**ABSTRACT**

Chronic Kidney Disease (CKD) or chronic renal disease has become a major issue with a steady growth rate. A person can only survive without kidneys for an average time of 18 days, which makes a huge demand for a kidney transplant and Dialysis. It is important to have effective methods for early prediction of CKD. Deep learning methods are effective in CKD prediction. Deep neural Network (DNN) is becoming a focal point in Machine Learning research. Its application is penetrating into different fields and solving intricate and complex problems. DNN is now been applied in health image processing to detect various ailment such as cancer and diabetes. In this project we can implement Multi-layer perceptron algorithm to classify the chronic diseases with diagnosis information. Multilayer Perceptron is a Neural Network that learns the relationship between linear and non-linear data. The Multilayer Perceptron was developed to tackle this limitation. It is a neural network where the mapping between inputs and output is non-linear. A Multilayer Perceptron has input and output layers, and one or more hidden layers with many neurons stacked together. And while in the Perceptron the neuron must have an activation function that imposes a threshold, like ReLU or sigmoid, neurons in a Multilayer Perceptron can use any arbitrary activation function. Based on this function, we can identify the chronic kidney disease from the datasets which is downloaded from KAGGLE website. Experimental results shows that the proposed system provide improved accuracy in disease prediction.

**ABSTRACT (TAMIL)**

நாள்பட்ட சிறுநீரக நோய் (CKD) அல்லது நாள்பட்ட சிறுநீரக நோய் ஒரு நிலையான வளர்ச்சி விகிதத்துடன் ஒரு பெரிய பிரச்சினையாக மாறியுள்ளது. ஒரு நபர் சராசரியாக 18 நாட்களுக்கு சிறுநீரகம் இல்லாமல் உயிர்வாழ முடியும், இது சிறுநீரக மாற்று மற்றும் டயாலிசிஸுக்கு பெரும் தேவையை உருவாக்குகிறது. சிகேடியை முன்கூட்டியே கணிக்க பயனுள்ள முறைகளைக் கொண்டிருப்பது முக்கியம். ஆழ்ந்த கற்றல் முறைகள் CKD கணிப்பதில் பயனுள்ளதாக இருக்கும். டீப் நியூரல் நெட்வொர்க் (டிஎன்என்) இயந்திர கற்றல் ஆராய்ச்சியில் ஒரு மையப் புள்ளியாக மாறி வருகிறது. அதன் பயன்பாடு பல்வேறு துறைகளில் ஊடுருவி சிக்கலான மற்றும் சிக்கலான சிக்கல்களைத் தீர்க்கிறது. புற்றுநோய் மற்றும் நீரிழிவு போன்ற பல்வேறு நோய்களைக் கண்டறிய டிஎன்என் இப்போது ஹெல்த் இமேஜ் பிராசஸிங்கில் பயன்படுத்தப்படுகிறது. இந்த திட்டத்தில், நோயறிதல் தகவலுடன் நாள்பட்ட நோய்களை வகைப்படுத்த பல அடுக்கு பெர்செப்ட்ரான் அல்காரிதம் செயல்படுத்தலாம். மல்டிலேயர் பெர்செப்ட்ரான் என்பது ஒரு நரம்பியல் நெட்வொர்க் ஆகும், இது நேரியல் மற்றும் நேரியல் அல்லாத தரவுகளுக்கு இடையிலான உறவைக் கற்றுக்கொள்கிறது. இந்த வரம்பைச் சமாளிக்க மல்டிலேயர் பெர்செப்ட்ரான் உருவாக்கப்பட்டது. இது ஒரு நரம்பியல் வலையமைப்பாகும், இதில் உள்ளீடுகள் மற்றும் வெளியீடுகளுக்கு இடையிலான மேப்பிங் நேரியல் அல்ல. ஒரு மல்டிலேயர் பெர்செப்ட்ரான் உள்ளீடு மற்றும் வெளியீட்டு அடுக்குகளைக் கொண்டுள்ளது, மேலும் பல நியூரான்கள் ஒன்றாக அடுக்கப்பட்ட ஒன்று அல்லது அதற்கு மேற்பட்ட மறைக்கப்பட்ட அடுக்குகள் உள்ளன. பெர்செப்ட்ரானில் இருக்கும் போது, ​​நியூரானானது, ReLU அல்லது sigmoid போன்ற வாசலைச் செலுத்தும் ஒரு செயல்படுத்தும் செயல்பாட்டைக் கொண்டிருக்க வேண்டும், பல அடுக்கு பெர்செப்ட்ரானில் உள்ள நியூரான்கள் எந்த தன்னிச்சையான செயல்படுத்தும் செயல்பாட்டையும் பயன்படுத்தலாம். இந்தச் செயல்பாட்டின் அடிப்படையில், KAGGLE இணையதளத்தில் இருந்து தரவிறக்கம் செய்யப்பட்ட தரவுத்தொகுப்பில் இருந்து நாள்பட்ட சிறுநீரக நோயை நாம் கண்டறியலாம். முன்மொழியப்பட்ட அமைப்பு நோய் முன்னறிவிப்பில் மேம்பட்ட துல்லியத்தை வழங்குகிறது என்பதை பரிசோதனை முடிவுகள் காட்டுகின்றன.

**CHAPTER 1**

**1. INTRODUCTION**

**1.1 BIG DATA**

**Big data** is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

**1.2 BIG DATA CHARACTERSTICS**

Big data refers to massive complex structured and unstructured data sets that are rapidly generated and transmitted from a wide variety of sources. These attributes make up the [three Vs of big data](https://linuxacademy.com/blog/cloud/what-is-big-data-the-three-vs/):

1. **Volume:** The huge amounts of data being stored.
2. **Velocity:**The lightning speed at which data streams must be processed and analyzed.
3. **Variety:** The different sources and forms from which data is collected, such as numbers, text, video, images, audio and text.

Big data is essentially the wrangling of the three Vs to gain insights and make predictions, so it's useful to take a closer look at each attribute.

**Volume:** Big data is enormous. While traditional data is measured in familiar sizes like megabytes, gigabytes and terabytes, big data is stored in petabytes and zettabytes.  To grasp the enormity of difference in scale, consider this comparison from the [Berkeley School of Information](https://datascience.berkeley.edu/big-data-infographic/): one gigabyte is the equivalent of a seven minute video in HD, while a single zettabyte is equal to 250 billion DVDs. This is just the tip of the iceberg. According to a report by EMC, the digital universe is [doubling in size every two years](https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm) and by 2020 is expected to reach 44 trillion zettabytes.  Big data provides the architecture handling this kind of data. Without the appropriate solutions for storing and processing, it would be impossible to mine for insights.

**Velocity**

From the speed at which it's created to the amount of time needed to analyze it, everything about big data is fast. Some have described it as trying to [drink from a fire hose](https://www.kinetica.com/blog/drinking-data-fire-hose/).  Companies and organizations must have the capabilities to harness this data and generate insights from it in real-time, otherwise it's not very useful. Real-time processing allows decision makers to act quickly, giving them a leg up on the competition.  While some forms of data can be batched processed and remain relevant over time, much of big data is streaming into organizations at a clip and requires immediate action for the best outcomes. Sensor data from health devices is a great example. The ability to instantly process health data can provide users and physicians with potentially life-saving information.

**Variety**

Roughly [95% of all big data is unstructured](https://www.mongodb.com/scale/unstructured-data-in-big-data), meaning it does not fit easily into a straightforward, traditional model. Everything from emails and videos to scientific and meteorological data can constitute a big data stream, each with their own unique attributes.

**1.3 USES OF BIG DATA**

The diversity of big data makes it inherently complex, resulting in the need for systems capable of processing its various structural and semantic differences.  Big data requires specialized NoSQL databases that can store the data in a way that doesn't require strict adherence to a particular model. This provides the flexibility needed to cohesively analyze seemingly disparate sources of information to gain a holistic view of what is happening, how to act and when to act. When aggregating, processing and analyzing big data, it is often classified as either [operational or analytical data](https://www.mongodb.com/big-data-explained) and stored accordingly. [Operational systems serve large batches of data](https://www.arkatechture.com/blog/the-difference-between-operational-and-analytical-data-systems) across multiple servers and includes such input as inventory, customer data and purchases — the day-to-day information within an organization. Analytical systems are more sophisticated than their operational counterparts, capable of handling complex data analysis and providing businesses with decision-making insights. These systems will often be integrated into existing processes and infrastructure to maximize the collection and use of data. Regardless of how it is classified, data is everywhere. Our phones, credit cards, software applications, vehicles, records, websites and the majority of “things” in our world are [capable of transmitting vast amounts of data](https://www.forbes.com/sites/bernardmarr/2017/03/14/the-complete-beginners-guide-to-big-data-in-2017/#70d762da7365), and this information is incredibly valuable.

Big data is used in nearly every industry to identify patterns and trends, answer questions, gain insights into customers, and tackle complex problems. Companies and organizations use the information for a multitude of reasons like growing their businesses, understanding customer decisions, enhancing research, making forecasts and targeting key audiences for advertising.

**1.4 APPLICATIONS OF BIG DATA**

### Finance

The finance and insurance industries utilize big data and predictive analytics for fraud detection, risk assessments, credit rankings, brokerage services and blockchain technology, among other uses. Financial institutions are also using big data to enhance their cybersecurity efforts and personalize financial decisions for customers.

### Healthcare

Hospitals, researchers and pharmaceutical companies are adopting big data solutions to improve and advance healthcare. With access to vast amounts of patient and population data, healthcare is enhancing treatments, performing more effective research on diseases like cancer and Alzheimer’s, developing new drugs, and gaining critical insights on patterns within population health.

### Media & Entertainment

If you've ever used Netflix, Hulu or any other streaming services that provides recommendations, you've witnessed big data at work.  Media companies analyze our reading, viewing and listening habits to build individualized experiences. Netflix even uses data on [graphics, titles and colors](https://www.wired.com/insights/2014/03/big-data-lessons-netflix/)to make decisions about customer preferences.

### Agriculture

[From engineering seeds to predicting crop yields](https://www.cio.com/article/3235141/big-data/4-ways-big-data-analytics-is-disrupting-the-agriculture-industry.html) with amazing accuracy, big data and automation is rapidly enhancing the farming industry. With the influx of data in the last two decades, information is more abundant than food in many countries, leading researchers and scientists to use big data to tackle hunger and malnutrition. With groups like the [Global Open Data for Agriculture & Nutrition](https://www.godan.info/news/open-access-open-data-help-solve-world-hunger-and-malnutrition) (GODAN) promoting open and unrestricted access to global nutrition and agricultural data, some progress is being made in the fight to end world hunger.

### More Application Areas

* [Advertising & marketing](https://builtin.com/data-science/big-data-marketing)
* [Business](https://builtin.com/data-science/big-data-business)
* [E-commerce & retail](https://builtin.com/data-science/big-data-ecommerce-retail-companies)
* [Education](https://builtin.com/data-science/big-data-in-education)
* [Internet of Things](https://builtin.com/data-science/iot-big-data-analytics-examples)
* [Sports](https://builtin.com/data-science/big-data-companies-sports)

**1.6 BIG DATA DATASETS**

### Structured

Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. **For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc.,**will be present in an organized manner.

### Unstructured

Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

### Semi-structured

Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data. Thus we come to the end of types of data. Lets discuss the characteristics of data.

## **1.7 ADVANTAGES OF BIG DATA (FEATURES)**

* One of the biggest advantages of Big Data is predictive analysis. Big Data analytics tools can predict outcomes accurately, thereby, allowing businesses and organizations to make better decisions, while simultaneously optimizing their operational efficiencies and reducing risks.
* By harnessing data from social media platforms using Big Data analytics tools, businesses around the world are streamlining their digital marketing strategies to enhance the overall consumer experience. Big Data provides insights into the customer pain points and allows companies to improve upon their products and services.
* Being accurate, Big Data combines relevant data from multiple sources to produce highly actionable insights. Almost [43% of companies](https://martech.zone/benefits-of-big-data/) lack the necessary tools to filter out irrelevant data, which eventually costs them millions of dollars to hash out useful data from the bulk. Big Data tools can help reduce this, saving you both time and money.
* Big Data analytics could help companies generate more sales leads which would naturally mean a boost in revenue. Businesses are using Big Data analytics tools to understand how well their products/services are doing in the market and how the customers are responding to them. Thus, the can understand better where to invest their time and money.
* With Big Data insights, you can always stay a step ahead of your competitors. You can screen the market to know what kind of promotions and offers your rivals are providing, and then you can come up with better offers for your customers. Also, Big Data insights allow you to learn customer behavior to understand the customer trends and provide a highly ‘personalized’ experience to them.

**1.8 BIG DATA ANALYTICS**

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today’s technology, it’s possible to analyze your data and get answers from it almost immediately – an effort that’s slower and less efficient with more traditional business intelligence solutions. Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today’s technology, it’s possible to analyze your data and get answers from it almost immediately – an effort that’s slower and less efficient with more traditional business intelligence solutions. Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. In his report Big Data in Big Companies, IIA Director of Research Tom Davenport interviewed more than 50 businesses to understand how they used big data. He found they got value in the following ways:

1. **Cost reduction.** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
2. **Faster, better decision making.** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they’ve learned.
3. **New products and services.** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers’ needs.

**1.9 TECHNOLOGIES IN BIG DATA**

[**Machine Learning**](https://www.sas.com/en_us/insights/analytics/machine-learning.html)**.** Machine learning, a specific subset of AI that trains a machine how to learn, makes it possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results – even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities – or avoiding unknown risks.

[**Data management**](https://www.sas.com/en_us/insights/data-management/data-management.html)**.** Data needs to be high quality and well-governed before it can be reliably analyzed. With data constantly flowing in and out of an organization, it's important to establish repeatable processes to build and maintain standards for data quality. Once data is reliable, organizations should establish a master data management program that gets the entire enterprise on the same page.

[**Data mining**](https://www.sas.com/en_us/insights/analytics/data-mining.html)**.** Data mining technology helps you examine large amounts of data to discover patterns in the data – and this information can be used for further analysis to help answer complex business questions. With data mining software, you can sift through all the chaotic and repetitive noise in data, pinpoint what's relevant, use that information to assess likely outcomes, and then accelerate the pace of making informed decisions.

[**Hadoop**](https://www.sas.com/en_us/insights/big-data/hadoop.html)**.** This open source software framework can store large amounts of data and run applications on clusters of commodity hardware. It has become a key technology to doing business due to the constant increase of data volumes and varieties, and its distributed computing model processes big data fast. An additional benefit is that Hadoop's open source framework is free and uses commodity hardware to store large quantities of data.

[**In-memory analytics**](https://www.sas.com/en_us/solutions/in-memory-analytics.html)**.** By analyzing data from system memory (instead of from your hard disk drive), you can derive immediate insights from your data and act on them quickly. This technology is able to remove data prep and analytical processing latencies to test new scenarios and create models; it's not only an easy way for organizations to stay agile and make better business decisions, it also enables them to run iterative and interactive analytics scenarios.

[**Predictive analytics**](https://www.sas.com/en_us/insights/analytics/predictive-analytics.html)**.** Predictive analytics technology uses data, statistical algorithms and machine-learning techniques to identify the likelihood of future outcomes based on historical data. It's all about providing a best assessment on what will happen in the future, so organizations can feel more confident that they're making the best possible business decision. Some of the most common applications of predictive analytics include fraud detection, risk, operations and marketing.

[**Text mining**](https://www.sas.com/en_us/software/text-miner.html)**.**With text mining technology, you can analyze text data from the web, comment fields, books and other text-based sources to uncover insights you hadn't noticed before. Text mining uses [machine learning](https://www.sas.com/en_us/insights/analytics/machine-learning.html) or [natural language processing](https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html) technology to comb through documents – emails, blogs, Twitter feeds, surveys, competitive intelligence and more – to help you analyze large amounts of information and discover new topics and term relationships.

**1.10 DATA MINING**

The data mining tutorial provides basic and advanced concepts of data mining. Our data mining tutorial is designed for learners and experts. Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called Knowledge Discovery in Database (KDD)*.* The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

**CHAPTER 2**

**2. LITERATURE SURVEY**

**2.1 REVIEW OF LITERATURE**

Literature review is the most important step in software development process, before developing the too, it’s necessary to determine the time factor, economy and company strength. Once these things are satisfied, the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start to build the tool the programmers need lot of external support. This support can be determined from senior programmers, from book or from websites. Before building the system and above consideration are consider for developing the system.

**2.2 TITLE: COMPARATIVE STUDY OF CHRONIC KIDNEY DISEASE PREDICTION USING KNN AND SVM**

**AUTHOR: PARUL SINHA**

Data mining deals with extraction of useful information from huge amounts of data. Many other terms are being used to understand data mining, such as mining of knowledge from databases, knowledge extraction, data analysis, and data archaeology. Basically, data mining is a crucial step in the process of knowledge discovery in databases, or KDD. The data mining techniques of classification, clustering and association helps in extracting knowledge from large amount of data. Machine Learning is a rising field concerned with the study of huge and multiple variable data. It is evolved from the study of pattern recognition and computational learning theory in artificial intelligence and involves computational methods, algorithms and techniques for analysis. In Medical Science’s perspective, Machine Learning promises to aid physicians make near-perfect diagnoses, opt the best medications for their patients, spot patients at high-risk for pitiable outcomes, and specifically improving patients’ physical condition while minimizing costs. In this paper, a new decision support system is implemented for prediction of CKD. Although the classifiers worked efficiently in prediction of other diseases also. In this paper, Chronic Kidney Disease is predicted using two different classifiers and a comparative study of their performance is done. From the analysis we found that, out of two classifiers SVM and KNN, KNN classifier performed better than the other. The rate of prediction of CKD is improved.

**2.3 TITLE: CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING MODELS**

**AUTHOR: S.REVATHY, B.BHARATHI**

The disability of the kidneys to perform their regular blood filtering function and others is called Chronic Kidney Disease (CKD). The term “chronic” describes the slow degradation of the kidney cells over a long period of time. This disease is a major kidney failure where the kidney sans blood filtering process and there is a heavy fluid buildup in the body. This leads to alarming increase of potasium and calcium salts in the body. Existence of high levels of these salts result in various other ailments in the body. This paper presented a prediction algorithm to predict CKD at an early stage. The dataset shows input parameters collected from the CKD patients and the models are trained and validated for the given input parameters. Decision tree, Random Forest and Support Vector Machine learning models are constructed to carry out the diagnosis of CKD.The performance of the models are evaluated based on the accuracy of prediction. The results of the research showed that Random Forest Classifier model better predicts CKD in comparison to Decision trees and Support Vector machines. The comparision can also be done based on the time of execution, feature set selection as the improvisation of this research. Random forest algorithm constructs multiple decision trees to act as an ensemble of classification and regression process. A number of decision trees are constructed using a random subsets of the training data sets. A large collection of decision trees provide higher accuracy of results. The runtime of the algorithm is comparatively fast and also accommodates missing data. Random forest randomizes the algorithm and not the training data set. The decision class is the mode of classes generated by decision trees.

**2.4 TITLE: CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING**

**AUTHOR: SATHIYA PRIYA S**

Computer vision has been one of the most remarkable breakthroughs for the machine learning and in particular for active healthcare applications. Machine learning allows to build the models to quickly analyse data and deliver results for the given data. Healthcare service providers can make better decisions on patient’s disease diagnosis and treatment for the particular disease with the help of machine learning. The massive quantities of data are analysed using machine learning. It delivers faster and more accurate results in order to identify the risks, it may also require additional time and resources to train it proper manner. Supervised machine learning algorithms can applied to predict the future events with the help of what has been learned in the past to new data using labelled examples. First the known training dataset is analysed, with that the learning algorithm produces an inferred function to make predictions about the output values. The prediction of chronic kidney disease is very important and now-a-days it is the leading cause of death. The performance of Decision tree method was found to be 99.25% accurate compared to naive Bayes method. Classification algorithm on chronic kidney disease dataset the performance was obtained as 99.33% Specificity and 99.20% Sensitivity. We are also further working on enhancing the performance of prediction system accuracy in neural network and deep learning algorithm. The prediction of chronic kidney disease is very important and now-a-days it is the leading cause of death. The performance of Decision tree method was found to be 99.25% accurate compared to naive Bayes method. Classification algorithm on chronic kidney disease dataset the performance was obtained as 99.33% Specificity and 99.20% Sensitivity. We are also further working on enhancing the performance of prediction system accuracy in neural network and deep learning algorithm.

**2.5 TITLE: PREDICTION OF CHRONIC KIDNEY DISEASES USING DEEP ARTIFICIAL NEURAL NETWORK TECHNIQUE**

**AUTHOR: HIMANSHU KRIPLANI**

The deep learning model built using artificial neural networks are considered very robust. The network consists of interconnected neurons which are mapped using mathematical functions to each other. To use a NN we have to first train the model which is finding of the proper weights involved in the mathematical function for mapping. In the second phase we test our model with already classified dataset to find the accuracy of our dataset. Neurons have a basis in Biology. This neuron in deep learning is termed as perceptron but both are used interchangeably. A single neuron takes inputs with weights associated with each weight. We use activation function for the output of the neuron. We trained a neural network model to predict the presence of chronic kidney disease. It uses a list of parameters. If this model can be well trained using a varied range of parameters, it may result in more accurate predictions. Clinics and hospitals can use this for faster and digitized methodology for prediction of chronic kidney disease. Of all the other models compared, deep neural network is the best one. The results would be more promising with increase in dataset. Thus it has outperformed other classifiers and is able to detect the chronic kidney disease more efficiently. But this AdaBoost is sensitive to noisy data, outliers and susceptible to the over fitting problem than proposed algorithms. Support vector machine has performed well but suffers from the determination of the parameters for a given value of the regularization and kernel parameters and choice of kernel. Here SVM moves the problem of over fitting from optimizing the parameters to model selection.

**2.6 TITLE: PREEMPTIVE DIAGNOSIS OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING TECHNIQUES**

**AUTHOR: REEM A. ALASSAF**

Chronic Kidney Disease (CKD) is a major public health concern with rising prevalence. Kidney disease is when the kidneys are damaged and could not filter the blood properly. This damage could cause the wastes to build up in the body. There are five stages of CKD, the most serious one is stage 5 because, at this stage, the kidneys are unable to do most of their functions. It is difficult to pinpoint the CKD stage of each patient especially at the early-stages. It also causes a high possibility of death within a short period of time, a patient must be hospitalized and appropriately cured. The most common causes of kidney disease are diabetes and high blood pressure. In this study, Saudi medical records were investigated for the first time in the process of diagnosing CKD using machine learning techniques. Authors used correlation coefficient and recursive feature elimination for feature selection. Then, four classification algorithms were explored, namely: ANN, SVM, Naïve Bayes, and k-NN. The performance of each of these classifiers was examined by the classification accuracy, precision, recall, and f-measure achieved by the classifier. ANN, SVM, and NB all achieved an accuracy of 98% while k-NN achieved an accuracy of 93.9%. Further research can be done to exceed the classification accuracy currently achieved, by using different classifiers or feature selection methods. To achieve models that can optimally solve a classification problem, the parameters of the learning algorithms have to be tuned. For that purpose, exhaustive grid search with cross validation was used. Grid search is used to set the search space with the specified parameters and their range of possible values. Then it generates all the possible combinations of the parameters’ values to search the parameters simultaneously. For each combination, 10- fold cross validation was used for validation and the average of the models’ accuracy is calculated.

**2.7 TITLE: DETECTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING ALGORITHMS WITH LEAST NUMBER OF PREDICTORS**

**AUTHOR: MARWA ALMASOUD**

This work examines the ability to detect CKD using machine learning algorithms while considering the least number of tests or features. We approach this aim by applying four machine learning classifiers: logistic regression, SVM, random forest, and gradient boosting on a small dataset of 400 records. In order to reduce the number of features and remove redundancy, the association between variables have been studied. A filter feature selection method has been applied to the remaining attributes and found that there are hemoglobin, albumin, and specific gravity have the most impact to predict the CKD. The classifiers have been trained, tested, and validated using 10-fold cross-validation. Higher performance was achieved with the gradient boosting algorithm by F1-measure (99.1 %), sensitivity (98.8%), and specificity (99.3%). This result is the highest among previous studies with less number of features and hence less cost. Therefore, we conclude that CKD can be detected with only three features. Also, we found that hemoglobin has the highest contribution in detecting CKD, whereas albumin has the lowest using RF and GB models. Since the data used in this research is small, in the future, we aim to validate our results by using big dataset or compare the results using another dataset that contains the same features. Also, in order to help in reducing the prevalence of CKD, we plan to predict if a person with CKD risk factors such as diabetes, hypertension, and family history of kidney failure will have CKD in the future or not by using appropriate dataset.

**2.8 TITLE: CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING METHODS**

**AUTHOR: IMESH UDARA EKANAYAKE**

The data distribution has properly covered the whole domain in CKD, but the general attributes like appetite, anaemia and pedal oedema are biased towards CKD. It is easy to achieve an accurate prediction using this data set but in the general context, it may lead to false positives as observed in the recall column. Further, the missing values which were completely missed at random made it impossible to achieve a perfect accuracy without filling them from a collaborative imputer instead of a constant. Considering the medical importance of the attributes, some of them have a lesser co-relation compared to others because of the stage they appear in the patient. When training the models it makes a huge impact on the accuracy. After training the model, it clearly shows that tree structures have higher accuracy than other classification algorithms, which can be justified from the distribution of the data set since the selected attributes have a clearer separation in the class except for serum creatinine attribute. Finally, when selecting the algorithm, some trained models have a bias towards some attributes, considering the causes of change of the nominal values of them, it has many different possibilities apart from CKD. Therefore, it motivates to rely less on one attribute and consider more when making the decision and based on that the extra tree classifier has been selected. This work suggests a new workflow including data preprocessing, missing values handling and features selection to predict CKD status as positive or negative. Furthermore, this work highlights the importance of incorporating the domain knowledge into feature selection when analysing clinical data related to CKD. Accordingly, it is worthwhile to explore the use of KNN imputer based approach to handle missing values in data sets related to multiple diseases in future. Furthermore, more insights into CKD can be gained by adding knowledge of genomics, water consumption patterns and food types into the analysis.

**2.9 TITLE: A NOVEL APPROACH TO PREDICT CHRONIC KIDNEY DISEASE USING MACHINE LEARNING ALGORITHMS**

**AUTHOR: BHAVYA GUDETI**

Aimed to diagnose Chronic Kidney Disease (CKD) at an earlier stage, this manuscript introduced a variety of machine learning algorithms. The models obtained from CKD patients are trained and authenticated with the mentioned input parameters. Support Vector Machine, Logistic Regression and knn are analyzed to conduct the study of CKD. The performances of those algorithms were determined primarily on the basis of precision. Our results exemplified that the Support Vector Machine algorithm predicts Chronic Kidney Disease better than Logistic Regression and K-Nearest Neighbors within the narrow limits of this medical scenario. The benefit of this approach is that the prediction process takes far less time and helps doctors to initiate treatment at the earliest for patients with CKD and further to classify larger population of patients within shorter span. Because the dataset used in this paper is tiny with 400 examples, we prefer to work with larger datasets in the future or compare the results of this dataset with a different dataset with the same. In addition, to help minimise the incidence of CKD, we try to predict if a person with this syndrome chances chronic risk factors such as hypertension, family history of kidney failure and diabetes using the appropriate dataset. Each classifier's results were evaluated using different evaluation parameters, and cross-checked against over-fitting with 10-fold cross-validation. The technique of nested crossvalidation has also helped to fine-tune the model parameters. The tests will be carried out using the Python 3.3 programming language through the Jupyter Notebook web application. Several Sciket-learning libraries were used, which is a free machine learning system platform for Python. Accuracy using F1-measurement, sensitivity, specificity and Area under Curve (AUC) are the assessment measures considered in this analysis. Each model produces different outputs; depending on its parameter values. Thus with the GB model we achieve the best efficiency in detection

**2.10 TITLE: APPLYING MACHINE LEARNING TECHNIQUES FOR PREDICTING THE RISK OF CHRONIC KIDNEY DISEASE**

**AUTHOR: K. R. ANANTHA PADMANABAN**

Data mining is becoming more popular nowadays in healthcare, as also in fraud, abuse detection etc. In classification is a more useful data mining function to handle items in a collection to target categories or classes. In kidney failure falls one among several classes viz heart disease, blindness etc which results due to chronic Diabetes. Dialysis is the only method to keep the kidneys function artificially and it is also painful and expensive process. According to World Health Organization about millions of people around the world are suffering from severe kidney disorder and its number is increasing every year. Therefore, an early diagnosing technique is immediately required so that precautions or controls can be taken before hand in time. For obtaining essential information from medical databases Data mining technique was found very much useful. By combining machine learning and statistical analysis intelligently, very useful information can be drawn from medical databases. Machine learning methods which coordinates various statistical analyses and databases helps us to extract hidden patterns and relationships from huge and multiple variable data. In order to ensure the chosen classifier’s accuracy the available test phases are verified. Moreover, these attributes like Specificity, sensitivity, and accuracy are common for disease detection. By applying Naïve Bayes and Decision tree techniques for our desired classification methods, we were able to achieve the result of identification of Kidney disorder at the early stages. We arrived to a conclusion that application of Data mining technique for different analysis of medical data is a good method. The performance of Decision tree method was found to be 91% accurate compared to naive Bayes method. Classification algorithm on diabetes dataset performance was obtained as94% Specificity and 95% Sensitivity. We also found that mining helps to retrieve correlations from attributes which are not direct indicators of the class which we are trying to predict. We are also further working on enhancing the performance of prediction system accuracy in neural network and clustering algorithm data analysis

**2.11 TITLE: A MACHINE LEARNING METHODOLOGY FOR DIAGNOSING CHRONIC KIDNEY DISEASE**

**AUTHOR: JIONGMING QIN**

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After unsupervised imputation of missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy. Hence, we speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. However, in the process of establishing the model, due to the limitations of the conditions, the available data samples are relatively small, including only 400 samples. Therefore, the generalization performance of the model might be limited. In addition, due to there are only two categories (ckd and notckd) of data samples in the data set, the model can not diagnose the severity of CKD. In the future, a large number of more complex and representative data will be collected to train the model to improve the generalization performance while enabling it to detect the severity of the disease. We believe that this model will be more and more perfect by the increase of size and quality of the data. The final decision is determined by the predictions of all trees in the disease diagnosis. SVM divides different kinds of samples by establishing a decision surface in a multidimensional space that comprises the predictors of the samples. KNN finds the nearest training samples by calculating the distances between the test sample and the training samples and then determines the diagnostic category by voting. Naive Bayes classifier calculates the conditional probabilities of the sample under the interval by the number of ckd and notckd samples in each different measurement interval. FNN can analyse non-linear relationships in the data sets due to its complex structure, and the sigmoid activation function was used in the hidden layer and the output layer

**CHAPTER 3**

**3. SYSTEM ANALYSIS AND DESIGN**

**3.1 EXISTING SYSTEM**

Chronic kidney disease (CKD) is a major public health concern around the world, with negative outcomes such as renal failure, cardiovascular disease, and early death. In machine learning, the quality and quantity of input data that is used for training the classifiers are very important. Most algorithms perform well when the prior probabilities of the target classes are similar. Data is said to be imbalanced if at least one of the target variable values has a significantly smaller number of instances when compared to the other values. Class imbalance is one of the vital issues in machine learning classification tasks. Machine learning algorithms trained on imbalanced data emphasize exploiting the total accuracy over the entire dataset leading to more attention being paid to the majority class samples. Due to this scenario, the minority class samples are poorly projected by the learning model. Both data mining and classification techniques are applied in chronic kidney disease prediction. Models like SVM, Decision tree, K-NN, Naive Bayes, neural networks are used for the prediction of the diseases. In random forest gave the best accuracy compared to decision tree and SVM. In Eleven techniques of decision trees like decision Stump, J48, CTC, LMT, NBTree, Random Forest, randomTree, REPTree, simple Cart, J48graft are applied to the dataset. Random forest outperforms other methods. Neural networks with three layers performed better compared with other models like SVM, DT, KNN, and Gradient Boost algorithm.

**3.1.1 DISADVANTAGES**

* Labelled data based disease classification
* Provide high number of false positive
* Binary classification can be occurred
* Computational complexity

**3.2 PROPOSED SYSTEM**

The kidney is a very important organ in the human body whose main functions include osmoregulation and excretion. All the harmful and unwanted materials from the body are gathered and excreted by the kidney and excretory system. In India, every year there are around 1 million cases of Chronic Kidney Disease (CKD), also called renal failure. It can cause a loss in kidney functionality and hence it can pose lots of danger. CKD is a rather slow and timely loss of kidney function over a very long duration of time. A person can develop permanent kidney failure. If one fails to diagnose CKD at an earlier stage then the patient can show symptoms of weak bones, anaemia, nerve damage, etc. Therefore it is the need of the hour to detect this disease in the early stages but since the symptoms become unpredictable and at times are not specific to the disease. Thus for diagnosis in people who show no symptoms at all machine learning can be of help. It uses an old CKD patient data set to train models. In this project we implement disease prediction with diagnosis information using deep learning algorithm. Incorporating the techniques of classification in these intelligent systems achieve at accurate diagnosis. Neural Networks has emerged as an important method of classification. Multi-layer Perceptron Neural Network with Back-propagation has been employed as the training algorithm in this work. This project proposes a diagnostic system for predicting heart disease with improved accuracy. The propagation algorithm has been repeated until minimum error rate was observed. And it is quite evident from the results presented in the previous section that the accuracy rate is maximized.

**3.2.1 ADVANTAGES**

* Accuracy is high
* Parallel processing
* Provide disease with diagnosis information
* Reduce number of false positive rate

**3.3 FEASIBILITY STUDY**

A feasibility study is an analysis that takes all of a project's relevant factors into account—including economic, technical, legal, and scheduling considerations—to ascertain the likelihood of completing the project successfully. Project managers use feasibility studies to discern the pros and cons of undertaking a project before they invest a lot of time and money into it.

**3.3.1 ECONOMICAL FEASIBILTY STUDY**

For any system if the expected benefits equal or exceed the expected costs, the system can be judged to be economically feasible. In economic feasibility, cost benefit analysis is done in which expected costs and benefits are evaluated. Economic analysis is used for evaluating the effectiveness of the proposed system. In economic feasibility, the most important is cost-benefit analysis.

**3.3.2 OPERATION FEASIBILITY STUDY**

Operational feasibility is dependent on human resources available for the project and involves projecting whether the system will be used if it is developed and implemented. Operational feasibility is a measure of how well a proposed system solves the problems, and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development.

**3.3.2 TECHNICAL FEASIBILITY STUDY**

A technical feasibility study assesses the details of how you intend to deliver a product or service to customers. Think materials, labor, transportation, where your business will be located, and the technology that will be necessary to bring all this together. It's the logistical or tactical plan of how your business will produce, store, deliver, and track its products or services.

**3.3.3 SOCIAL FEASIBILITY STUDY**

Social feasibility is one of the feasibility study where the acceptance of the people is considered regarding the product to be launched. It describes the effect on users from the introduction of the new system considering whether there will be a need for retraining the workforce. It describes how you propose to ensure user co-operation before changes are introduced.

**3.4 SYSTEM SPECIFICATION**

A system requirement specification also known as software requirement specification is a document that describes the feature and behaviour of the system or software application

**3.4.1 HARDWARE REQUIREMENTS**

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design.

* Processor : Intel processor 2.6.0 GHZ
* RAM : 1GB
* Hard disk : 160 GB
* Compact Disk : 650 Mb
* Keyboard : Standard keyboard
* Monitor : 15 inch color monitor

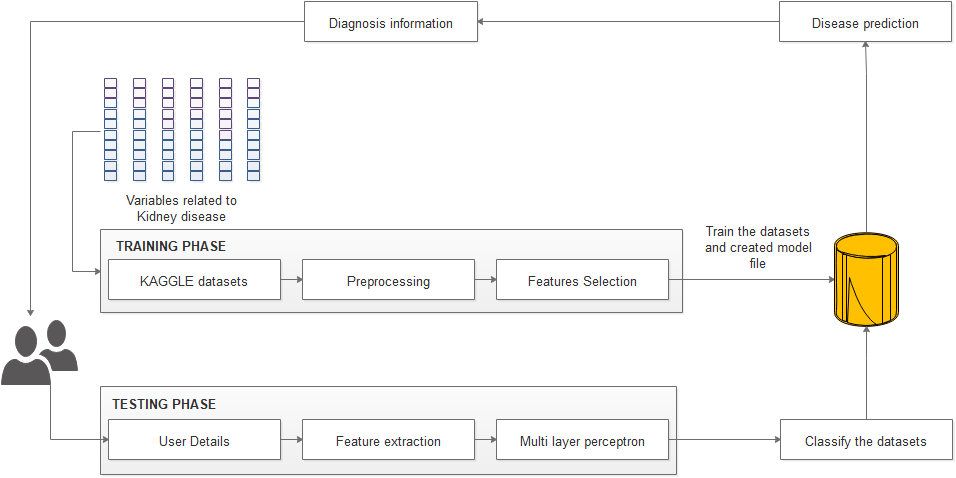
**3.4.2 SOFTWARE REQUIREMENTS**

The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is useful in estimating cost, planning team activities and performing tasks throughout the development activity.

* Operating System : Windows OS
* Front End : PYTHON
* Back End : MYSQL
* IDE : PYCHARM
* Application : WINDOWS APPLICATION

**3.5 SYSTEM DESIGN**

**3.5.1 SYSTEM ARCHITECTURE DIAGRAM**



**3.5.2 DATA FLOW DIAGRAM**

A two-dimensional diagram explains how data is processed and transferred in a system. The graphical depiction identifies each source of data and how it interacts with other data sources to reach a common output. Individuals seeking to draft a data flow diagram must identify external inputs and outputs, determine how the inputs and outputs relate to each other, and explain with graphics how these connections relate and what they result in. This type of diagram helps business development and design teams visualize how data is processed and identify or improve certain aspects.

**Data flow Symbols:**

|  |  |
| --- | --- |
| **Symbol** | **Description** |
| http://cpanel.stpaulsscience.org/gceict/specifications/ocr/unit3/sdlc/dfd/entity.jpg | An **entity**. A source of data or a destination for data. |
| http://cpanel.stpaulsscience.org/gceict/specifications/ocr/unit3/sdlc/dfd/process.jpg | A **process** or task that is performed by the system. |
| http://cpanel.stpaulsscience.org/gceict/specifications/ocr/unit3/sdlc/dfd/store.jpg | A **data store**, a place where data is held between processes. |
| http://cpanel.stpaulsscience.org/gceict/specifications/ocr/unit3/sdlc/dfd/flow.jpg | A **data flow**. |

**LEVEL 0**

Input: Kidney disease datasets

Output: Stored in database

**LEVEL 1**

Input: Datasets

Output: Preprocessed datasets

**LEVEL 2**

Features Selection

Input: Preprocessed data

Output: Model file created

**LEVEL 3**

MLP algorithm

Input: Trained Database

Output: Disease prediction

**LEVEL 4**

MLP algorithm

Input: Kidney disease name

Output: Accuracy

**3.5.3 USECASE DIAGRAM**

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different [use cases](https://en.wikipedia.org/wiki/Use_case) in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well.

****

**3.5.4 CLASS DIAGRAM**

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application. Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modeling of object oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages. Class diagram shows a collection of classes, interfaces, associations, collaborations, and constraints. It is also known as a structural diagram.

****

**3.5.5 SEQUENCE DIAGRAM**

A Sequence diagram is an [interaction diagram](https://en.wikipedia.org/wiki/Interaction_diagram) that shows how objects operate with one another and in what order. It is a construct of a [message sequence chart](https://en.wikipedia.org/wiki/Message_sequence_chart). A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios

****

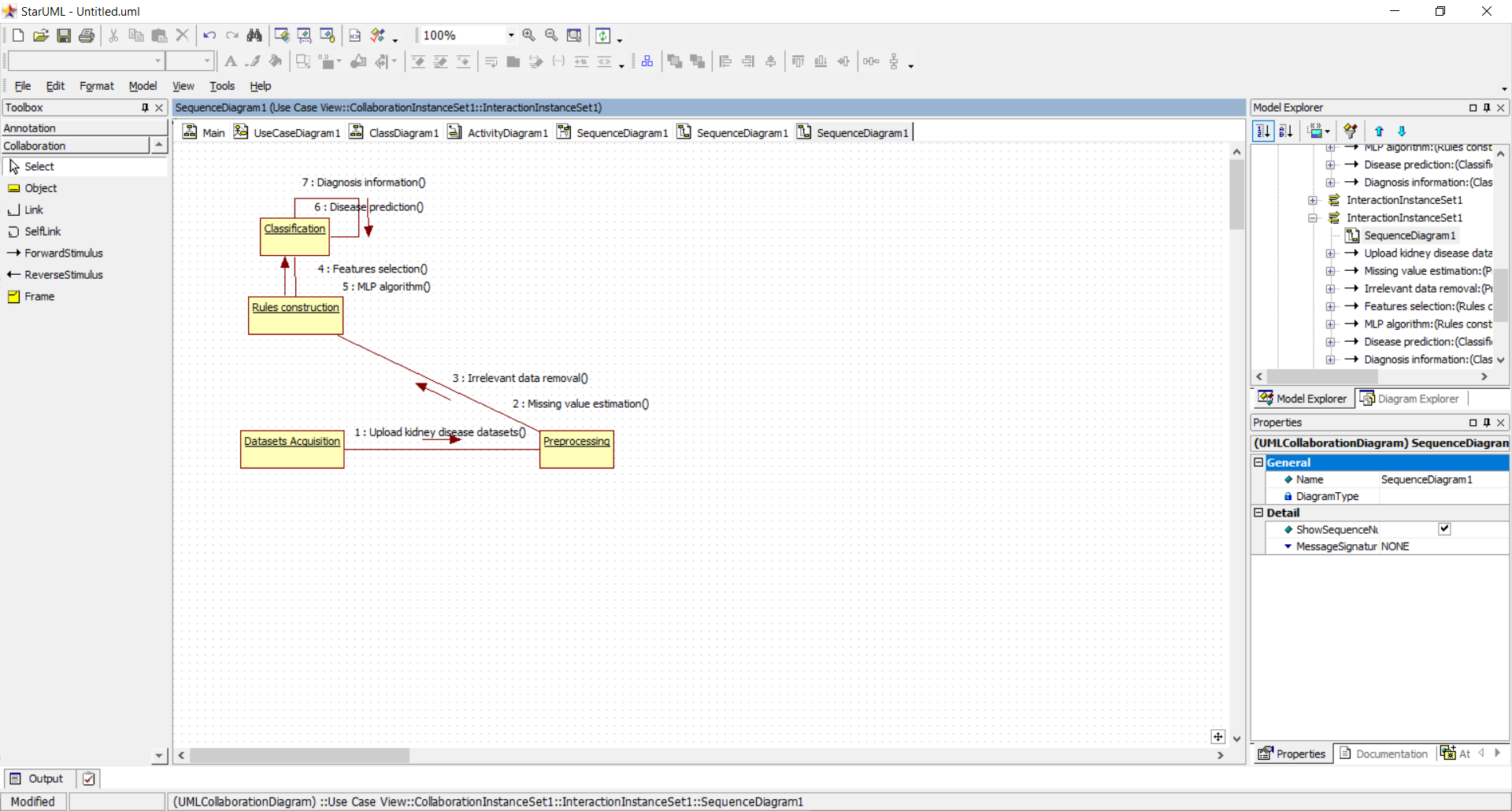
**3.5.5 ACTIVITY DIAGRAM**

Activity diagram is another important diagram in UML to describe dynamic aspects of the system. Activity diagram is basically a flow chart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. So the control flow is drawn from one operation to another. This flow can be sequential, branched or concurrent. Activity diagrams deals with all type of flow control by using different elements like fork, join etc.

****

**3.5.6 COLLABORATION DIAGRAM**

A collaboration diagram, also called a communication diagram or interaction diagram, is an illustration of the relationships and interactions among [software](http://searchsoa.techtarget.com/definition/software) [object](http://searchsoa.techtarget.com/definition/object)s in the Unified Modelling Language (UML). The concept is more than a decade old although it has been refined as modelling paradigms have evolved



**CHAPTER 4**

**4. SYSTEM TESTING**

**4.1 GENERAL**

Software testing is a method of assessing the functionality of a software program. There are many different types of software testing but the two main categories are dynamic testing and static testing. Dynamic testing is an assessment that is conducted while the program is executed; static testing, on the other hand, is an examination of the program's code and associated documentation. Dynamic and static methods are often used together.

Testing is a set activity that can be planned and conducted systematically. Testing begins at the module level and work towards the integration of entire computers based system. Nothing is complete without testing, as it is vital success of the system.

Testing Objectives:

There are several rules that can serve as testing objectives, they are

1. Testing is a process of executing a program with the intent of finding an error
2. A good test case is one that has high probability of finding an undiscovered error.
3. A successful test is one that uncovers an undiscovered error.

If testing is conducted successfully according to the objectives as stated above, it would uncover errors in the software. Also testing demonstrates that software functions appear to the working according to the specification, that performance requirements appear to have been met.

There are three ways to test a program

1. For Correctness
2. For Implementation efficiency
3. For Computational Complexity.

Tests for correctness are supposed to verify that a program does exactly what it was designed to do. This is much more difficult than it may at first appear, especially for large programs.

Tests used for implementation efficiency attempt to find ways to make a correct program faster or use less storage. It is a code-refining process, which reexamines the implementation phase of algorithm development. Tests for computational complexity amount to an experimental analysis of the complexity of an algorithm or an experimental comparison of two or more algorithms, which solve the same problem.

The data is entered in all forms separately and whenever an error occurred, it is corrected immediately. A quality team deputed by the management verified all the necessary documents and tested the Software while entering the data at all levels. The development process involves various types of testing. Each test type addresses a specific testing requirement. The most common types of testing involved in the development process are:

• Unit Test.

• Functional Test

• Integration Test

**4.2 TYPES OF TESTING**

**4.2.1 UNIT TESTING**

The first test in the development process is the unit test. The source code is normally divided into modules, which in turn are divided into smaller units called units. These units have specific behaviour. The test done on these units of code is called unit test. Unit test depends upon the language on which the project is developed. Unit tests ensure that each unique path of the project performs accurately to the documented specifications and contains clearly defined inputs and expected results.

**4.2.2 FUNCTIONAL TESTING**

Functional test can be defined as testing two or more modules together with the intent of finding defects, demonstrating that defects are not present, verifying that the module performs its intended functions as stated in the specification and establishing confidence that a program does what it is supposed to do.

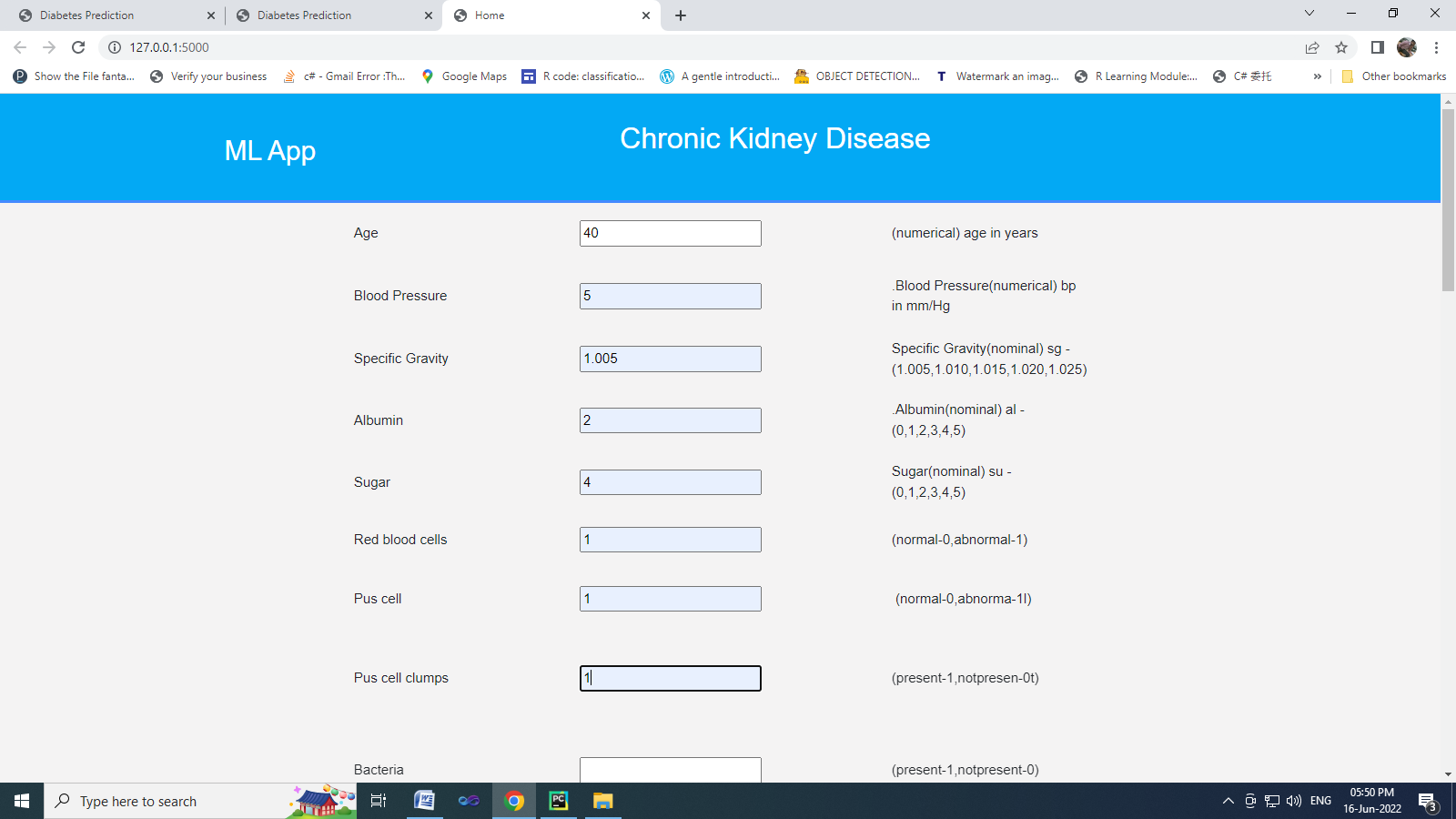
**4.2.3 INTEGRATION TESTING**

In integration testing modules are combined and tested as a group. Modules are typically code modules, individual applications, source and destination applications on a network, etc. Integration Testing follows unit testing and precedes system testing. Testing after the product is code complete. Betas are often widely distributed or even distributed to the public at large in hopes that they will buy the final product when it is released.

**4.3 TEST CASES**

* **TEST CASE 1: TEST THE DATASETS**
* **INPUT:** Load Pre-processed dataset
* **EXPECTED OUTPUT:** Type the correct format datasets

**OUTPUT**



**CHAPTER 5**

**5. SOFTWARE DESCRIPTION**

**5.1 FRONT END: PYTHON**

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. In July 2018, Van Rossum stepped down as the leader in the language community. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of Python's other implementations. Python and CPython are managed by the non-profit Python Software Foundation. Rather than having all of its functionality built into its core, Python was designed to be highly extensible. This compact modularity has made it particularly popular as a means of adding programmable interfaces to existing applications. Van Rossum's vision of a small core language with a large standard library and easily extensible interpreter stemmed from his frustrations with ABC, which espoused the opposite approach. While offering choice in coding methodology, the Python philosophy rejects exuberant syntax (such as that of Perl) in favor of a simpler, less-cluttered grammar. As Alex Martelli put it: "To describe something as 'clever' is not considered a compliment in the Python culture."Python's philosophy rejects the Perl "there is more than one way to do it" approach to language design in favour of "there should be one—and preferably only one—obvious way to do it".

Python's developers strive to avoid premature optimization, and reject patches to non-critical parts of CPython that would offer marginal increases in speed at the cost of clarity.[ When speed is important, a Python programmer can move time-critical functions to extension modules written in languages such as C, or use PyPy, a just-in-time compiler. CPython is also available, which translates a Python script into C and makes direct C-level API calls into the Python interpreter. An important goal of Python's developers is keeping it fun to use. This is reflected in the language's name a tribute to the British comedy group Monty Python and in occasionally playful approaches to tutorials and reference materials, such as examples that refer to spam and eggs (from a famous Monty Python sketch) instead of the standard for and bar.

A common neologism in the Python community is pythonic, which can have a wide range of meanings related to program style. To say that code is pythonic is to say that it uses Python idioms well, that it is natural or shows fluency in the language, that it conforms with Python's minimalist philosophy and emphasis on readability. In contrast, code that is difficult to understand or reads like a rough transcription from another programming language is called unpythonic. Users and admirers of Python, especially those considered knowledgeable or experienced, are often referred to as Pythonists, Pythonistas, and Pythoneers. 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.

Python’s initial development was spearheaded by Guido van Rossum in the late 1980s. Today, it is developed by the Python Software Foundation. Because Python is a multiparadigm language, Python programmers can accomplish their tasks using different styles of programming: object oriented, imperative, functional or reflective. Python can be used in Web development, numeric programming, game development, serial port access and more.

There are two attributes that make development time in Python faster than in other programming languages:

1. Python is an interpreted language, which precludes the need to compile code before executing a program because Python does the compilation in the background. Because Python is a high-level programming language, it abstracts many sophisticated details from the programming code. Python focuses so much on this abstraction that its code can be understood by most novice programmers.
2. Python code tends to be shorter than comparable codes. Although Python offers fast development times, it lags slightly in terms of execution time. Compared to fully compiling languages like C and C++, Python programs execute slower. Of course, with the processing speeds of computers these days, the speed differences are usually only observed in benchmarking tests, not in real-world operations. In most cases, Python is already included in Linux distributions and Mac OS X machines.

**5.2 BACK END: MY SQL**

MySQL is the world's most used open source [relational database management system](http://en.wikipedia.org/wiki/Relational_database_management_system) (RDBMS) as of 2008 that run as a server providing multi-user access to a number of databases. The MySQL development project has made its [source code](http://en.wikipedia.org/wiki/Source_code) available under the terms of the [GNU General Public License](http://en.wikipedia.org/wiki/GNU_General_Public_License), as well as under a variety of [proprietary](http://en.wikipedia.org/wiki/Proprietary_software) agreements. MySQL was owned and sponsored by a single [for-profit](http://en.wikipedia.org/wiki/Business) firm, the [Swedish](http://en.wikipedia.org/wiki/Sweden) company [MySQL AB](http://en.wikipedia.org/wiki/MySQL_AB), now owned by [Oracle Corporation](http://en.wikipedia.org/wiki/Oracle_Corporation).

MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack—LAMP is an acronym for "[Linux](http://en.wikipedia.org/wiki/Linux), [Apache](http://en.wikipedia.org/wiki/Apache_HTTP_Server), MySQL, [Perl](http://en.wikipedia.org/wiki/Perl)/[PHP](http://en.wikipedia.org/wiki/PHP)/[Python](http://en.wikipedia.org/wiki/Python_%28programming_language%29)." [Free-software](http://en.wikipedia.org/wiki/Free_software)-open source projects that require a full-featured database management system often use MySQL.For commercial use, several paid editions are available, and offer additional functionality. Applications which use MySQL databases include: [TYPO3](http://en.wikipedia.org/wiki/TYPO3), [Joomla](http://en.wikipedia.org/wiki/Joomla), [Word Press](http://en.wikipedia.org/wiki/WordPress), [phpBB](http://en.wikipedia.org/wiki/PhpBB), [MyBB](http://en.wikipedia.org/wiki/MyBB), [Drupal](http://en.wikipedia.org/wiki/Drupal) and other software built on the [LAMP](http://en.wikipedia.org/wiki/LAMP_%28software_bundle%29) software stack. MySQL is also used in many high-profile, large-scale [World Wide Web](http://en.wikipedia.org/wiki/World_Wide_Web) products, including Wikipedia, Google(though not for searches), [Imagebook](http://en.wikipedia.org/wiki/Facebook)[Twitter](http://en.wikipedia.org/wiki/Twitter), [Flickr](http://en.wikipedia.org/wiki/Flickr), [Nokia.com](http://en.wikipedia.org/wiki/Nokia), and [YouTube](http://en.wikipedia.org/wiki/YouTube).

**Inter images**

MySQL is primarily an RDBMS and ships with no GUI tools to administer MySQL databases or manage data contained within the databases. Users may use the included command line tools, or use MySQL "front-ends", desktop software and web applications that create and manage MySQL databases, build database structures, back up data, inspect status, and work with data records. The official set of MySQL front-end tools, MySQL Workbench is actively developed by Oracle, and is freely available for use.

**Graphical**

The official MySQL Workbench is a free integrated environment developed by MySQL AB, that enables users to graphically administer MySQL databases and visually design database structures. MySQL Workbench replaces the previous package of software, [MySQL GUI Tools](http://en.wikipedia.org/wiki/MySQL_GUI_Tools). Similar to other third-party packages, but still considered the authoritative MySQL frontend, MySQL Workbench lets users manage database design & modeling, SQL development (replacing MySQL Query Browser) and Database administration (replacing MySQL Administrator).MySQL Workbench is available in two editions, the regular free and open source Community Edition which may be downloaded from the MySQL website, and the proprietary Standard Edition which extends and improves the feature set of the Community Edition.

**CHAPTER 6**

**6. SYSTEM IMPLEMENTATION**

**6.1 MODULES**

* Datasets Acquisition
* Preprocessing
* Features Selection
* Classification
* Disease diagnosis

**6.2 MODULES DESCRIPTION**

**6.2.1 DATASETS ACQUISITION**

A data set (or dataset, although this spelling is not present in many contemporary dictionaries like Merriam-Webster) is a collection of [data](https://en.wikipedia.org/wiki/Data). Most commonly a data set corresponds to the contents of a single [database table](https://en.wikipedia.org/wiki/Table_(database)), or a single statistical [data matrix](https://en.wikipedia.org/wiki/Data_matrix_(multivariate_statistics)), where every [column](https://en.wikipedia.org/wiki/Column_(database)) of the table represents a particular variable, and each [row](https://en.wikipedia.org/wiki/Row_(database)) corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. The module conducted using the CKD dataset. There are 400 rows and 24 columns in this dataset. The output column “class” has a value of either “1” or “0.” The value “0” indicates that the patient is not a CKD patient, while the value “1” shows that the patient is a CKD patient.

**6.2.2 PREPROCESSING**

Data pre-processing is an important step in the [data mining] process. The phrase ["garbage in, garbage out"](https://en.wikipedia.org/wiki/GIGO) is particularly applicable to data mining and [machine learning](https://en.wikipedia.org/wiki/Machine_learning) projects. Data-gathering methods are often loosely controlled, resulting in [out-of-range](https://en.wikipedia.org/w/index.php?title=Range_error&action=edit&redlink=1) values, impossible data combinations, [missing values](https://en.wikipedia.org/wiki/Missing_values), etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and [quality of data](https://en.wikipedia.org/wiki/Data_quality) is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then [knowledge discovery](https://en.wikipedia.org/wiki/Knowledge_discovery) during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. In this module, we can eliminate the irrelevant values and also estimate the missing values of data. Finally provide structured datasets.

**6.2.3 FEATURES SELECTION**

Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. A related term, feature engineering (or feature extraction), refers to the process of extracting useful information or features from existing data. Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often uni-variate and consider the feature independently, or with regard to the dependent variable. It can be used to construct the multiple heart diseases. In this module, select the multiple features from uploaded datasets. And train the datasets with CKD or NON-CKD and generate the model file for future classification

**6.2.4 CLASSIFICATION**

In this module implement classification algorithm to predict the heart diseases. And using deep learning algorithm such as Multi-layer perceptron algorithm to predict the diseases. A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. It (MLP) consists of multiple layers of nodes in a directed graph, and each layer is fully connected to the next one. Each node is a neuron with a nonlinear activation function except for the input nodes. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modified form of the standard linear perceptron and can distinguish data that are not linearly separable. If a multilayer perceptron (MLP) has a simple on-off mechanism i.e. linear activation function in all neurons ,to determine whether or not a neuron fires, then it is easily proved with linear algebra that any number of layers can be reduced to the standard two-layer input-output model. The gradient techniques are then applied to the optimization methods to adjust the weights to minimize the loss function in the network. Hence, the algorithm requires a known and a desired output for all inputs in order to compute the gradient of loss function. Usually, the generalization of MultiLayerd Feed Forward Networks is done using delta rule which possibly makes a chain of iterative rules to compute gradients for each layer. Back Propagation Algorithm necessitates the activation function to be different between the neurons. The ongoing researches on parallel, distributed computing and computational neuroscience are currently implemented with the concepts of MultiLayer Perceptron using a Back Propagation Algorithm. MLP Back Propagation Algorithm has also gained focus in pattern recognition domain. They are so convenient in research, because of their ability in solving complex problems, and also for their fitness approximation results even with critical predictions. MLP is one of the Neural Network models, has the same architecture of Feed-Forward back Propagation for Supervised training. The multilayer perceptron is the most known and most frequently used type of neural network. User can provide the features and automatically predict the diseases.

**6.2.5 DISEASE DIAGNOSIS**

Medical decision support system is a decision-support program which is designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patients’ data. In this module, provide the diagnosis information based on predicted diseases. Proposed system provides improved accuracy in disease prediction. Risk factors are conditions or habits that make a person more likely to develop a disease.

**6.3 DETAILED DESIGN**

A back propagation is a feed forward artificial neural network structure that plots sets of input data onto a set of appropriate outputs. It contains of numerous layers of nodes in a directed graph, and each layer is fully connected to the next one. Each node is a neuron with a nonlinear activation function excluding for the input nodes. Back propagation exploits a supervised learning technique called back propagation for training the network. If back propagation has a simple on-off mechanism i.e. linear activation function in all neurons, to regulate whether or not a neuron fires, then it is easily proved with linear algebra that any number of layers can be reduced to the standard two-layer input-output model. The gradient techniques are then practical to the optimization methods to regulate the weights to diminish the loss function in the network. Feed Forward Networks is done using delta rule which possibly makes a chain of iterative rules to compute gradients for each layer. Back Propagation Algorithm requires the activation function to be different between the neurons. The ongoing investigates on parallel, distributed computing and computational neuroscience are currently implemented with the concepts of Back Propagation They are so convenient in study, because of their ability in solving complex problems, and also for their fitness approximation results even with critical predictions. Back propagation is one of the Neural Network models, has the same architecture of Feed-Forward back Propagation for Supervised training. The back propagation is the most known and most frequently used type of neural network. User can provide the features and inevitably predict the diseases. The algorithm steps are follows

Step 1: Randomly set the weights and biases.

Step 2: Feed the training sample.

Step 3: Propagate the inputs forward; compute the net input and output of each unit in the hidden and output layers.

Step 4: Back propagate the error to the intermediate layer.

Step 5: Update weights and biases to replicate the propagated errors.

Training and learning functions are mathematical measures used to automatically regulate the network's weights and biases.

Step 6: Stop condition

**CHAPTER 7**

**7. CONCLUSION AND FUTURE ENHANCEMENT**

**7.1 CONCLUSION**

Chronic Kidney Disease, also known as chronic kidney failure, is a gradual loss of kidney function. The kidneys filter waste and fluids from our blood which are excreted through urine. The kidneys balance the salts and minerals such as calcium, phosphorus, sodium, and potassium that circulate in the blood. They also make hormones that help control blood pressure, make red blood cells and keeps the bones strong. When chronic kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in our body. The early stages of chronic kidney disease is characterised by few signs or symptoms, which makes it difficult to identify the disease. The symptoms may become apparent only when the kidney function is significantly impaired. If left untreated, Chronic kidney disease can progress to end stage kidney failure, which is fatal without artificial filtering (dialysis) or a kidney transplant. In the work presented, neural networks is used for diagnosis of chronic kidney disease. With the above results we have achieved our objective to find the best model for CKD diagnosis. The multilayer perceptron with back propagation algorithm is a good model for diagnosis of CKD, Its accuracy is 91%. The error rate is also considerably low. Thus we have come to conclusion that multilayer perceptron trained with back propagation is one of the most suitable and efficient algorithm for kidney disease diagnosis. Limitation of the application is we do not have data of strength because of size of data set and missing attribute values

**7.2 FUTURE ENHANCEMENT**

In future, we can extend the framework to implement various deep learning algorithm to improve the accuracy in disease prediction and also analyse multiple types of datasets such as image and so on.

**CHAPTER 8**

**8. APPENDIX**

**8.1 SOURCE CODE**

**MAIN.PY**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import pickle

import warnings

warnings.filterwarnings('ignore')

df = pd.read\_csv("./Data/kidney\_disease.csv")

df.head()

df.shape

df.info()

df.drop('id', axis = 1, inplace = True)

df[['htn','dm','cad','pe','ane']] = df[['htn','dm','cad','pe','ane']].replace(to\_replace={'yes':1,'no':0})

df[['rbc','pc']] = df[['rbc','pc']].replace(to\_replace={'abnormal':1,'normal':0})

df[['pcc','ba']] = df[['pcc','ba']].replace(to\_replace={'present':1,'notpresent':0})

df[['appet']] = df[['appet']].replace(to\_replace={'good':1,'poor':0,'no':np.nan})

df['classification'] = df['classification'].replace(to\_replace={'ckd':1.0,'ckd\t':1.0,'notckd':0.0,'no':0.0})

df['wc']=df['wc'].replace(["\t6200","\t8400","\t?"],[6200,8400, np.nan])

df['pcv']=df['pcv'].replace(["\t43","\t?"],[43,np.nan])

df['rc']=df['rc'].replace(["\t?"],[np.nan])

df = df.fillna(method='ffill')

df = df.fillna(method='backfill')

# Further cleaning

df['pe'] = df['pe'].replace(to\_replace='good',value=0)

df['appet'] = df['appet'].replace(to\_replace='no',value=0)

df['cad'] = df['cad'].replace(to\_replace='\tno',value=0)

df['dm'] = df['dm'].replace(to\_replace={'\tno':0,'\tyes':1,' yes':1, '':np.nan})

df['classification'].value\_counts()

target\_true\_count = len(df.loc[df['classification'] == 1])

target\_false\_count = len(df.loc[df['classification'] == 0])

#target\_true\_count, target\_false\_count

sns.countplot(x = 'classification',data = df)

from sklearn.model\_selection import train\_test\_split

feature\_columns = ['age', 'bp', 'sg', 'al', 'su', 'rbc', 'pc', 'pcc', 'ba', 'bgr', 'bu','sc', 'sod',

'pot', 'hemo', 'pcv', 'wc', 'rc', 'htn', 'dm', 'cad','appet', 'pe', 'ane']

predicted\_class = ['classification']

X = df[feature\_columns]

y = df[predicted\_class]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.30, random\_state=10)

from sklearn.neural\_network import MLPClassifier

from sklearn.metrics import classification\_report

classifier = MLPClassifier(random\_state=3)

classifier.fit(X\_train, y\_train)

y\_pred = classifier.predict(X\_test)

print(classification\_report(y\_test, y\_pred))

filename = 'Model/prediction-rfc-model.pkl'

pickle.dump(classifier, open(filename, 'wb'))

**APP.PY**

from flask import Flask, render\_template, flash, request, session,send\_file

from flask import render\_template, redirect, url\_for, request

import sys

import pickle

import numpy as np

app = Flask(\_\_name\_\_)

app.config['DEBUG']

app.config['SECRET\_KEY'] = '7d441f27d441f27567d441f2b6176a'

@app.route("/")

def homepage():

return render\_template('home.html')

@app.route("/result", methods=['GET', 'POST'])

def result():

if request.method == 'POST':

t1 = request.form['t1']

t2 = request.form['t2']

t22 = request.form['t22']

t23 = request.form['t23']

t24 = request.form['t24']

t25 = request.form['t25']

t26 = request.form['t26']

t27 = request.form['t27']

t28 = request.form['t28']

t29 = request.form['t29']

t210 = request.form['t210']

t211 = request.form['t211']

t212 = request.form['t213']

t213 = request.form['t213']

t214 = request.form['t214']

t215 = request.form['t215']

t216 = request.form['t216']

t217 = request.form['t217']

t218 = request.form['t218']

t219 = request.form['t219']

t220 = request.form['t220']

t221 = request.form['t221']

t222 = request.form['t222']

t223 = request.form['t223']

filename = 'Model/prediction-rfc-model.pkl'

classifier = pickle.load(open(filename, 'rb'))

data = np.array([[t1,t2, t22, t23, t24,t25, t26,t27,t28,t29,t210,

t211,

t212,

t213,

t214,

t215,

t216,

t217,

t218,

t219,

t220,

t221,t222,t223]])

my\_prediction = classifier.predict(data)

print(my\_prediction[0])

Answer = ''

tre = ''

if my\_prediction == 1:

Answer = 'CKD'

tre = 'Vitamin, Calcium reducer, Bone marrow stimulant, Diuretic and Dietary supplement'

elif my\_prediction == 0:

Answer = 'NON-CKD'

tre = ''

return render\_template('home.html', res=Answer,tre=tre)

if \_\_name\_\_ == '\_\_main\_\_':

app.run(debug=True, use\_reloader=True)

**8.2 SCREENSHOTS**

**TRAIN THE DATA**

RangeIndex: 400 entries, 0 to 399

Data columns (total 26 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 id 400 non-null int64

1 age 391 non-null float64

2 bp 388 non-null float64

3 sg 353 non-null float64

4 al 354 non-null float64

5 su 351 non-null float64

6 rbc 248 non-null object

7 pc 335 non-null object

8 pcc 396 non-null object

9 ba 396 non-null object

10 bgr 356 non-null float64

11 bu 381 non-null float64

12 sc 383 non-null float64

13 sod 313 non-null float64

14 pot 312 non-null float64

15 hemo 348 non-null float64

16 pcv 330 non-null object

17 wc 295 non-null object

18 rc 270 non-null object

19 htn 398 non-null object

20 dm 398 non-null object

21 cad 398 non-null object

22 appet 399 non-null object

23 pe 399 non-null object

24 ane 399 non-null object

25 classification 400 non-null object

dtypes: float64(11), int64(1), object(14)

memory usage: 81.4+ KB

precision recall f1-score support

0.0 0.88 0.81 0.85 37

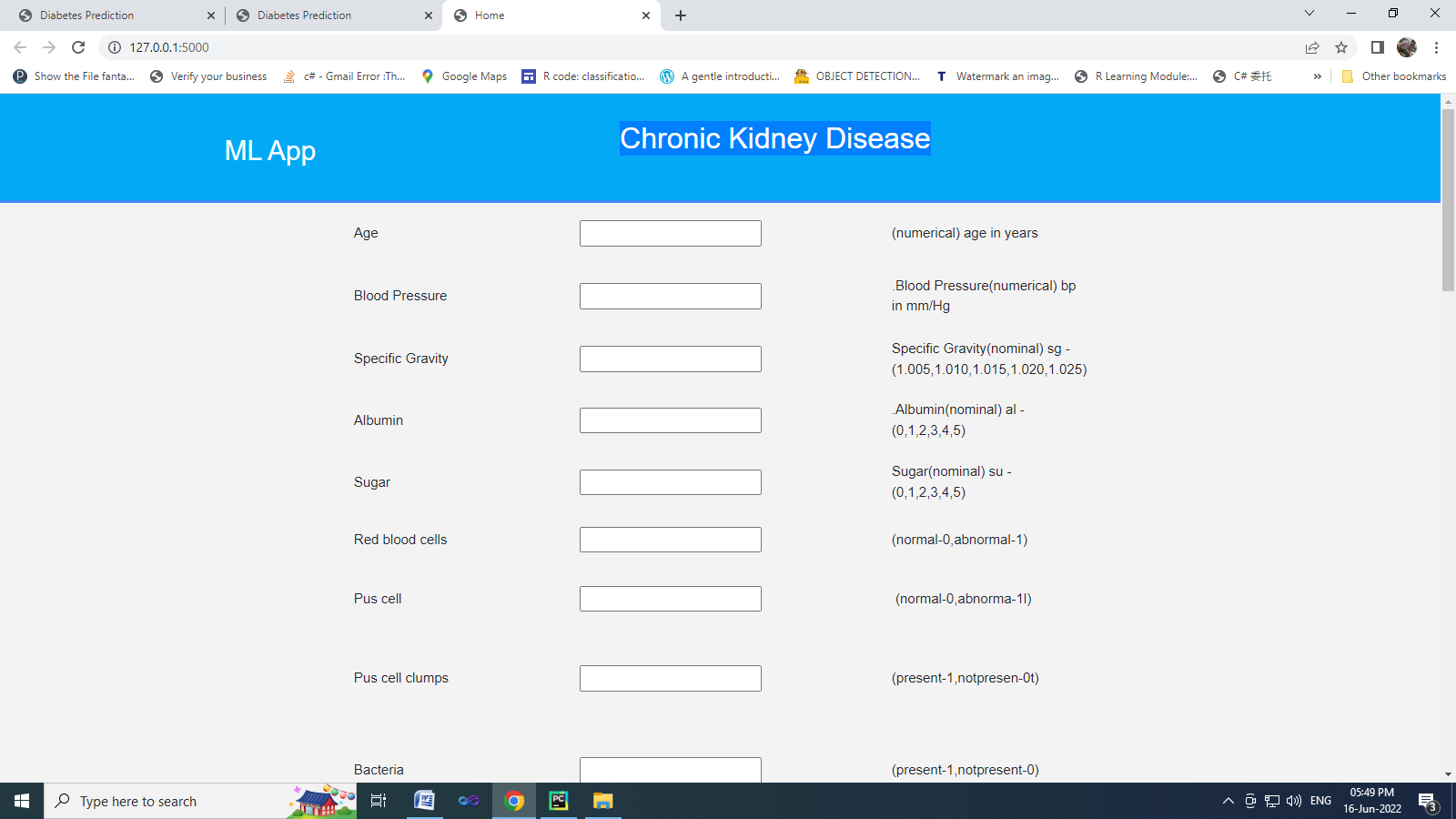
1.0 0.92 0.95 0.93 83

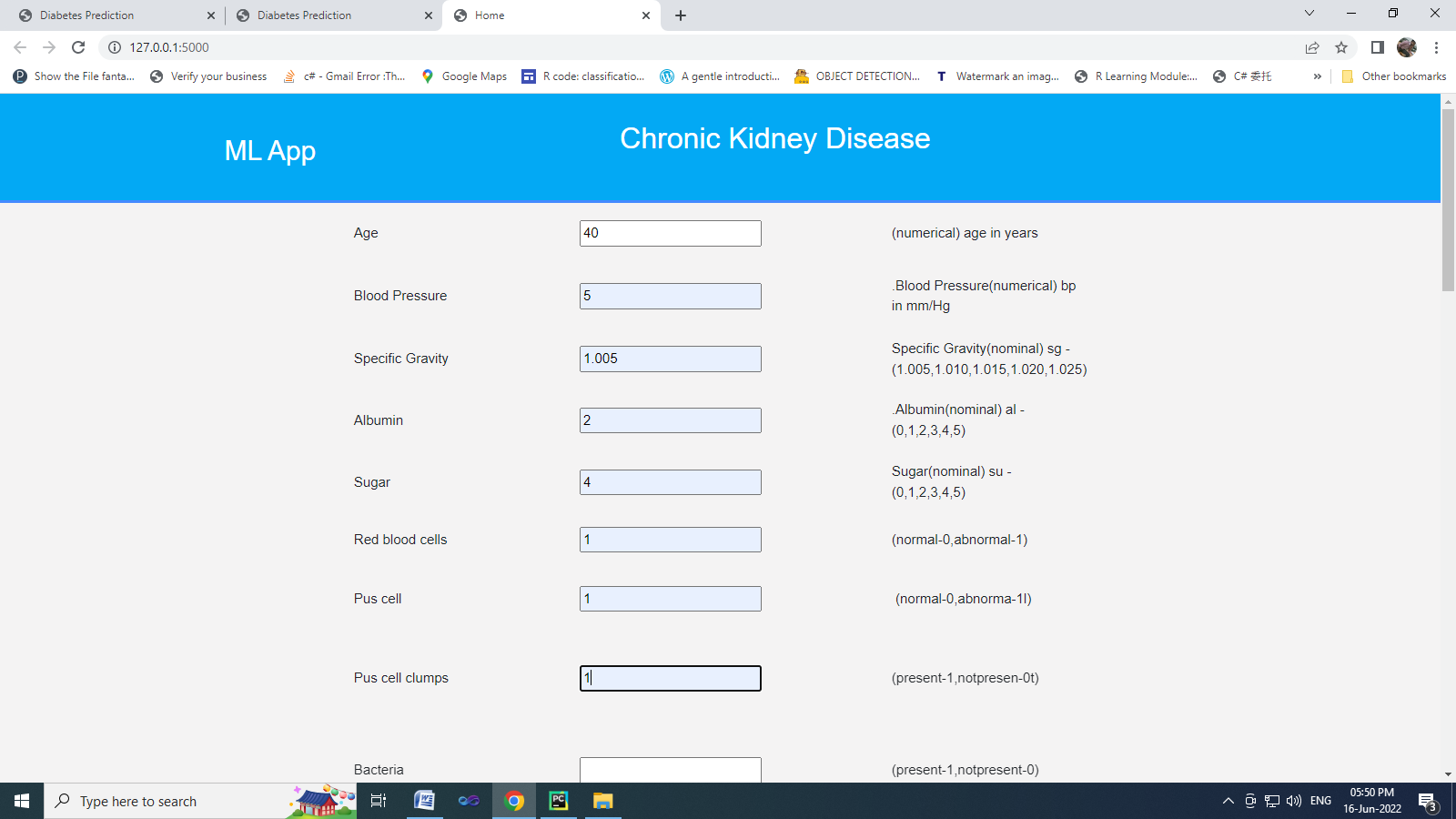
accuracy 0.91 120

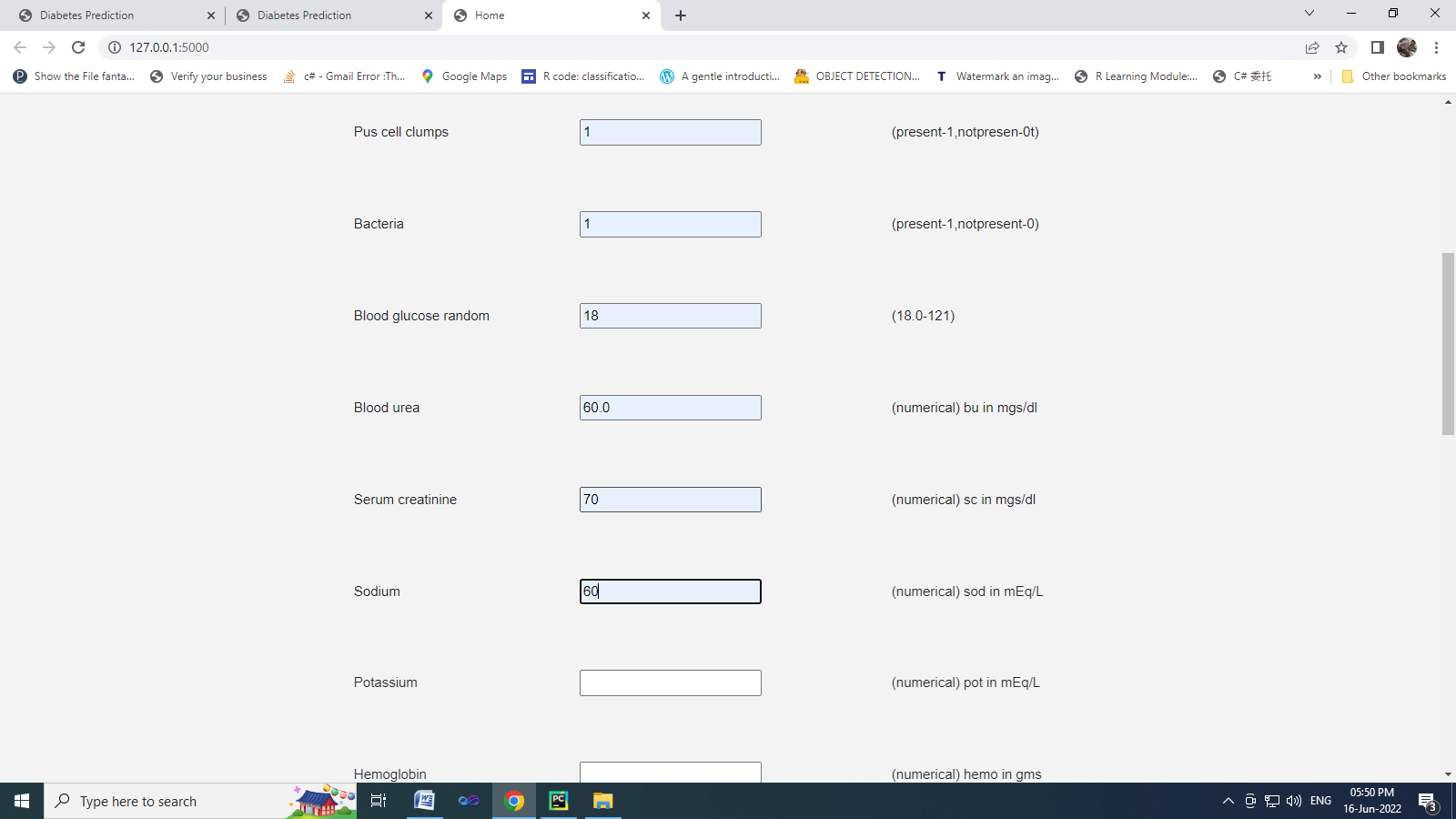
macro avg 0.90 0.88 0.89 120

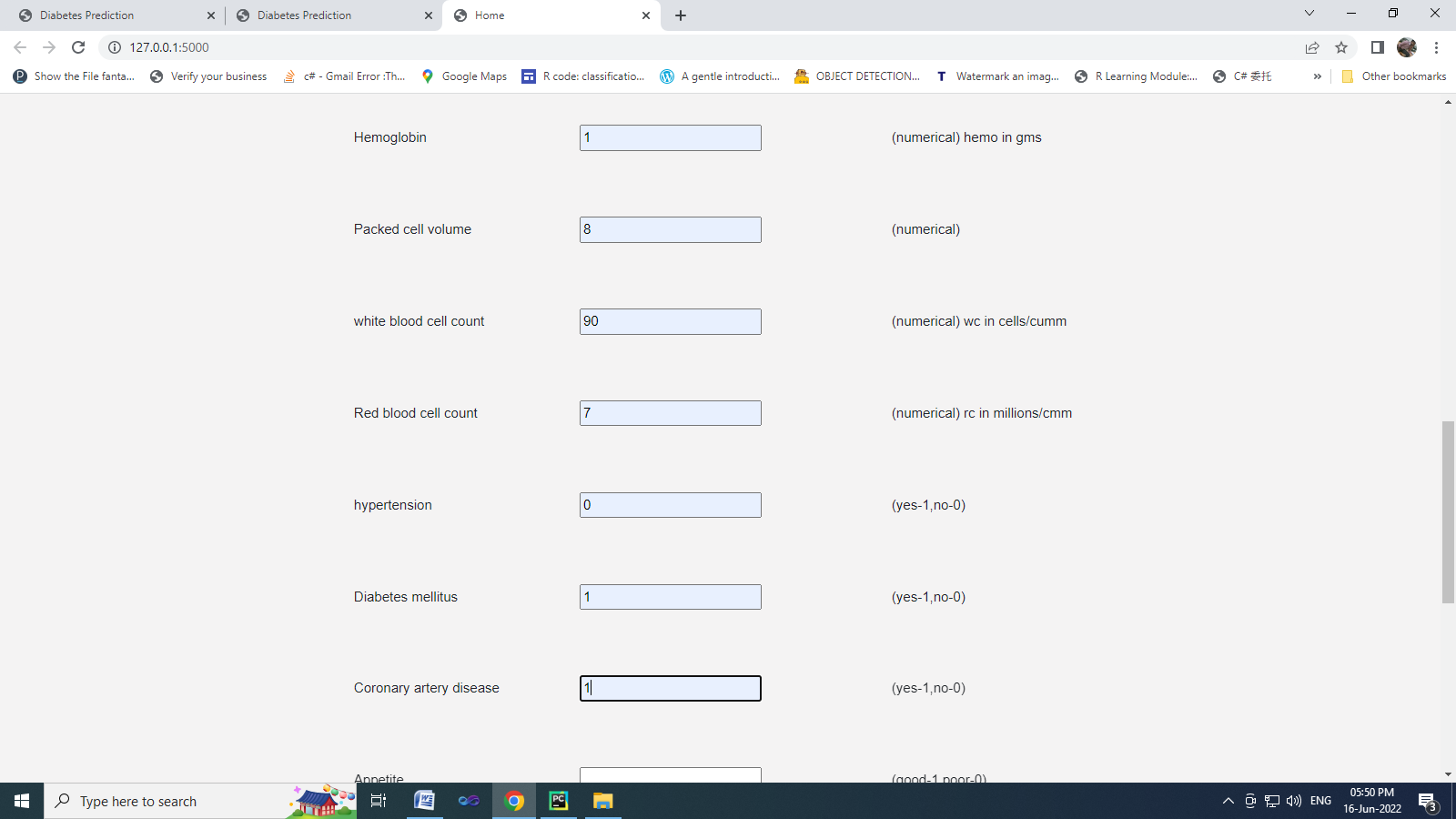
weighted avg 0.91 0.91 0.91 120

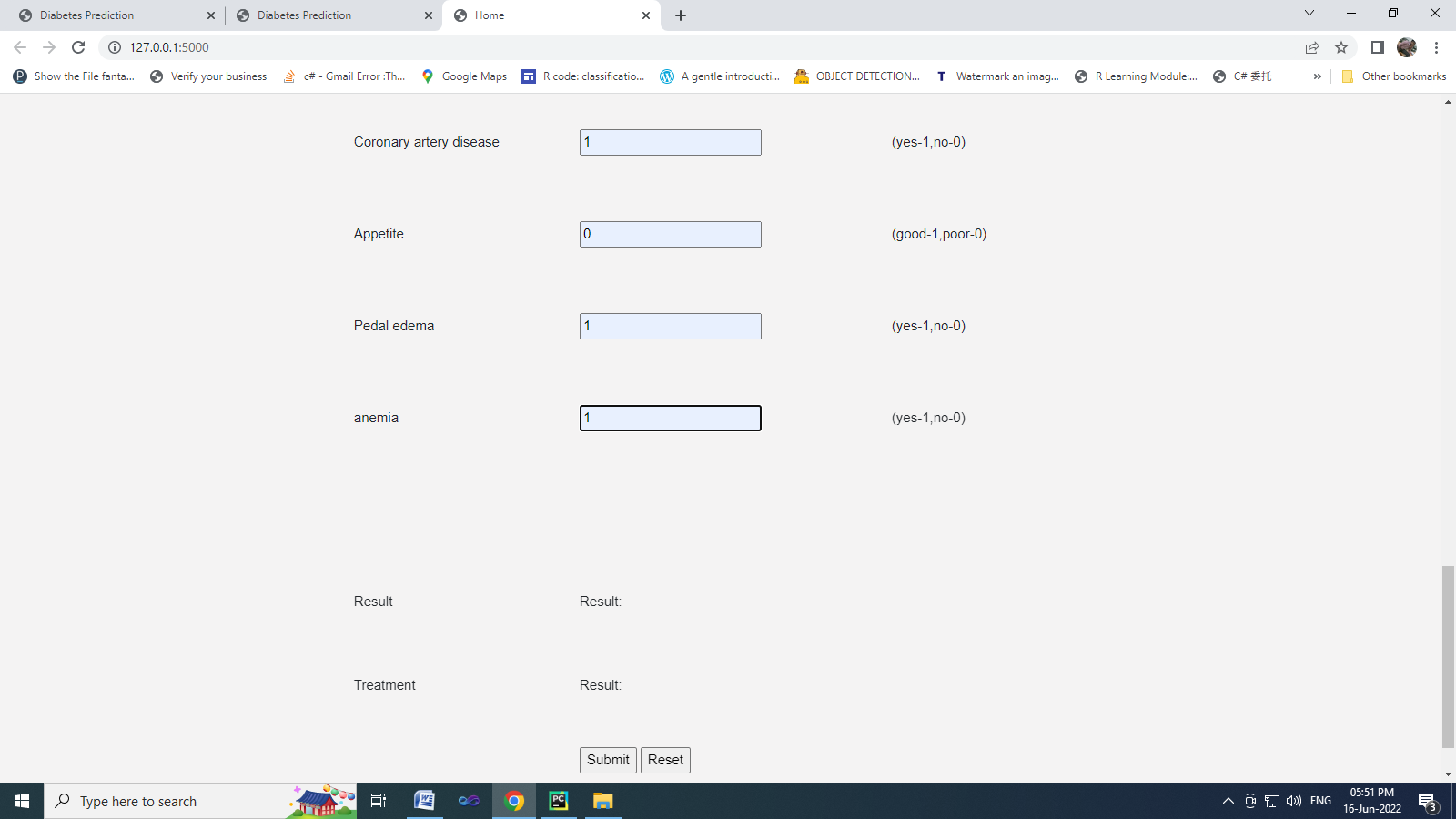
**TEST THE DATA**

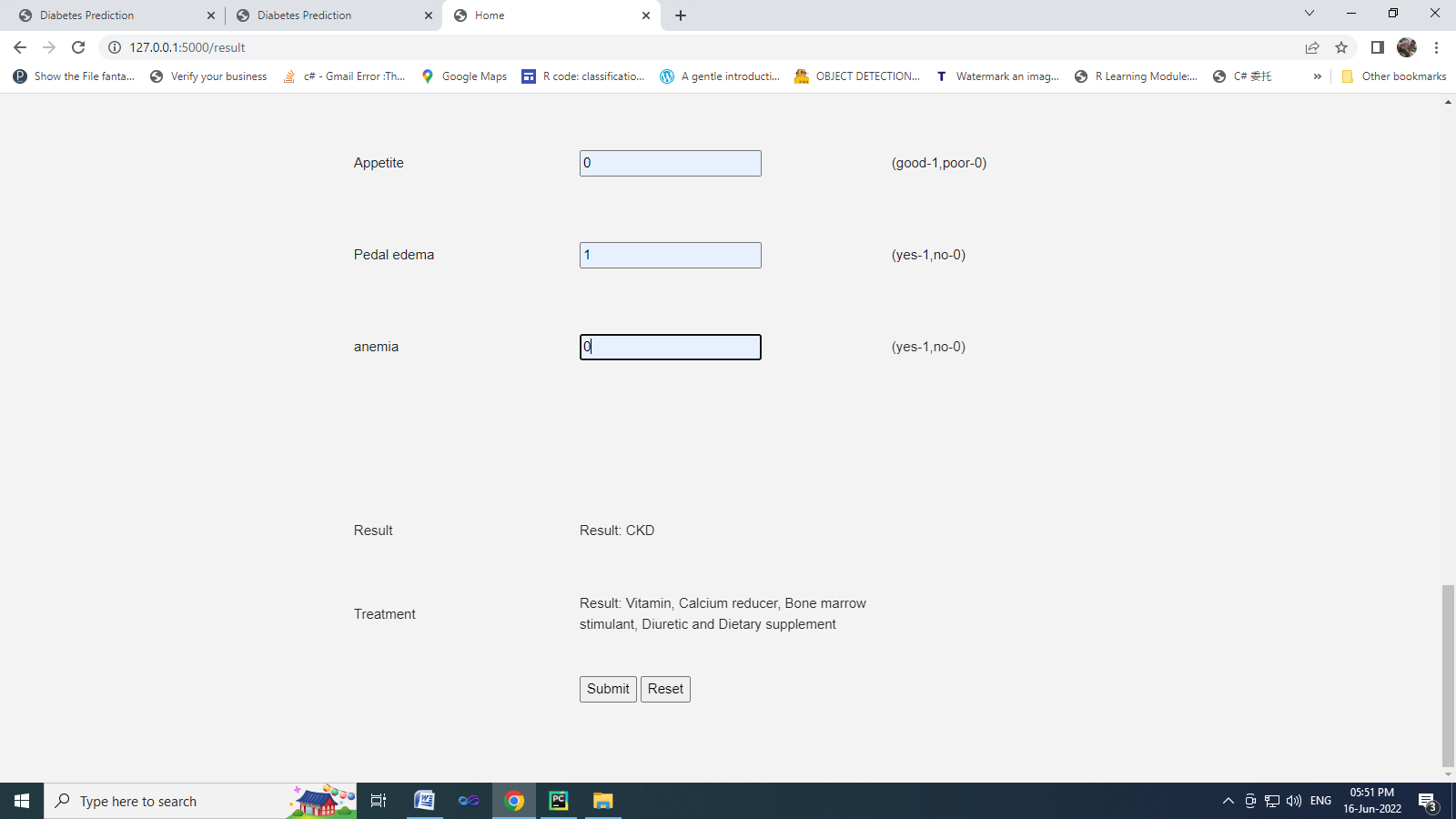












**CHAPTER 9**

**9. REFERENCES**

[1] Sinha, Parul, and Poonam Sinha. "Comparative study of chronic kidney disease prediction using KNN and SVM." International Journal of Engineering Research and Technology 4.12 (2015): 608-12.

[2] Revathy, S., et al. "Chronic kidney disease prediction using machine learning models." International Journal of Engineering and Advanced Technology 9.1 (2019): 6364-6367.

[3] Scholar, P. G. "Chronic kidney disease prediction using machine learning." International Journal of Computer Science and Information Security (IJCSIS) 16.4 (2018).

[4] Kriplani, Himanshu"Prediction of chronic kidney diseases using deep artificial neural network technique." Computer aided intervention and diagnostics in clinical and medical images. Springer, Cham, 2019. 179-187.

[5] Alassaf, Reem A., et al. "Preemptive diagnosis of chronic kidney disease using machine learning techniques." 2018 international conference on innovations in information technology (IIT). IEEE, 2018.

[6] Almasoud, Marwa, and Tomas E. Ward. "Detection of chronic kidney disease using machine learning algorithms with least number of predictors." International Journal of Soft Computing and Its Applications 10.8 (2019).

[7] Ekanayake, Imesh Udara, and Damayanthi Herath. "Chronic kidney disease prediction using machine learning methods." 2020 Moratuwa Engineering Research Conference (MERCon). IEEE, 2020.

[8] Gudeti, Bhavya, et al. "A novel approach to predict chronic kidney disease using machine learning algorithms." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020.

[9] Padmanaban, K. Anantha, and G. Parthiban. "Applying machine learning techniques for predicting the risk of chronic kidney disease." Indian Journal of Science and Technology 9.29 (2016): 1-6.

[10] Qin, Jiongming, et al. "A machine learning methodology for diagnosing chronic kidney disease." IEEE Access 8 (2019): 20991-21002.