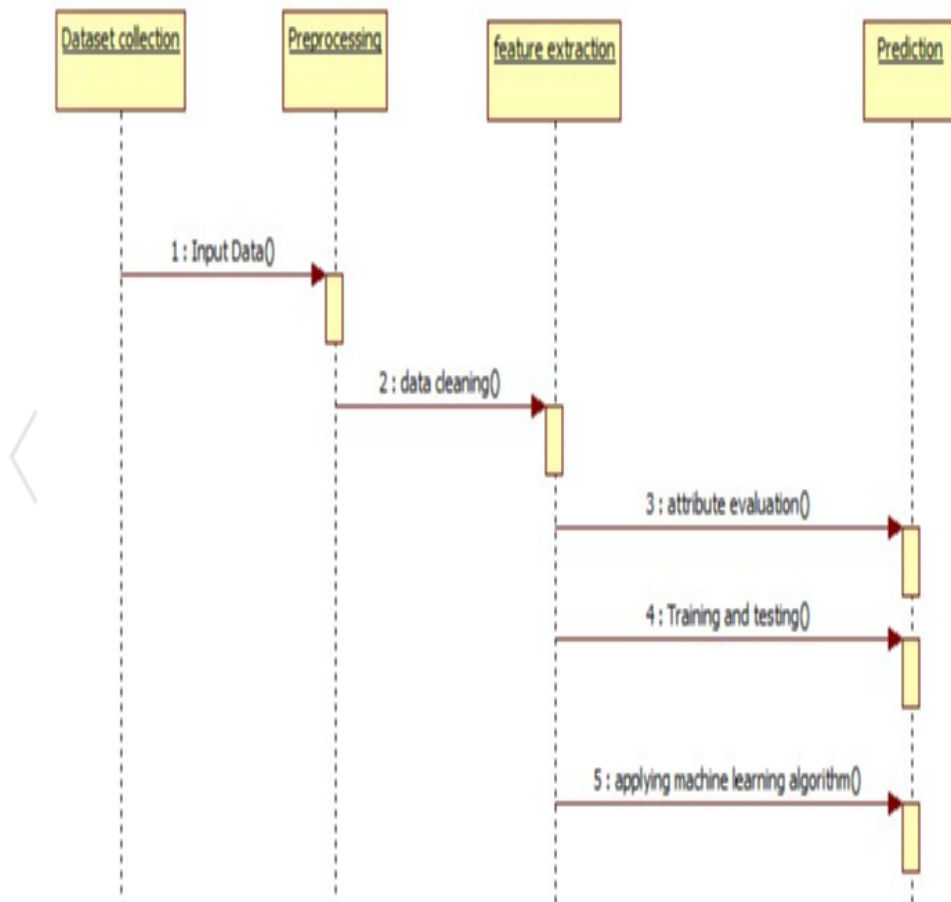
29

system can be depicted.



SEQUENCE DIAGRAM:

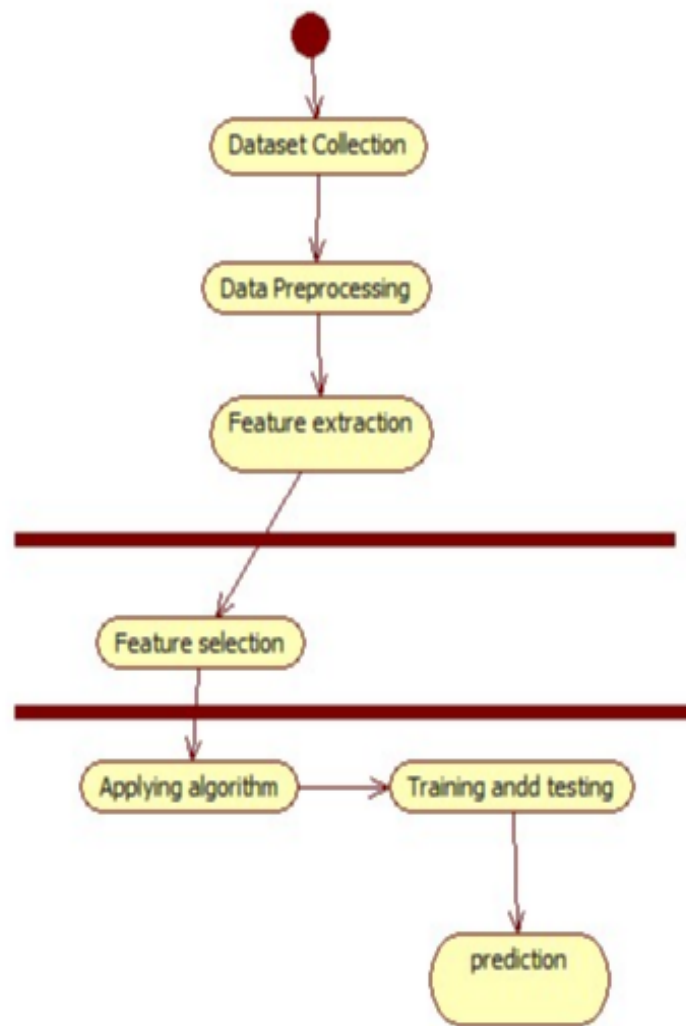
A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



31

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



CHRONIC KIDNEY DISEASE DATASET DESCRIPTION

The dataset used in this study was obtained from the UCI machine learning

32

repository. The dataset contains data of 500 instances and a total of 25 attributes

36

including the class.

3.7 MODULE

Data Preprocessing

In this stage where the data that is distorted, or encoded is brought to such a state that the machine can easily analyze it. A dataset can be observed as a group of data objects. Data objects are labeled by a number of features, that ensures the basic features of an object. In the dataset there may be missing values, they can either be eliminated or estimated. The most common method of dealing with missing values is filling them in with mean, median or mode value of respective feature.

Feature Selection

Feature Selection is the next step in the process. This technique reduces the amount of data to be loaded by extracting only the features those are required. The features are age, blood pressure, sugar, red blood cell, sodium level, hemoglobin, anemia etc. If the number of features increase the level of accuracy will also increase.

Feature Extraction

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set.

Building and Training Model:

After feature selection location and month attribute are used for training. The dataset is divided into pair of x_{train} , y_{train} and x_{test} , y_{test} . The algorithms model is imported from sklearn. Building model is done using `model.Fit(xtrain, ytrain)`. This phase would involve supervised classification methods like Random Forest Classifiers. We used a k-fold cross-validation method to test the network performance. Two different approaches were implemented, and both consisted of 10-fold cross-validation. The first approach was to randomly divide the data into 10 approximately equal portions so that each tumor category was equally present in each portion, referred to as record-wise cross-validation. The second approach was to randomly divide the data into 10 approximately equal portions where the data from a single subject could only be found in one of the sets. Each set, therefore, contained data from a couple of subjects regardless of the tumor class, referred to as subject-wise cross-validation. The second approach was implemented to test the generalization capability of the network in medical diagnostics. The generalization capability in clinical practice represents the ability to predict the diagnosis based on the data obtained from subjects from which there are no observations in the training process. Therefore, observations from individuals in the training set must not appear in the test set. If this is not the case, complex predictors can pick up a confounding relationship between identity and diagnostic status and so produce unrealistically high prediction accuracy. In order to compare the performance of our network with other state-of-the-art methods, we also tested our network without k-fold cross-validation (one test). In all the above-mentioned methods, two data portions were used for the test, two for validation, and six for training. Both datasets, normal and augmented, were tested using all the methods. The network was trained using an Adam

optimizer, with a mini-batch size equal to 16 and data shuffling in every

34

iteration. The early-stop condition that affects when the process of network training will stop corresponds to one epoch. More specifically, it was tuned to finish the training process after the one epoch, when the loss starts to increase. The regularization factor was set to 0.004, and the initial learning rate to 0.0004.

Classification

For classification we use Support Vector Machine(SVM),Decision tree, Random Forest to predict the disease and its performance. As a first step we have to import the libraries for classification and prediction. We import SVM and datasets from the scikit-learn library. NumPy for carrying out efficient mathematical computations. Accuracy-score from sklearn.metrics to predict the accuracy of the model. We have divided the data into training and testing sets. Now is the time to train our SVM on the training data. scikit-learn contains the SVM library,Decision tree, Random Forest, which contains built-in classes for various SVM algorithms. Since we are going to perform a classification task, we will use the algorithms class works by mapping data to a high-dimensional feature space so that data points can be classified, even when the data are not otherwise linearly separable.

Support Vector Machine

In machine learning, Support Vector Machine (SVM) are supervised learning models with related learning algorithms that examine data used for classification and regression analysis. SVM training algorithm builds a model that assigns new examples to one category or the other, making it a nonprobabilistic binary linear classifier. SVM works by mapping data to a high-dimensional feature space so that data points can be classified, even when the data are not otherwise

linearly separable.

35

The followings are important concepts in SVM –

Support Vectors – Datapoints that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.

Hyperplane – As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

Margin – It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

Random forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

40

We can understand the working of Random Forest algorithm with the help of following steps

Step 1 – First, start with the selection of random samples from a given dataset.

Step 2 – Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

36

Step 3 – In this step, voting will be performed for every predicted result.

Step 4 – At last, select the most voted prediction result as the final prediction result.

Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

Root node: The topmost node in a tree.

Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node

Splitting: It is a process of dividing a node into two or more sub-nodes.

Parent and Child Node: A node, which is divided into sub-nodes is called the parent node of sub-nodes whereas sub-nodes are the child of the parent node.

Decision Node: When a sub-node splits into further sub-nodes, then it is called a decision node.

SOFTWARE TESTING

General

In a generalized way, we can say that the system testing is a type of testing in which the main aim is to make sure that system performs efficiently and seamlessly. The process of testing is applied to a program with the main aim to discover an unprecedented error, an error which otherwise could have damaged

the future of the software. Test cases which brings up a high possibility of discovering an error is considered successful. This successful test helps to answer the still unknown errors.

TEST CASE

Testing, as already explained earlier, is the process of discovering all possible weak-points in the finalized software product. Testing helps to counter the working of sub-assemblies, components, assembly and the complete result. The

37

software is taken through different exercises with the main aim of making sure that software meets the business requirement and user-expectations and doesn't fail abruptly. Several types of tests are used today. Each test type addresses a specific testing requirement.

Testing Techniques

A test plan is a document which describes approach, its scope, its resources and the schedule of aimed testing exercises. It helps to identify almost other test item, the features which are to be tested, its tasks, how will everyone do each task, how much the tester is independent, the environment in which the test is taking place, its technique of design plus the both the end criteria which is used, also rational of choice of theirs, and whatever kind of risk which requires emergency planning. It can be also referred to as the record of the process of test planning. Test plans are usually prepared with signification input from test engineers.

(I) UNIT TESTING

In unit testing, the design of the test cases is involved that helps in the validation of the internal program logic. The validation of all the decision branches and internal code takes place. After the individual unit is completed it takes place.

Plus it is taken into account after the individual unit is completed before integration. The unit test thus performs the basic level test at its component stage and test the particular business process, system configurations etc. The unit test ensures that the particular unique path of the process gets performed precisely to the documented specifications and contains clearly defined inputs with the results which are expected.

(II) INTEGRATION TESTING

These tests are designed to test the integrated software items to determine whether if they really execute as a single program or application. The

38

testing is event driven and thus is concerned with the basic outcome of field. The Integration tests demonstrate that the components were individually satisfaction, as already represented by successful unit testing, the components are apt and fine. This type of testing is specially aimed to expose the issues that come-up by the components combination.

(III) FUNCTIONAL TESTING

The functional tests help in providing the systematic representation that functions tested are available and specified by technical requirement, documentation of the system and the user manual.

(IV) SYSTEM TESTING

System testing, as the name suggests, is the type of testing in which ensure that the software system meet the business requirements and aim. Testing of the configuration is taken place here to ensure predictable result and thus analysis of it. System testing is relied on the description of process and its flow, stressing on pre driven process and the points of integration.

(V) WHITE BOX TESTING

The white box testing is the type of testing in which the internal components of the system software is open and can be processed by the tester. It is therefore a complex type of testing process. All the data structure, components etc. are tested by the tester himself to find out a possible bug or error. It is used in situation in which the black box is incapable of finding out a bug. It is a complex type of testing which takes more time to get applied.

(VI) BLACK BOX TESTING

The black box testing is the type of testing in which the internal components of the software is hidden and only the input and output of the system is the key for the tester to find out a bug. It is therefore a simple type of testing. A programmer with basic knowledge can also process this type of testing. It is less

39

time consuming as compared to the white box testing. It is very successful for software which are less complex are straight-forward in nature. It is also less costly than white box testing.

(V) ACCEPTANCE TESTING

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

CHAPTER 5

CONCLUSION

Through the proposed algorithms SVM, Random Forest and Decision Tree are possible to identify the difficulty to diagnose kidney disease, especially when taking into account chronic kidney disease. It is necessary to develop more works and studies in this area in order to improve the identification of CKD, in this way assisting the doctors to identify this illness early and providing to the patient the possibility of recovering the renal function. As future work, we pretend to use neural networks and deep learning as a classifiers and evaluating their performance turn the classification into a multi-class classification.

REFERENCES

- Chronic Kidney Disease Prediction using Data Mining,J. Snegha, International Conference on Emerging Trends, 2020.
- Multiple Pheromone table based on Ant Colony Optimization for Clustering, Hindawi, Research article, Kai-Cheng Hu, 2015.
- Chronic Kidney Disease Prediction Using Back Propagation Neural Network Algorithm, Advances in Intelligent Systems and Computing, Nilesh Borisagar, 2017

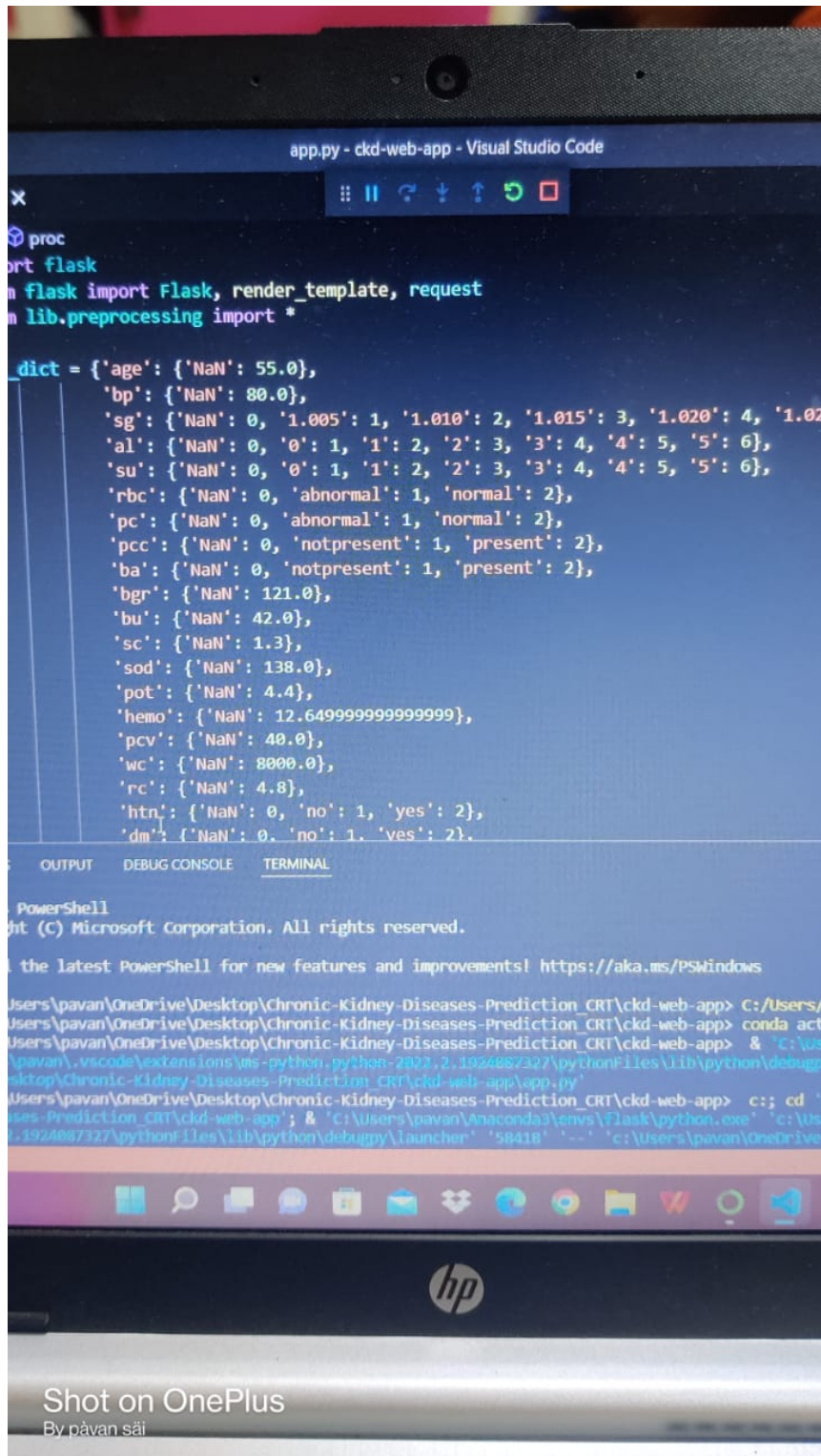
41

- Chronic Kidney Disease Prediction using Machine Learning, Reshma S , Salma Shaji , S R Ajina , Vishnu Priya S R, Janisha A, 2020, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020)
- Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Arif-Ul-Islam, Ant-Miner and J48 Decision Tree, 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE).

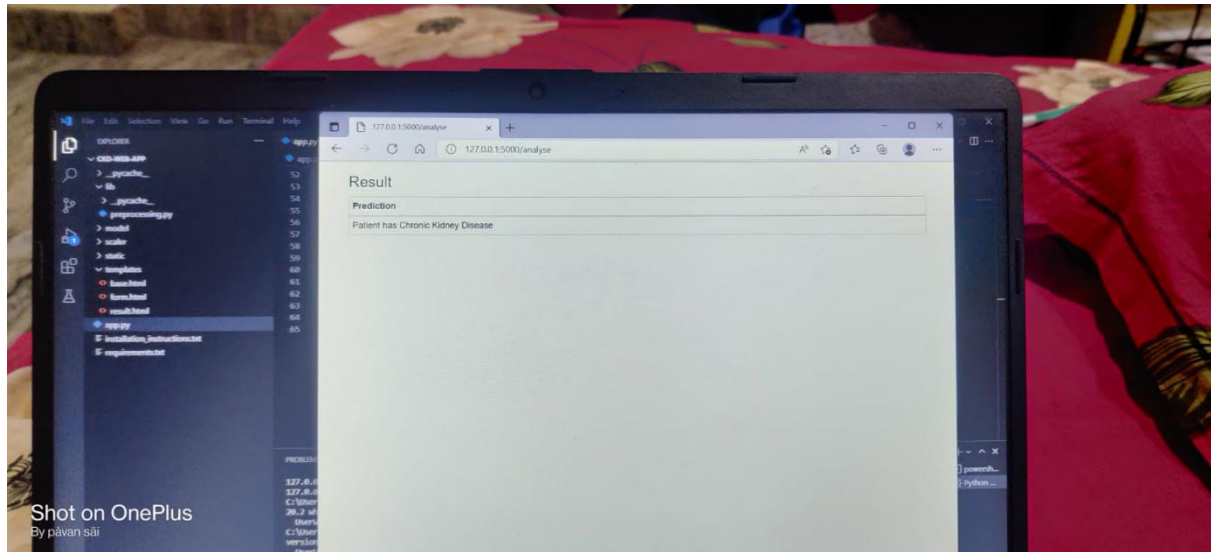
APPENDICE

42

47



B. Screenshots



45

51