A

Major Project

On

CITY HEALTH PREDICTION MODEL USING RANDOM FOREST CLASSIFICATION METHOD

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

Ch Pooja (17601A0524)

D Snehith Kumar (17601A0530)

G Abhiram (17601A0536)

Under the Guidance of

M. ANUSHA REDDY

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi)
Recognized Under Section 2(f) & 12(B) of the UGCAct.1956,
Kandlakoya (V), Medchal Road, Hyderabad-501401.

2017-21

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "**CITY HEALTH PREDICTION MODEL USING RANDOM FOREST CLASSIFICATION METHOD**" being submitted by **CH POOJA(17601A0524), D SNEHITH KUMAR(17601A0530) & G ABHIRAM (17601A0536)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2020-21.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Ms. M. Anusha Reddy
Assistant Professor
INTERNAL GUIDE

Dr. A. Raji Reddy
DIRECTOR

Dr. K. Srujan Raju
HoD

EXTERNAL EXAMINER

Submitted for viva voice Examination held on ______________

ACKNOWLEGDEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide

Ms. M. Anusha Reddy, Assistant Professor for her exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by her shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) Coordinators: **Mr. J. Narasimha Rao, Mr. B. P. Deepak Kumar, Mr. K. Murali, Dr. Suwarna Gothane and Mr. B. Ramji** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju,** Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We would like to express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

CH POOJA (17601A0524)
D SNEHITH KUMAR (17601A0530)
G ABHIRAM (17601A0536)

ABSTRACT

City Health Office in Indonesia is creating a health report every year, describing the condition of the city public health. The report is used as the source of determining the city health index. The construction of a city health development index is important to produce an objective formula. In this study, the classification method Random Forest is used to developing a proper model for prediction and analysis of the health index of a city. The goal of this work is to find a prediction model to make a more accurate prediction and reducing errors in dealing with the city health index. The performance of the model is evaluated by using three parameters: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The research shows that the model of Random Forest with a 15 percent data test by using 200 decision trees gives the best results.

LIST OF FIGURES

LIST OF SCREENSHOTS

TABLE OF CONTENTS

1. INTRODUCTION

1. INTRODUCTION

Most of the American citizens die every year because of errors present in the health care system, and thousands of people deteriorate from nonfatal burns owing to the constant cause. Health Information Technology (IT) Framework recommended few strategies, like collaboration, knowing consumer selection of clinicians and organizations, and IT adoption.

Healthcare is one of the fastest growing sectors today and is currently in the core of a complete global overhaul and transformation. Russell Reynolds and Associates cites that global healthcare costs, currently estimated at $6 trillion to $7 trillion, are projected to reach more than $12 trillion within just seven years. This trend is also exemplified domestically, in the United States. The total spending on healthcare in the United States increased up to 5.3 percent and has topped $3 trillion nationwide. Additionally, healthcare spending in the United States represents 17 percent of the total gross domestic product (GDP); our healthcare costs are rising at rates close to double of our economic growth rate. In addition to a rise in the amount consumers are spending on healthcare, the federal government has been forced to pay more and more for healthcare as costs become too high for patients to afford. The amount of money the federal government has allocated for healthcare spending has increased by 11.7 percent in 2014 to an incredible $844 billion in 2015. This rise in federal funding represents the significant disparity between the cost of healthcare and the financial burden on consumers. Given this rapid growth in costs, a number of actions must be taken to ensure the costs of healthcare do not further spiral out of control. The need for patient-physician communication, follow-up appointments, and the availability of specialists have also become painfully apparent. Innovation and technological solutions may be the solution to fix the issues with our modern-day healthcare system. These innovations range from swallowable microchips that alert doctors when medication has been taken to large scale data analysis to determine which medications are most effective. However, recently, machine learning has been identified as having major technological application in the healthcare realm. While such technologies will probably never completely replace physicians, they can transform the healthcare sector, benefiting both patients and providers.

The field of medicine has taken significant strides in its advancement; the development of vaccinations, antibiotics, and even the concept of sterilization have disrupted the industry and caused a cascade-like effect on all patients and doctors involved in healthcare. Needless to say, the human population has progressed a great degree from past medical care. The healthcare industry is made up of preventive, diagnostic, remedial, and therapeutic sectors. Each of these sectors work together to provide a comprehensive, holistic experience for the modern-day patients.

Some major trends are occurring in the industry today; the first of which involves the transition to patientcentered care. Organizational changes have caused a transition from being focused on hospital care to being more reliant on preventative and outpatient centric care. As medicine reaches the apex of its transition to consumerism, there will be an inherent need to provide consumers with the tools to make intelligent decisions [8]. A fundamental aspect of patient orientated care is regarding the strength of the relationship between the physicians and their patients. James Rikert provides the following example, “(A patient’s) primary malignancy was lung cancer. During the course of treatment… (she had seen) … a pulmonologist for her symptoms. He had performed pulmonary function tests, prescribed inhalers, and told her to return if her symptoms did not improve. She never went back, and the cancer was later found by her family doctor, by which time it was metastatic. A fundamental issue in the example above and in healthcare system in the past has been that of not developing a good relationship between the health provider and the patient. The way that a modern-day physician interacts with their patients determines their satisfaction with the care, their response to medication, and their tendency to schedule follow up appointments. In the status quo, doctors have too many patients to ensure that each patient is in good health, even after the appointment. Additionally, doctors are being urged to see a greater number of patients in a smaller amount of time. As the need to establish relationships with patients increases in tandem with the rising number of patients and need for care, doctors will have to utilize new technology to accomplish their goal of patient-centered care.

The hazardous improvement of health-related information given new opportunities for developing recuperate of a patient, Machine learning shows a vital performance in health-care and these are mostly enforced to healthcare, which includes computer-aided diagnosis, image registration, image annotation, image-guided medical aid, and image database retrieval, multimodal image fusion, medical image segmentation, where deficiency might be incurable.

Machine learning has probably limited social impacts in the health-care field [8]. Machine learning provides the solution for decreasing the increasing price of health-care and serving to create an improved patient-clinician communication. ML solutions will be used for an inordinateness of health-relevant uses; some include serving to clinicians identify a lot of customized prescriptions and therapy for patients and additionally serving to patients identify once and if they must record follow up appointments.

Currently, in health-care, a huge quantity of information has become accessible. It contains EMRs that consist of information which may be either unstructured or structured. Structured health information is the data that’s simple to analyze in a database and they will carry a set of statistics and classes as well as however not restricted to patient weights, and even generic symptoms like stomach pain, headache, etc. The bulk of medical knowledge is unstructured information within the variety of numerous completely different notes, images, audio and video recording, reports, and discharge summaries. It’s terribly exhausting to quantify and analyze a conversation between the supplier and the patient; the conversation is incredibly personalized and might take many alternative directions. The algorithms of Machine learning are useful in identifying complicated patterns within prosperous and huge data. This facility is especially well-suited to clinical applications, particularly those people who rely on advanced genomics and proteomics measurements. It is often used utilized in numerous illness diagnosing and detection. In medical applications, machine learning algorithms will manufacture higher decisions regarding treatment plans for patients by suggestions of implementing useful health-care system.

Healthcare management is utilizing this method to forecast wait times for patients in exigency department waiting for places. These models use factors like patient information, discomfort levels, exigency department charts, and even the layout of the hospital room itself to conclude wait times.

Using the prognostic model, clinics will think hospital room admissions. So machine learning application could profit patients by decreasing price, rising accuracy, or diffusing experience that is in brief offer.

Literature Survey:

Big data and analytics have been causing disruptions in major industry segments. In recent years, big data has become a new ubiquitous term. Big Data refers to large, complex datasets that are beyond the capabilities of traditional data management systems of storing, managing, and processing in a timely and economical manner. Big data technologies can handle structured, semi- structured, and unstructured data in petabytes and more. Healthcare is one of the verticals that can significantly benefit from the increasing amounts of data and its availability. Entities- including health care providers, pharmaceutical companies, research institutions, and government agencies-have begun to compile massive amounts of data from research, clinical trials, and public health and insurance programs. The consolidation of data from various sources has significant potential. In the past, doctors have been treating patients based on symptoms; however, physicians are beginning to diagnose and treat patients with a concept known as evidence-based medicine. This involves reviewing large amounts of data aggregated from clinical trials and other treatment pathways on the large scale and making decisions based on the best information available. For example, if a patient comes in with a particular case of the flu, a physician in the past would rely on what he or she knew about the flu in general or what other doctors in the area knew. With big data technologies, a physician can look at nationwide trends on what course of treatment would work best for the patient to prescribe the best medications. The aggregation of individual data sets that would otherwise prove meaningless provides doctors with the information needed to make better, more holistic medical decisions.

In today’s connected world, data across sectors are growing exponentially. As the volumes of data increase, new novel ways to interact with and to extract meaning from the data are emerging. In the past, human intervention has been used to parse through the data; however, this is inefficient and a large number of hidden patterns within the data are not found. This is externally important is the healthcare sector. As Thomas H. Davenport writes in the Wall Street Journal, “Humans can typically create one or two good models a week (while) machine learning can create thousands of models a week".

Machine Learning is a particular method of data analytics that automates model building, as it relates to the development of models. With machines learning to utilize certain algorithms, they can find hidden insights from data; it is important to note that in machine learning, we are not telling the machines where to look. The iterative nature of machine learning allows the machine to adapt its methods and outputs as it is exposed to new situations and data.

BenevolentAI: BenevolentAI was founded in 2013 by Ken Mulvany founder of Proximagen. BenevolentBio is focused on applying technology in the bioscience industries. The initial focus has been on human health – generating new ideas that have the potential to improve the lives of millions and deliver better medicines to patients faster in currently overlooked areas such as orphan disease and rare cancers. BenevolentTech is developing an advanced artificial intelligence platform that helps scientists make new discoveries and redefines how scientists gain access to, and use, all the data available to drive innovation. The technology is built upon a deep judgement system that learns and reasons from the interaction between human reasoning and data. Butterfly Network: Butterfly Network is transforming diagnostic and therapeutic imaging with devices, deep learning, and the cloud. Butterfly Network operates at the intersection of engineering and medicine by bringing together world-class talent in computer science, physics, mechanical engineering, electrical engineering and medicine.

Digital Reasoning Systems: Digital Reasoning is a global leader in using artificial intelligence to understand human communications. Its cognitive computing platform, Synthesys, automates key tasks and uncovers transformative insights across vast amounts of human communications for many of the world’s most elite companies, organizations, and agencies.

Flatiron Health: Flatiron Health is a health care technology company and operator of the OncologyCloud platform. Integrating across the entire clinical data spectrum, Flatiron Health allows cancer care providers and life science companies to gain deep business and clinical intelligence through its web-based platform.

H2O.ai: H2O is a provider of an open source based predictive analytics platform for data scientists and application developers who need scalable and fast machine learning for smart business applications. These applications include smart home appliances, self-driving cars, personalized digital content, smart assistants, and others.

iCarbonX: CarbonX is developing an artificial intelligence platform to facilitate research related to the treatment of diseases, preventive care, and precision nutrition. This approach is considered an essential element in enabling the future development of personalized medicine.

Pathway Genomics: Pathway Genomics, founded in 2008, provides physicians and their patients with accurate genetic information to improve or maintain health and wellness. The company's mobile health applications merge artificial intelligence and deep learning with personal genetic information that provides personalized health and wellness guidance.

WellTok: WellTok combines knowledge of the healthcare industry and social networking technology in its CafeWell.com channel to achieve levels of consumer engagement for healthcare population managers through Social Health Management. WellTok's software/Internet products focus on providing a complete, integrated solution that includes the engaging social health network CafeWell; actionable member and group analytics; and integration with enterprise information systems.

Another startup in this space is AliveCor - the Silicon Valley-based maker of the Kardia Mobile, a portable electrocardiogram device - is now betting that artificial intelligence will help doctors monitor patients' heart conditions. Its machine learning algorithms will automatically flag abnormal ECGs, leading to early detection of common heart arrhythmias and helping prevent strokes.

Durable population-wide improvements in health and health equity require the active engagement not only of health care and public health but of other sectors as well. Initiatives like the County Health Rankings and Roadmaps have helped advance widespread understanding of drivers responsible for better health and health equity by reporting health-related data from diverse sectors at the county level. Most US cities, however, are not contiguous with county boundaries: of the 500 largest US cities, 16% bridge more than 1 county, and most others constitute only a relatively small portion of a county’s population (∼30%, on average). County data are thus insufficient for many municipal leaders seeking to initiate health improvement initiatives in their jurisdictions. Although the health departments of some large cities have robust surveillance and analytic capabilities, most local health departments do not.

In recent times, the application of computational or machine intelligence in medical diagnostics has become quite common. Machine intelligence aided decision systems are often being adopted to assist (but not to replace) a physician in diagnosing the disease of a patient. A physician typically accumulates her knowledge based on patient symptoms and the confirmed diagnoses. Thus diagnostic accuracy is highly dependent on a physician's experience. Since it is now relatively easy to acquire and store a large amount of information digitally, the deployment of computerized medical decision support systems has become a viable approach to assisting physicians to swiftly and accurately diagnose patients. Such a system can be seen as a classification task as the goal is to make a prediction (i.e., diagnosis) on a new case based on the available records and features (of previously known cases). Such classification tasks are considered to be one of the most challenging tasks in medical informatics. While various statistical techniques may be applied in medical data classification, the major drawback of these approaches is that they depend on some assumptions (e.g., related to the properties of the relevant data) for their successful application. To know the properties of the dataset is a difficult task and is sometimes is not feasible. On the other hand, soft computing based approaches are less dependent on such knowledge. A number of soft computing based classifiers have been proposed and analyzed in the literature to classify medical data accurately. Abbass et al. proposed a system with the pareto-differential evaluation algorithm with a local search scheme, termed the Memetic ParetoArtificial Neural Network (MPANN), to diagnose breast cancer. Subsequently, Kiyan et al. presented a statistical neural network- based approach to diagnose breast cancer. In Ref., Karabatak et al. developed an expert system for detecting breast cancer, where, to reduce the dimensions of the dataset, Association Rules (AR) were used. Peng et al. proposed a hybrid feature selection approach to address the issues of high dimensionality of biomedical data, and experimented on the breast cancer dataset. Fana et al. combined case-based data clustering and a fuzzy decision tree to design a hybrid model for medical data classification. The model was executed on two datasets, WBC and liver disorders. Azar et al. proposed three classification methods, namely, radial basis function (RBF), multilayer perceptron (MLP), and probabilistic neural network (PNN), and experimented on a breast cancer dataset. In their experiments, PNN showed better performance than MLP.

We have also conducted an ablation study on each dataset to verify the importance of the features selected for the model and to confirm the positive contribution thereof on the classification task. Demonstrated these results. These findings further strengthen our claim that our general methodology of feature ranking and selection does play a strong role in a robust model construction. As a by-product, we also have a suggestion concerning the feature importance in each dataset (from classification point of view).

For example, for the Diabetes dataset, our experiments suggest that Features 3 and 4 have less contribution/importance in the context of disease prediction. This is interesting, as Feature 3 is ‘diastolic blood pressure (mm Hg)’and Feature 4 is ‘triceps skin fold thickness (mm)’. On the other hand, across all ranking algorithms, Feature 2 (‘plasma glucose concentration a 2 h in an oral glucose tolerance test’) has been ranked as the most important feature, which is in accord from a medical per- spective. As another example, for the Heart Disease dataset, all rankers have ranked Feature 12 as the most important feature.

Python:

Python is a programming language, which means it’a a language both people and computers can understand. Python was developed by a Dutch software engineer named Guido van Rossum, who created the language to solve some problems he saw in computer languages of the time.

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, and a syntax that allows programmers to express concepts in fewer lines of code, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

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. C Python, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. C Python is managed by the non-profit Python Software Foundation.

You Can Use Python for Pretty Much Anything

One significant advantage of learning Python is that it’s a general-purpose language that can be applied in a large variety of projects.

Below are just some of the most common fields where Python has found its use:

- Data science
- Scientific and mathematical computing
- Web development
- Computer graphics
- Basic game development
- Mapping and geography (GIS software)

Python Is Widely Used in Data Science

Python’s ecosystem is growing over the years and it’s more and more capable of the statistical analysis.

It’s the best compromise between scale and sophistication (in terms od data processing).

Python emphasizes productivity and readability.

Python is used by programmers that want to delve into data analysis or apply statistical techniques (and by devs that turn to data science)

There are plenty of Python scientific packages for data visualization, machine learning, natural language processing, complex data analysis and more. All of these factors make Python a great tool for scientific computing and a solid alternative for commercial packages such as MatLab. The most popular libraries and tools for data science are:

Pandas: a library for data manipulation and analysis. The library provides data structures and operations for manipulating numerical tables and time series.

NumPy: the fundamental package for scientific computing with Python, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays.

SciPy: a library used by scientists, analysts, and engineers doing scientific computing and technical computing.

Being a free, cross-platform, general-purpose and high-level programming language, Python has been widely adopted by the scientific community. Scientists value Python for its precise and efficient syntax, relatively flat learning curve and the fact that it integrates well with other languages (e.g. C/C++).

As a result of this popularity there are plenty of Python scientific packages for data visualization, machine learning, natural language processing, complex data analysis and more. All of these factors make Python a great tool for scientific computing and a solid alternative for commercial packages such as MatLab.

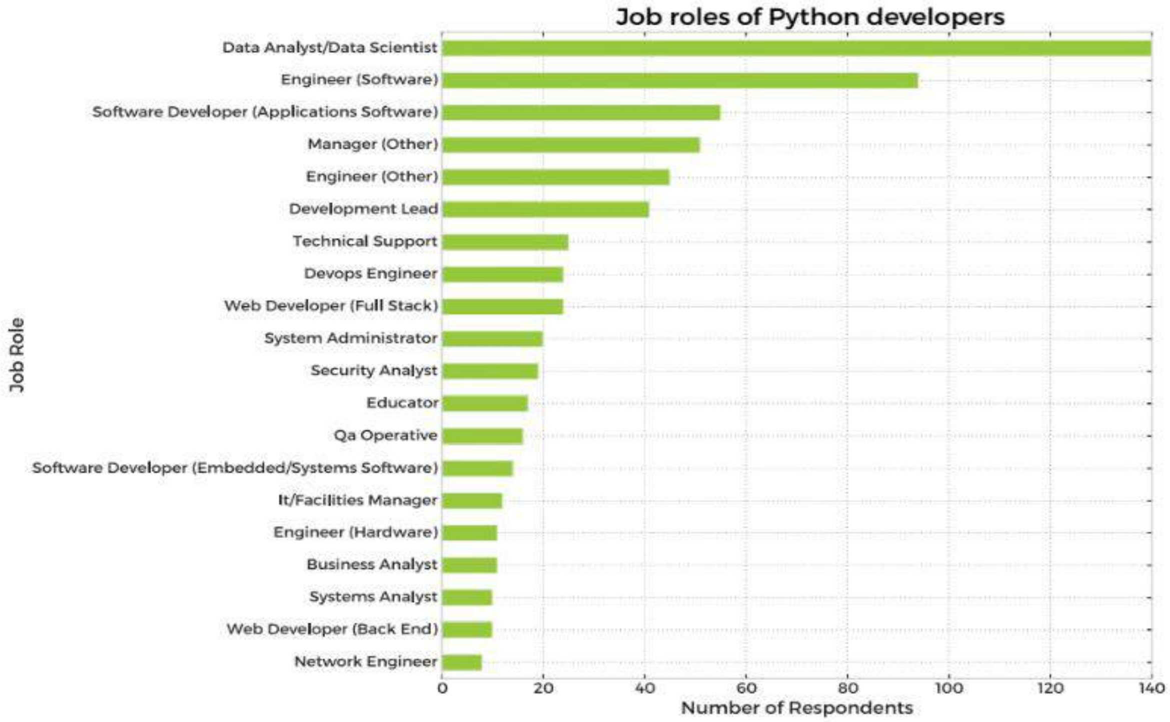


Figure 1.1: Job Roles of Python Developers

Here's our list of the most popular Python scientific libraries and tools

Astropy:

The Astropy Project is a collection of packages designed for use in astronomy. The core astropy package contains functionality aimed at professional astronomers and astrophysicists, but may be useful to anyone developing astronomy software.

Biopython:

Biopython is a collection of non-commercial Python tools for computational biology and bioinformatics. It contains classes to represent biological sequences and sequence annotations, and it is able to read and write to a variety of file formats.

Cubes:

Cubes is a light-weight Python framework and set of tools for the development of reporting and analytical applications, Online Analytical Processing (OLAP), multidimensional analysis and browsing of aggregated data.

DEAP:

DEAP is an evolutionary computation framework for rapid prototyping and testing of ideas. It incorporates the data structures and tools required to implement most common evolutionary computation techniques such as genetic algorithm, genetic programming, evolution strategies, particle swarm optimization, differential evolution and estimation of distribution algorithm.

SCOOP:

SCOOP is a Python module for distributing concurrent parallel tasks on various environments, from heterogeneous grids of workstations to supercomputers.

PsychoPy:

PsychoPy is a package for the generation of experiments for neuroscience and experimental psychology. PsychoPy is designed to allow the presentation of stimuli and collection of data for a wide range of neuroscience, psychology and psychophysics experiments.

Pandas:

Pandas is a library for data manipulation and analysis. The library provides data structures and operations for manipulating numerical tables and time series.

Mlpy:

Mlpy is a machine learning library built on top of NumPy/SciPy, the GNU Scientific Libraries. Mlpy provides a wide range of machine learning methods for supervised and unsupervised problems and it is aimed at finding a reasonable compromise between modularity, maintainability, reproducibility, usability and efficiency.

Matplotlib:

Matplotlib is a python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib allows you to generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, and more.

NumPy:

NumPy is the fundamental package for scientific computing with Python, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays.

NetworkX:

NetworkX is a library for studying graphs which helps you create, manipulate, and study the structure, dynamics, and functions of complex networks.

TomoPy:

TomoPy is an open-sourced Python toolbox to perform tomographic data processing and image reconstruction tasks. TomoPy provides a collaborative framework for the analysis of synchrotron tomographic data with the goal to unify the effort of different facilities and beamlines performing similar tasks.

Theano:

Theano is a numerical computation Python library. Theano allows you to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently.

SymPy:

SymPy is a library for symbolic computation and includes features ranging from basic symbolic arithmetic to calculus, algebra, discrete mathematics and quantum physics. It provides computer algebra capabilities either as a standalone application, as a library to other applications, or live on the web.

SciPy:

SciPy is a library used by scientists, analysts, and engineers doing scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.

Scikit-learn:

Scikit-learn is a machine learning library. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Scikit-image:

Scikit-image is a image processing library. It includes algorithms for segmentation, geometric transformations, color space manipulation, analysis, filtering, morphology, feature detection, and more.

ScientificPython:

ScientificPython is a collection of modules for scientific computing. It contains support for geometry, mathematical functions, statistics, physical units, IO, visualization, and parallelization.

SageMath:

SageMath is mathematical software with features covering many aspects of mathematics, including algebra, combinatorics, numerical mathematics, number theory, and calculus. SageMath uses the Python, supporting procedural, functional and object-oriented constructs.

Veusz:

Veusz is a scientific plotting and graphing package designed to produce publication-quality plots in popular vector formats, including PDF, PostScript and SVG.

Graph-tool:

Graph-tool is a module for the manipulation and statistical analysis of graphs.

SunPy:

SunPy is a data-analysis environment specializing in providing the software necessary to analyze solar and heliospheric data in Python.

Bokeh:

Bokeh is a Python interactive visualization library that targets modern web browsers for presentation. Bokeh can help anyone who would like to quickly and easily create interactive plots, dashboards, and data applications. Its goal is to provide elegant, concise construction of novel graphics in the style of D3.js, but also deliver this capability with high-performance interactivity over very large or streaming datasets.

TensorFlow:

TensorFlow is an open source software library for machine learning across a range of tasks, developed by Google to meet their needs for systems capable of building and training neural networks to detect and decipher patterns and correlations, analogous to the learning and reasoning which humans use. It is currently used for both research and production at Google products, often replacing the role of its closed-source predecessor, DistBelief.

Nilearn:

Nilearn is a Python module for fast and easy statistical learning on NeuroImaging data. Nilearn makes it easy to use many advanced machine learning, pattern recognition and multivariate statistical techniques on neuroimaging data for applications such as MVPA (Mutli-Voxel Pattern Analysis), decoding, predictive modelling, functional connectivity, brain parcellations, connectomes.

Dmelt:

DataMelt, or DMelt, is a software for numeric computation, statistics, analysis of large data volumes ("big data") and scientific visualization. The program can be used in many areas, such as natural sciences, engineering, modeling and analysis of financial markets. DMelt can be used with several scripting languages including Python/Jython, BeanShell, Groovy, Ruby, as well as with Java.

Python-weka-wrapper:

Weka is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The python-weka-wrapper package makes it easy to run Weka algorithms and filters from within Python.

Dask:

Dask is a flexible parallel computing library for analytic computing composed of two components: 1) dynamic task scheduling optimized for computation, optimized for interactive computational workloads, and 2) Big Data collections like parallel arrays, dataframes, and lists that extend common interfaces like NumPy, Pandas, or Python iterators to larger-than-memory or distributed environments.

Python Saves Time:

Even the classic “Hello, world” program illustrates this point:

```
print("Hello, world")
```

For comparison, this is what the same program looks like in Java:

```
public class HelloWorld {

    public static void main(String[] args) {

        System.out.println("Hello, world");

    }

}
```

All the Big Names Use Python

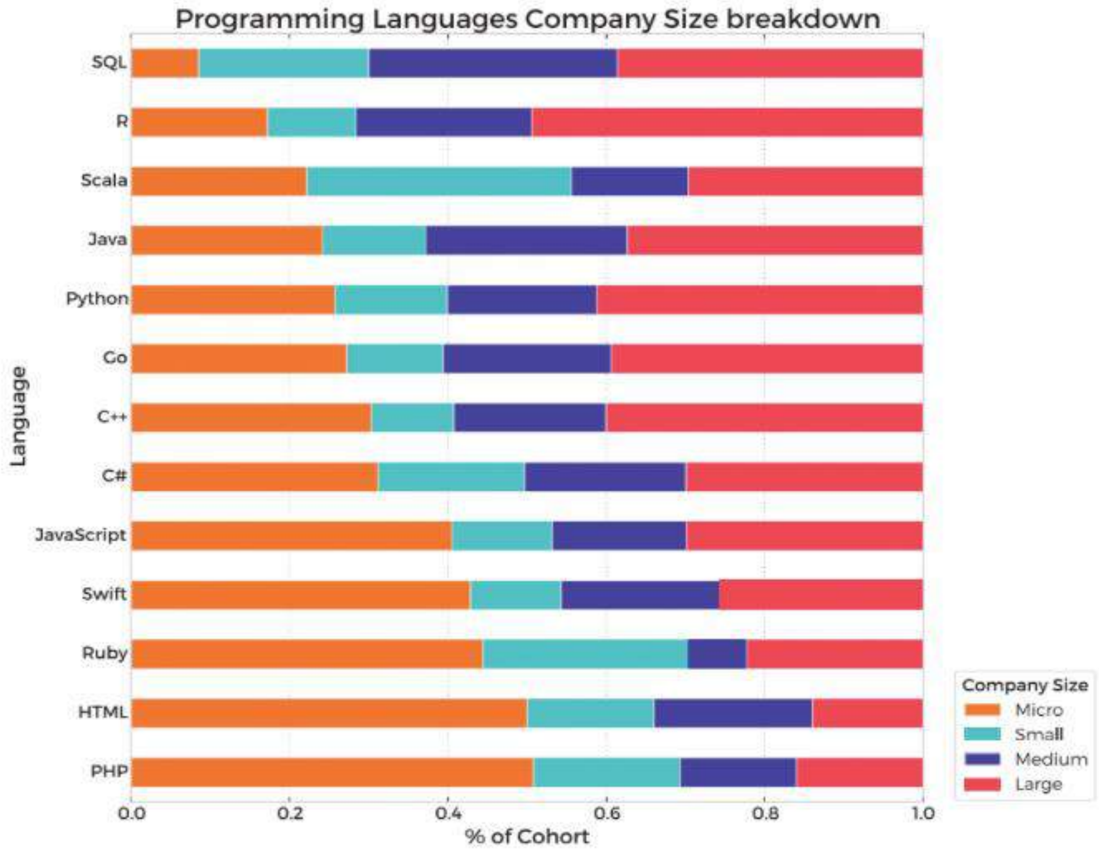


Figure 1.2: Programming Languages Company Size Breakdown

Python Keywords and Identifier:

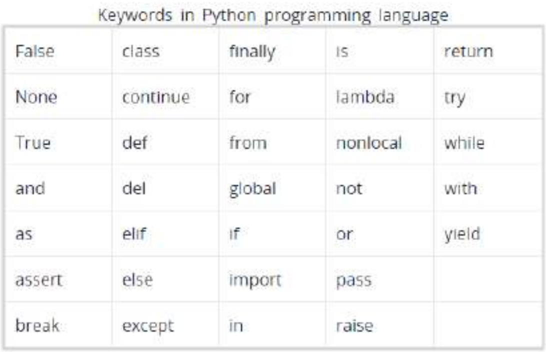
Keywords are the reserved words in Python.

We cannot use a keyword as variable name, function name or any other identifier. They are used to define the syntax and structure of the Python language.

In Python, keywords are case sensitive.

There are 33 keywords in Python 3.3. This number can vary slightly in course of time.

All the keywords except True, False and None are in lowercase and they must be written as it is. The list of all the keywords is given below.



Keywords in Python programming language

False	class	finally	is	return
None	continue	for	lambda	try
True	def	from	nonlocal	while
and	del	global	not	with
as	elif	if	or	yield
assert	else	import	pass	
break	except	in	raise	

Figure 1.3: Keywords in Python Language

Identifier is the name given to entities like class, functions, variables etc. in Python. It helps differentiating one entity from another.

Rules for writing identifiers:

Identifiers can be a combination of letters in lowercase (a to z) or uppercase (A to Z) or digits (0 to 9) or an underscore (_). Names like myClass, var_1 and print_this_to_screen, all are valid example.

An identifier cannot start with a digit. 1variable is invalid, but variable1 is perfectly fine.

Keywords cannot be used as identifiers.

```
>>> global = 1
  File "<interactive input>", line 1
    global = 1
           ^
```

SyntaxError: invalid syntax

We cannot use special symbols like !, @, #, $, % etc. in our identifier.

```
>>> a@ = 0
  File "<interactive input>", line 1
    a@ = 0
     ^
SyntaxError: invalid syntax
```

Identifier can be of any length.

Python is a broadly utilized abnormal state programming dialect for universally useful programming Python highlights a dynamic sort framework and programmed memory administration and backings various programming ideal models, including object-arranged, basic, utilitarian programming, and procedural styles. It has an expansive and complete standard library. Python mediators are accessible for some working frameworks, permitting Python code to keep running on a wide assortment of framework.

Figure 1.4: Python Logo

2. SYSTEM ANALYSIS

2. SYSTEM ANALYSIS

2.1 PROBLEM DEFINITION

A model aimed at predicting the adoption of technology in the health care sector has had a tremendously positive impact on medical processes along with the practices in which health care professionals engage..

2.2 EXISTING SYSTEM

Many attempts have been by researchers in the past for predicting flight delays using Machine Learning, Deep Learning and Big Data approaches..

Kalliguddi(author) constructed regression models like Decision Tree Regressor, Random Forest regressor on flight data for predicting both departure and arrival delays.

The main issues is to find the error rate in terms od predictions and reducting the error factor in the model.

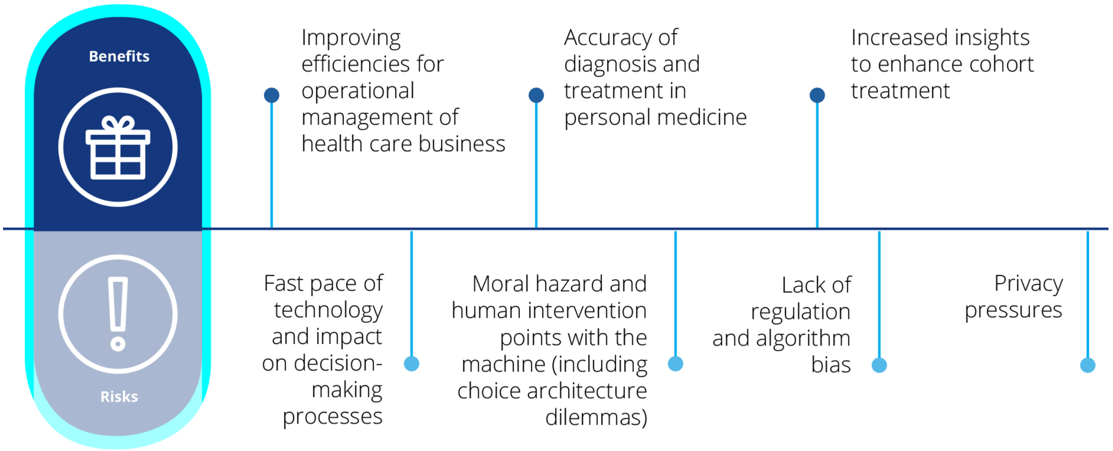
2.2.1 LIMITATIONS OF EXISTING SYSTEM

Predictive algorithms enable computers to recognize patterns in data and draw deductions from those patterns that show the likelihood of particular events occurring in the future.

This kind of algorithm is used in many types of activities, ranging from detection of credit card fraud and the optimization of search engines to stock market analysis and speech recognition.

FIGURE 1

Benefits and risks associated with predictive analytics in health care


Benefits
Improving efficiencies for operational management of health care business
Accuracy of diagnosis and treatment in personal medicine
Increased insights to enhance cohort treatment
Risks
Fast pace of technology and impact on decision-making processes
Moral hazard and human intervention points with the machine (including choice architecture dilemmas)
Lack of regulation and algorithm bias
Privacy pressures

Source: Deloitte analysis.

Deloitte Insights | www.deloitte.com/insights

.

2.3 PROPOSED SYSTEM

Health forecasting can be translated into effective interventions with individual patients, the analytic tools will be useless. So healthcare organizations must develop the infrastructure and the culture required to turn the data into action. That infrastructure must provide the ability to generate timely reports and use automation tools to apply intervention strategies across a patient population.

2.3.1 ADVANTAGES OF THE PROPOSED SYSTEM

Improving efficiencies for operational management of health care business operations

Accuracy of diagnosis and treatment in personal medicine

Increased insights to enhance cohort treatment
The current interest in predictive modeling is part of a larger trend to employ business and clinical intelligence (B&CI) applications in healthcare.

For optimal use in chronic disease management, predictive analytics should be applied to longitudinal rather than episodic data.

This requires getting patients involved. For example, patients might be asked to fill out online functional status surveys at regular intervals.

In select healthcare settings, remote monitoring data may also be routinely available.

Project Planning:

Step1: Data set collection.

Step2: Feature Extraction

Step3: Data processing and corelation between the features

Step4: apply regression.

Step5: Predict the results and evlation of Error factors.

The reason for this SRS record is to distinguish the necessities and functionalities for Intelligent Network Backup Tool. The SRS will characterize how our group and the customer consider the last item and the attributes or usefulness it must have. This record additionally makes a note of the discretionary prerequisites which we intend to execute yet are not required for the working of the venture.

This stage assesses the required necessities for the Images Processing for an orderly method for assessing the prerequisites a few procedures are included. The initial step associated with dissecting the prerequisites of the framework is perceiving the idea of framework for a solid examination and all the case are defined to better comprehend the investigation of the dataset.

INTENDED AUDIENCE AND READING SUGGESTIONS

This record is proposed for extend engineers, directors, clients, analyzers and documentation journalists. This report goes for examining plan and execution imperatives, conditions, framework highlights, outside interface prerequisites and other non utilitarian necessities.

IDENTIFICATION OF NEEDS

The first and imperative need for a business firm or an association is to know how they are performing in the market and parallelly they have to know how to conquer their rivals in the market.

To do as such we have to investigation our information in view of all the accessible variables

2.4 FEASIBILITY STUDY

A credibility contemplate expects to fair-mindedly and soundly uncover the qualities and inadequacies of a present business or proposed meander, openings and threats present in nature, the benefits required to bring through, and in the long run the prospects for advance. In its most clear terms, the two criteria to judge believability are incurred significant injury required and motivator to the fulfilled.

An inside and out arranged feasibility ponder should give a recorded establishment of the business or wander, a delineation of the thing or organization, accounting explanations, purposes of enthusiasm of the operations and organization, publicizing examination and game plans, budgetary data, authentic necessities and cost duties. All things considered, plausibility looks at go before specific change and wander utilization. There are three sorts of attainability

- Economical Feasibility
- Technical Feasibility
- Operational Feasibility

2.4.1 ECONOMICAL FEASIBILITY

The electronic structure manages the present existing system's data stream and technique absolutely and should make each one of the reports of the manual structure other than a substantial gathering of the other organization reports. It should be filled in as an electronic application with specific web server and database server. Advance a segment of the associated trades happen in different ranges. Open source programming like TOMCAT, JAVA, MySQL and Linux is used to restrict the cost for the Customer. No extraordinary wander need to manage the instrument.

2.4.2 TECHNICAL FEASIBILITY

Surveying the particular probability is the trickiest bit of a believability consider. This is in light of the fact that, starting at the present moment, not a lot of point by point layout of the system, making it difficult to get to issues like execution, costs on (by excellence of the kind of development to be passed on) et cetera.

Different issues must be considered while doing a particular examination. Grasp the differing progressions required in the proposed system. Before starting the wander, we should be clear about what are the advances that are to be required for the change of the new system. Check whether the affiliation by and by has the required advancements. Is the required development open with the affiliation?

In case so is the utmost sufficient?

For instance – "Will the present printer have the ability to manage the new reports and structures required for the new system?"

2.4.3 OPERATIONAL FEASIBILITY

Proposed wanders are profitable just if they can be changed into information systems that will meet the affiliations working necessities. Simply communicated, this trial of probability asks with reference to whether the structure will work when it is made and presented. Are there genuine obstacles to Implementation? Here are questions that will help test the operational achievability of a wander.

- Is there sufficient help for the wander from organization from customers? In case the present structure is particularly cherished and used to the extent that individuals won't have the ability to see purposes behind change, there may be resistance.
- Are the present business methodologies qualified to the customer? If they are not, Users may welcome a change that will accomplish a more operational and supportive systems.

- Have the customer been locked in with the orchestrating and change of the wander? Early commitment decreases the chances of impenetrability to the structure.

2.5 HARDWARE AND SOFTWARE REQUIREMENTS

2.5.1 SOFTWARE REQUIREMENTS

Operating System	:	Windows
Framework	:	Jupyter
Language	:	Python
IDE	:	Anaconda

2.5.2 HARDWARE REQUIREMENTS

Processor	:	Pentium 4
Hard disc	:	500GB
RAM	:	4GB

System with all standard accessories like monitor, keyboard, mouse, etc.

3. ARCHITECTURE

3. ARCHITECTURE

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

This section describes the system in narrative form using non-technical terms. It should provide a high-level system architecture diagram showing a subsystem breakout of the system, if applicable. The high-level system architecture or subsystem diagrams should, if applicable, show interfaces to external systems. Supply a high-level context diagram for the system and subsystems, if applicable. Refer to the requirements trace ability matrix (RTM) in the Functional Requirements Document (FRD), to identify the allocation of the functional requirements into this design document.

This section describes any constraints in the system design (reference any trade-off analyses conducted such, as resource use versus productivity, or conflicts with other systems) and includes any assumptions made by the project team in developing the system design.

The organization code and title of the key points of contact (and alternates if appropriate) for the information system development effort. These points of contact should include the Project Manager, System Proponent, User Organization, Quality Assurance (QA) Manager, Security Manager, and Configuration Manager, as appropriate.

3.1 PROJECT ARCHITECTURE

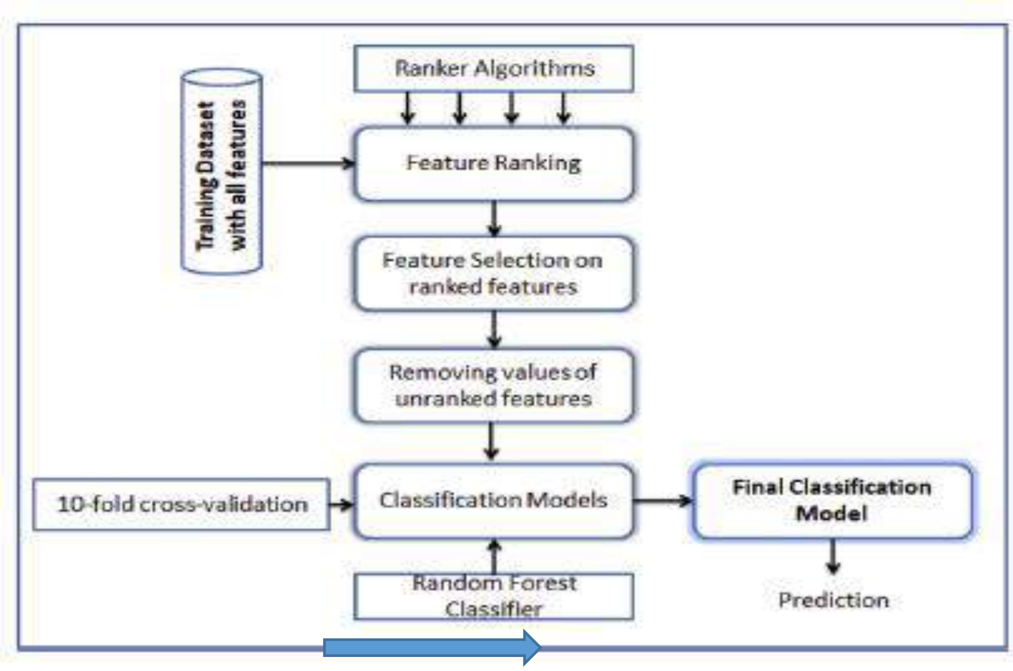

Training Dataset with all features
Ranker Algorithms
Feature Ranking
Feature Selection on ranked features
Removing values of unranked features
10-fold cross-validation
Classification Models
Final Classification Model
Random Forest Classifier
Prediction

Figure 3.1: Project Architecture of City Health Prediction Model Using Random Forest Classification Method

3.2 MODULES

DATA AND METHODOLOGY:

The data is obtained from Copernicus website which is a free database. It has satellite images from all over the world. It updates regularly and images from a certain date can be obtained. The image obtained is then further processed to get the desired data. Ground Truth data is referred to as the data that is collected on field by observation. In this step, data is collected by visiting some fields that have known objects. These fields may include vegetation fields, buildings, roads, rivers etc.

PROCESSING:

After stacking the image is now ready to be processed. The processing takes place in a software called ENVI. First the ground truth Regions Of Interest are loaded on the stacked image. The image having Ground truth ROIs loaded over it is shown in figure-1. When the Ground truth ROIs appear on the image, a mask is made in order to select only the desired image from the empty background. After making mask, the desired machine learning algorithms are run on the image. In this project we have used both supervised and unsupervised algorithms. The supervised algorithms include neural network and support vector machine while unsupervised algorithm includes K-means clustering. These particular algorithms are selected because they are the most popular and widely used algorithms that are present for use in the ENVI software. Secondly it was aimed to compare one of the most popular algorithm from unsupervised classification to two most popular supervised classification algorithms

POST PROCESSING:

After the data is processed and image is classified then begins the phase of post processing. In post processing we make confusion matrix. The confusion matrix gives the accuracy and statistics of the result.

UML DIAGRAMS

UML (Unified Modeling Language) is a standard vernacular for choosing, envisioning, making, and specifying the collectibles of programming structures. UML is a pictorial vernacular used to make programming blue prints. It is in like way used to exhibit non programming structures similarly like process stream in a gathering unit and so forth.

UML is not a programming vernacular yet rather instruments can be utilized to make code in different tongues utilizing UML graphs. UML has an incite relationship with question composed examination and outline. UML expect a fundamental part in portraying trade viewpoints of a structure.

3.3 USE CASE DIAGRAM

The use case graph is for demonstrating the direct of the structure. This chart contains the course of action of use cases, performing pros and their relationship. This chart might be utilized to address the static perspective of the structure.

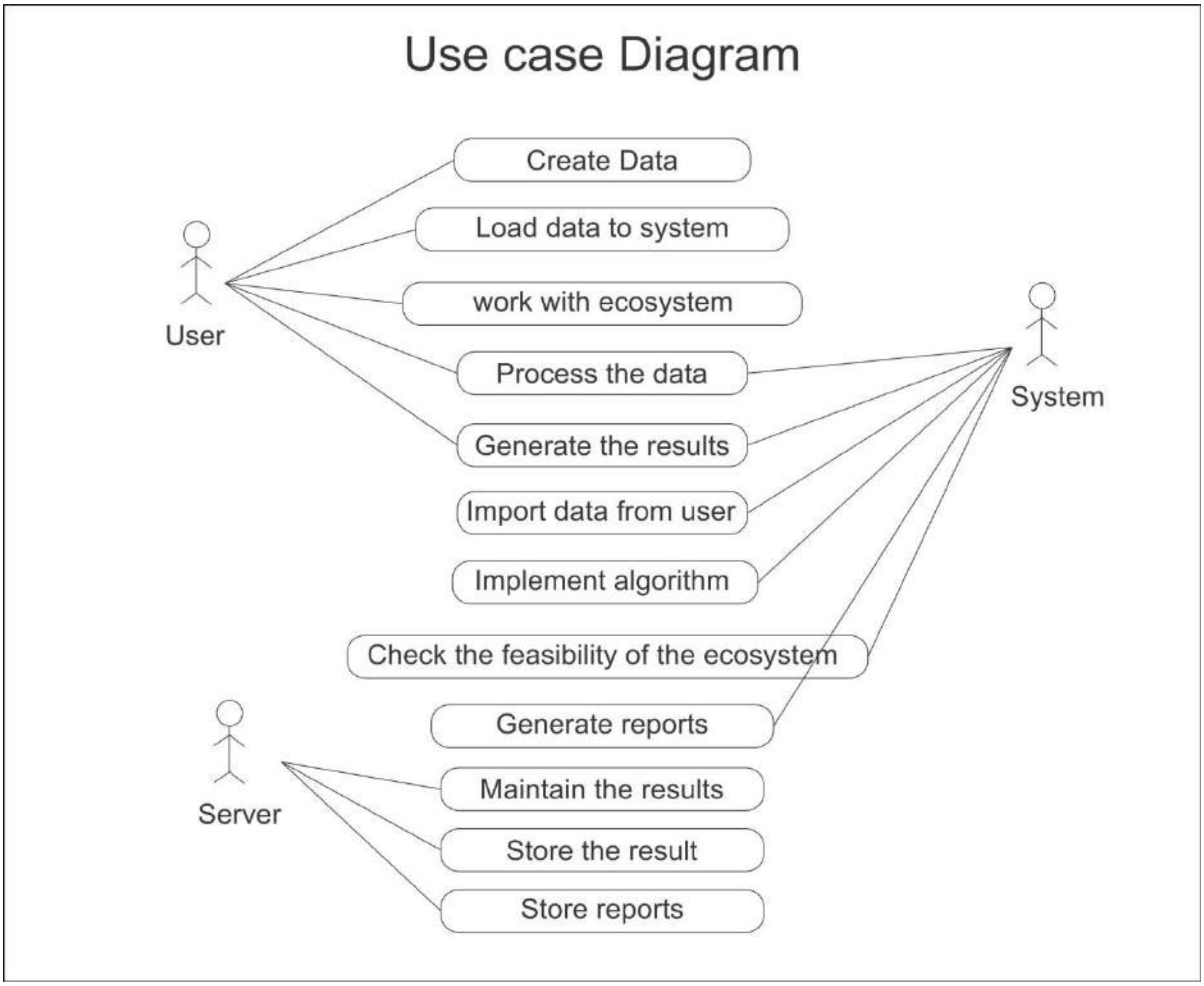


Figure 3.2: Use Case Diagram for City Health Prediction Model Using Random Forest Classification Method

In the above diagram, the performing specialists are customer, structure, client, server, Hadoop and data cleaning. The client exchanges the data to the system which disengages the data into squares and gives the data to Hadoop. By then Hadoop does the data cleaning which is just performing data connection and data repairing, by then the results will be secured. These results can be seen using Hadoop and can be secured in server for future reason. The gained results can be created as reports by the customer.

3.4 CLASS DIAGRAM

The class graph is the most normally pulled in layout UML. It addresses the static course of action perspective of the structure. It solidifies the strategy of classes, interfaces, joint attempts and their affiliations.

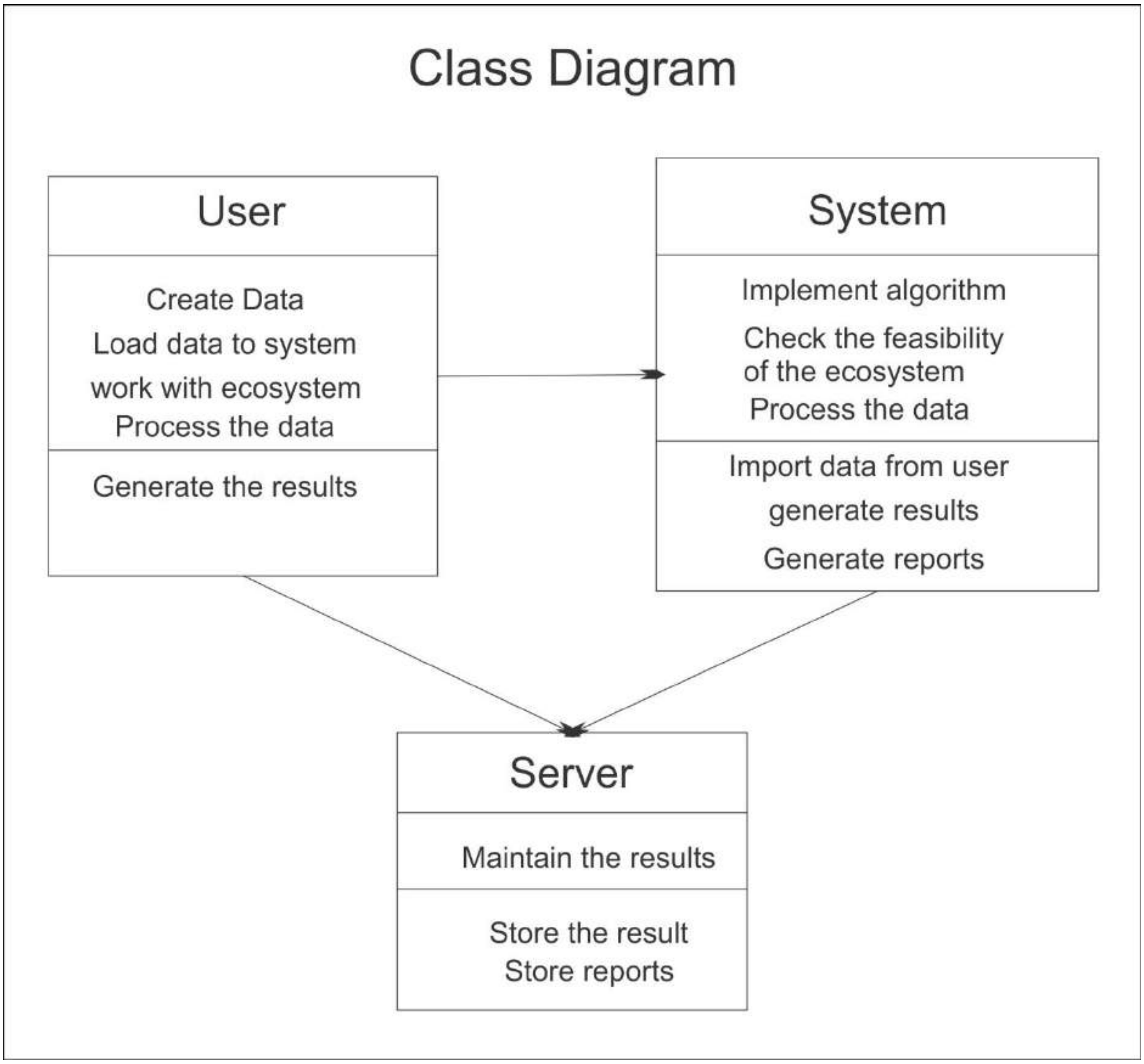


Figure 3.3: Class Diagram for City Health Prediction Model Using Random Forest Classification Method

In the above class diagram, the relationship that is the dependence between each one of the classes is sketched out. Additionally, even the operations performed in each and every class is similarly appeared.

3.5 SEQUENCE DIAGRAM

This is a cooperation design which tends to the time requesting of messages. It includes set of parts and the messages sent and gotten by the instance of parts. This chart is utilized to address the dynamic perspective of the structure.

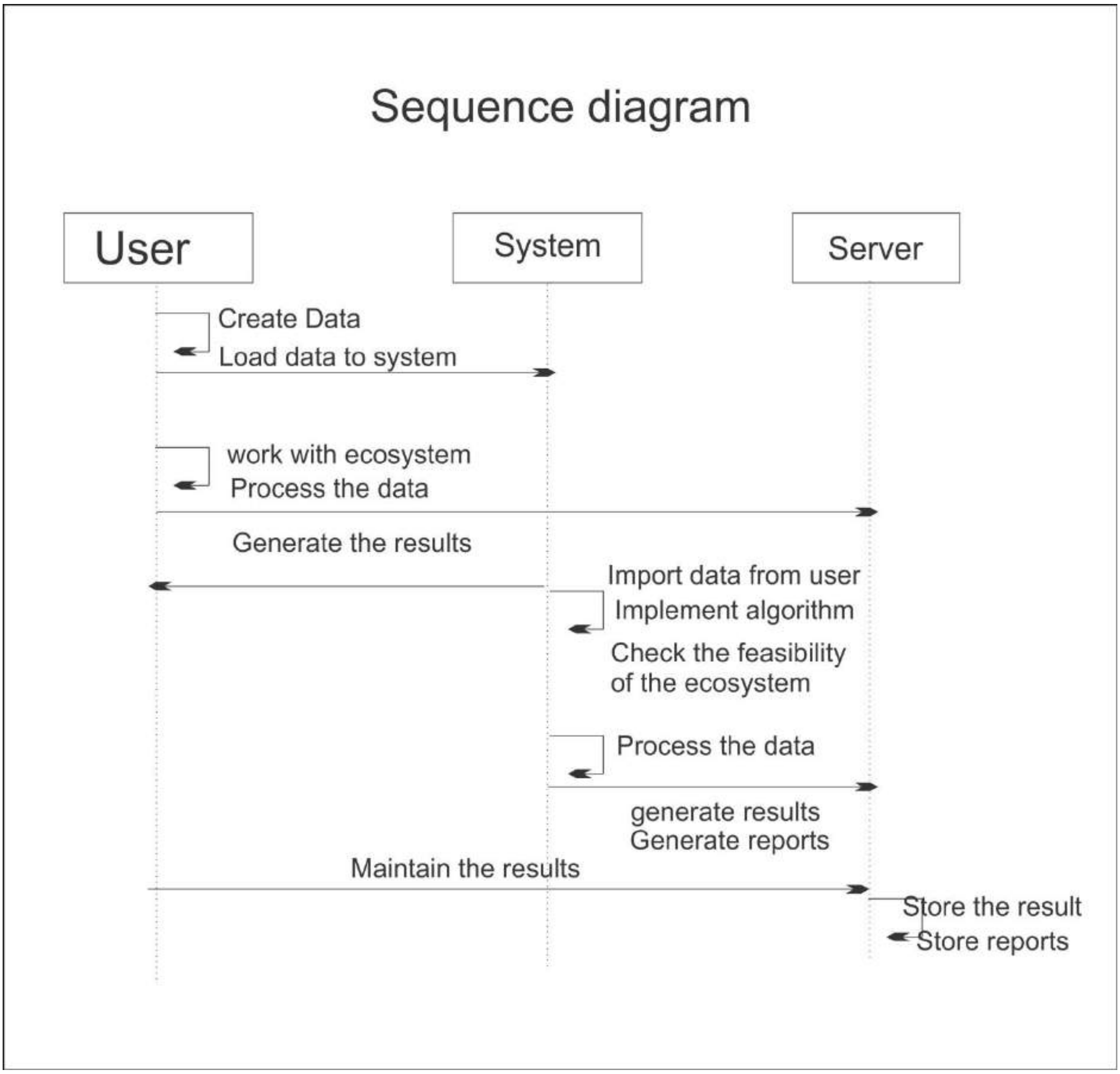


Figure 3.4: Sequence Diagram for City Health Prediction Model Using Random Forest Classification Method

A succession outline indicates question communications masterminded in time arrangement. In the above graph, there are five articles cooperating with each other. Each protest has a vertical dashed line which speaks to the presence of a question over some undefined time frame. This graph has additionally a tall, thin rectangle which is called center of control that demonstrates the timeframe amid which a protest is playing out an activity, either specifically or through a subordinate system.

3.6 COLLABORATION DIAGRAM

This is a support format, which tends to the principal relationship of articles that send and get messages. It incorporates set of parts, connectors that interface the parts and the messages sent and get by those parts. This graph is utilized to address the dynamic perspective of the framework.

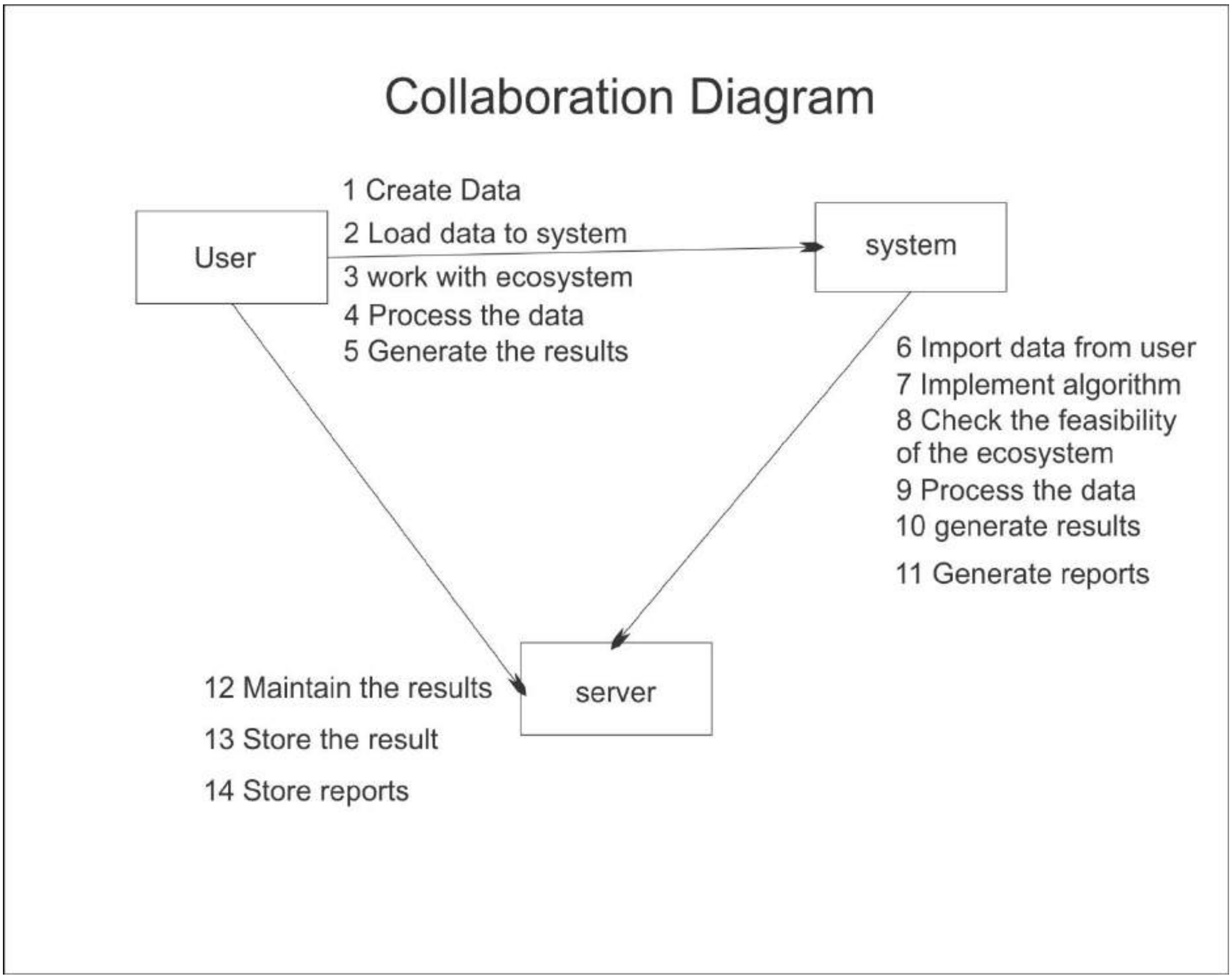


Figure 3.5: Collaboration Diagram for City Health Prediction Model Using Random Forest Classification Method

The joint effort outline contains articles, way and arrangement number. In the above graph, there are five questions specifically customer, client, framework, Hadoop and server. These items are connected to each other utilizing a way. A succession number show the time request of a message.

3.7 ACTIVITY DIAGRAM

The state graph contains the game-plan of states, occasions and exercises. This graph is noteworthy for tending to the lead of the interface, class and made effort. The key centralization of state outline is to show the occasion sort out lead of the request. The state follows diagram the dynamic perspective of the framework.

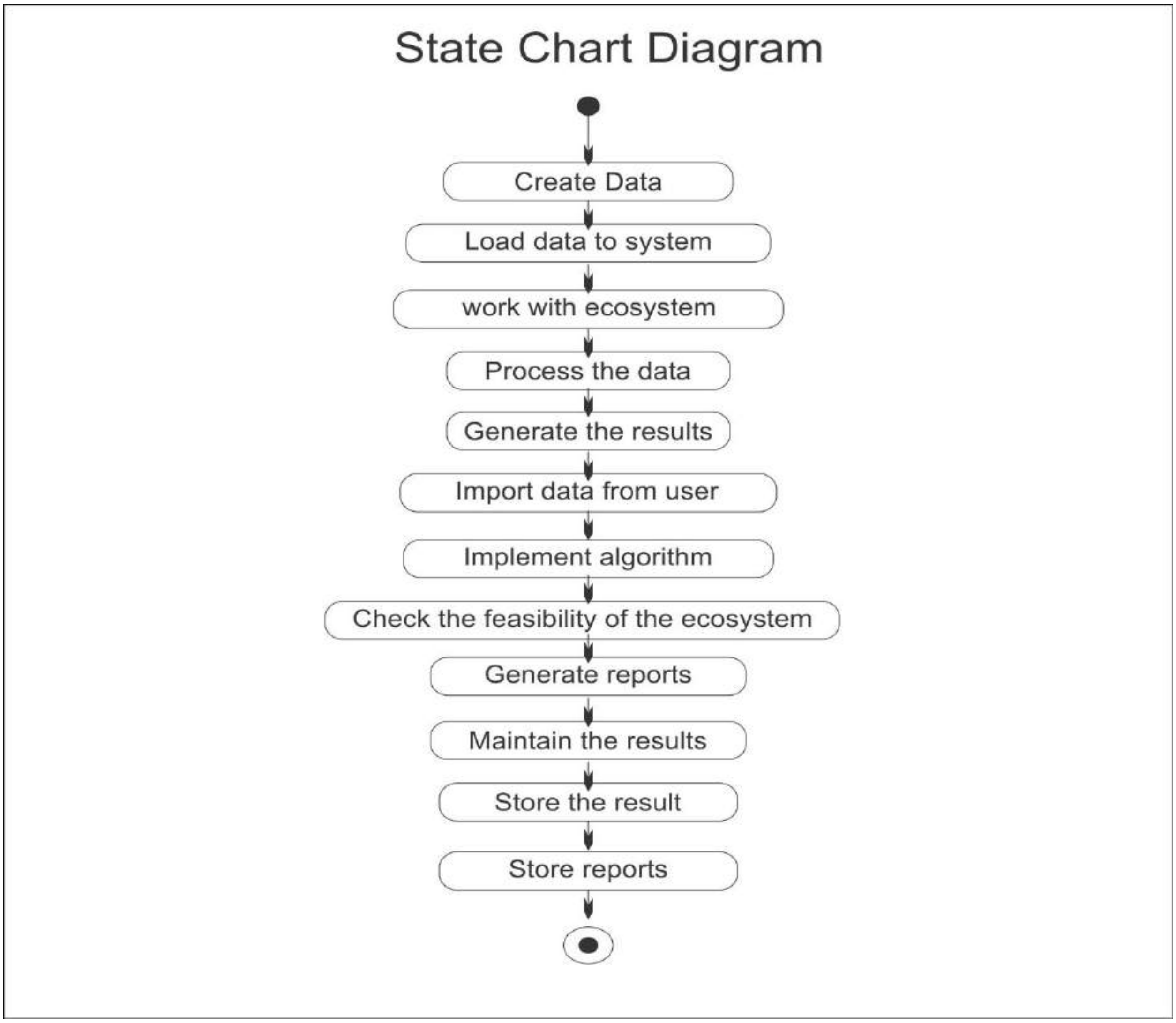


Figure 3.6: Activity Diagram for City Health Prediction Model Using Random Forest Classification Method

A state outline graph contains two components called states and progress. States speak to circumstances amid the life of a question. We can without much of a stretch outline a state in Smart Draw by utilizing a rectangle with adjusted corners. Change is a strong bolt speaks to the way between various conditions of a question. Name the change with the occasion that activated it and the activity those outcomes from it.

3.8 COMPONENT DIAGRAM

The imperative portion of part format is segment. This diagram demonstrates within parts, connectors and ports that understand the piece. Precisely when section is instantiated, duplicates of inside parts are besides instantiated.

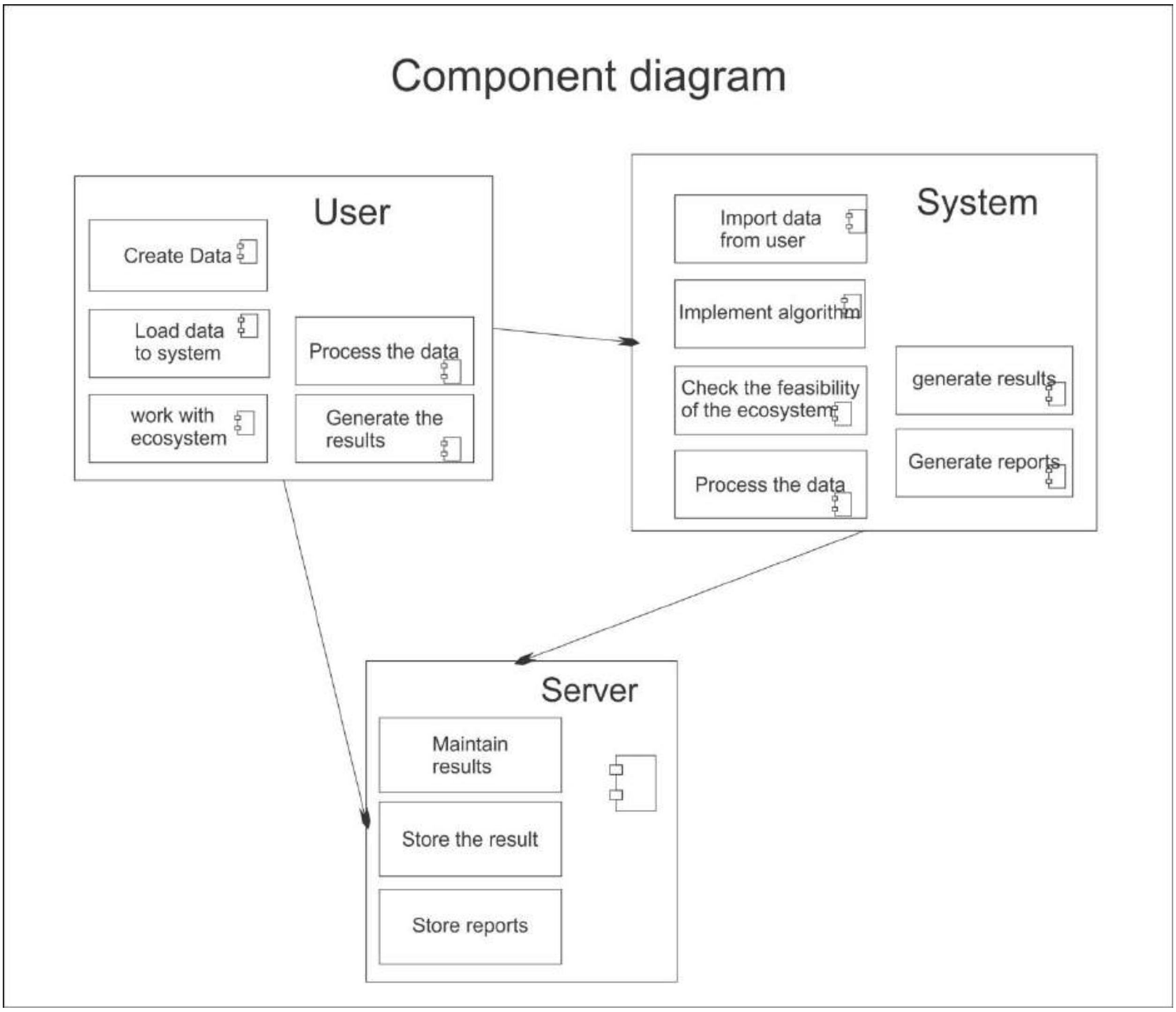


Figure 3.7: Component Diagram for City Health Prediction Model Using Random Forest Classification Method

A part outline is spoken to utilizing segment. A part is a physical building piece of the framework. It is spoken to as a rectangle with tab. Part outline portrays the inward handling of the venture. The information is sent to the Hadoop where sqoop is utilized for information cleaning and the reports are produced utilizing hive.

3.9 DEPLOYMENT DIAGRAM

The fundamental fragment in game-plan layout is a middle point. The strategy of focus focuses and their relationship with other is tended to utilizing sending plot. The sending outline is identified with the area diagram, that is one focus purpose obviously of activity format frequently includes no short of what one sections. This outline is in like way critical for tending to the static perspective of the framework.

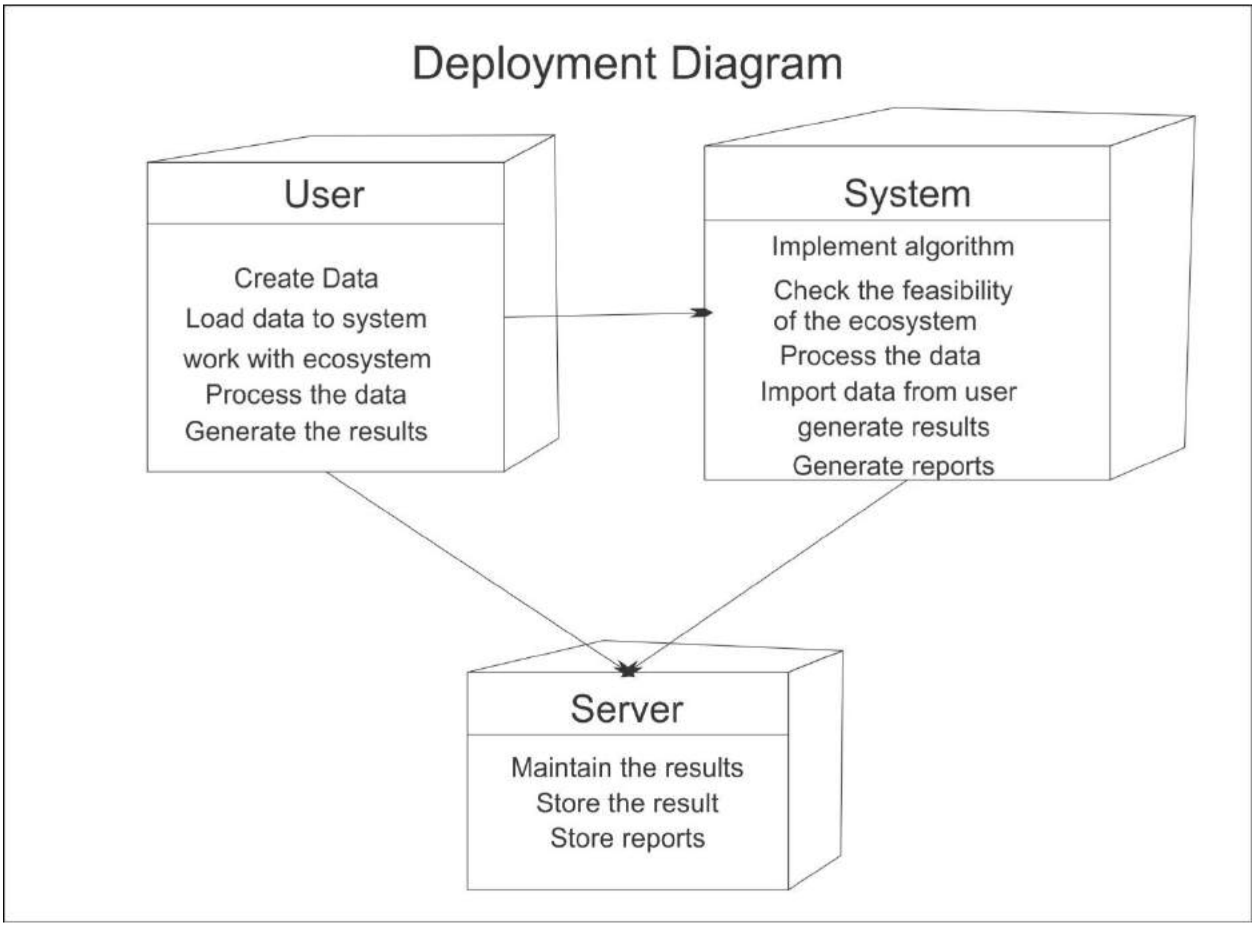


Figure 3.8: Deployment Diagram for City Health Prediction Model Using Random Forest Classification Method

An arrangement graph is spoken to utilizing hub. A hub is a physical asset that executes code parts. They are likewise used to portray run time handling of hubs. The information is sent to the Hadoop where sqoop is utilized for information cleaning and the reports are produced utilizing hive.

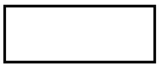
3.10 DATA FLOW DIAGRAMS

An information stream design (DFD) is a graphical portrayal of the "stream" of information through a data framework, demonstrating its strategy edges. A DFD is a significant part of the time utilized as a preparatory stroll to make an overview of the framework, which can later be cleared up. DFDs can in like way be utilized for the depiction of information prepare. A DFD indicates what sort of data will be sense of duty regarding and yield from the structure, where the information will begin from and go to, and where the information will be secured. It doesn't demonstrate data about the organizing of process or data about whether strategy will work in game-plan or in parallel.

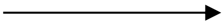
DFD Symbols:

In the DFD, there are four symbols

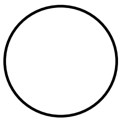
- A square defines a source or destination of system data.



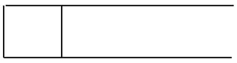
- An arrow identifies data flow. It is the pipeline through which the information flows.



- A circle represents a process that transforms incoming data flow into outgoing data flow.



- An open rectangle is a data store, data at rest or a temporary repository of data.



Level 0: System input/ output level

A level 0 DFD describes the system wide boundaries, dealing input to and output flow from the system and major processes.

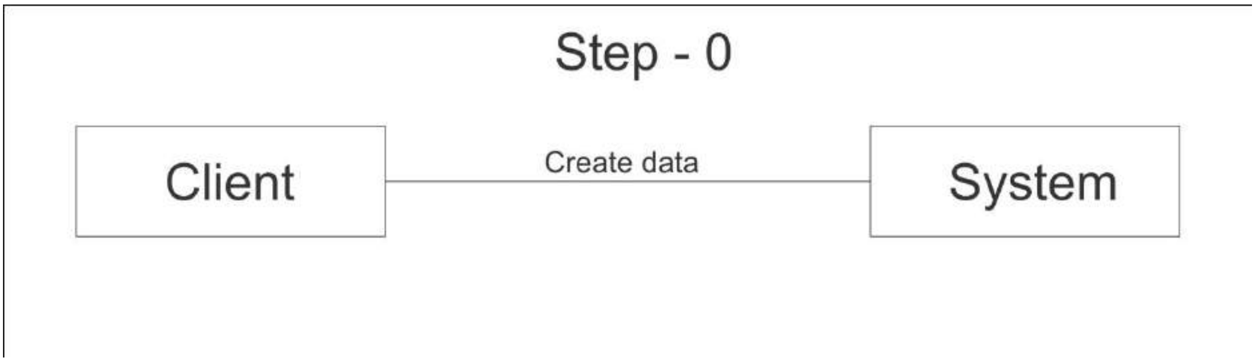


Figure 3.9: Level 0 DFD

DFD Level 0 is in like way called a Context Diagram. It's a urgent review of the entire structure or process being bankrupt down or appeared. It's required to be an at first watch, demonstrating the framework as a particular surprising state handle, with its relationship to outside substances.

Level 1: Sub system level data flow

Level 1 DFD delineates the accompanying level of purposes of enthusiasm with the data stream between subsystems. The Level 1 DFD exhibits how the system is secluded into sub-structures (shapes), each of which oversees no less than one of the data streams to or from an outside pro, and which together give most of the helpfulness of the system as a rule.

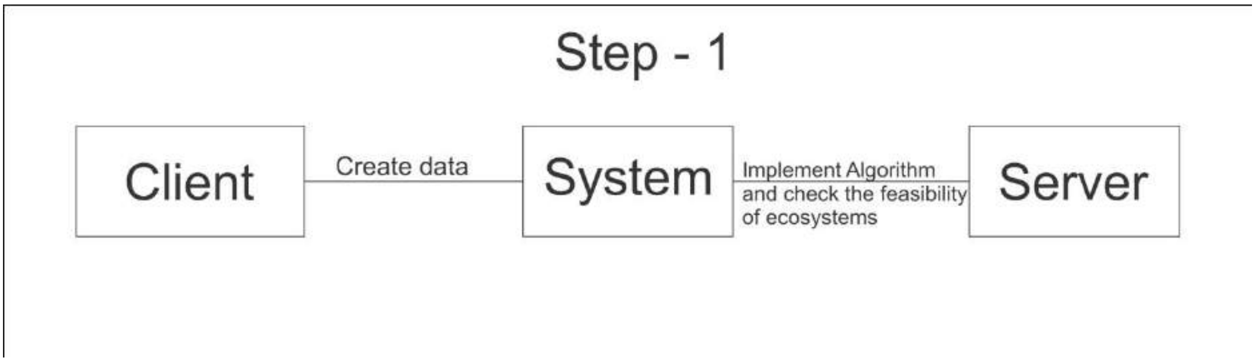


Figure 3.10: Level 1 DFD

Level 2: File level detail data flow

Plausibility and danger examination are connected here from various perspectives. The level 2 DFD elucidates the fundamental level of understanding about the system's working.

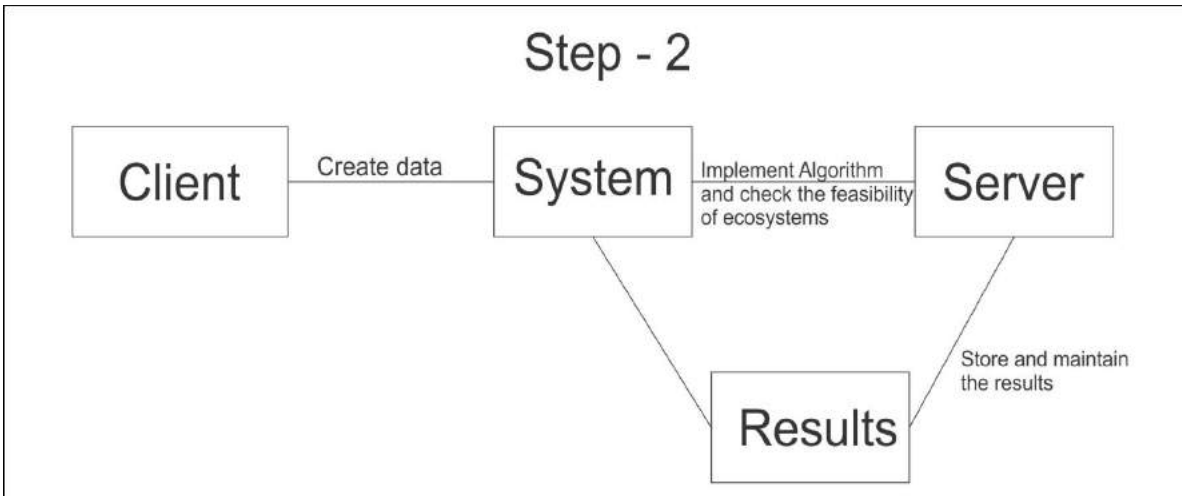

Step - 2
Client
Create data
System
Implement Algorithm and check the feasibility of ecosystems
Server
Store and maintain the results
Results

Fig 3.11 Level 2 DFD

4. IMPLEMENTATION

4. IMPLEMENTATION

4.1 SAMPLE CODE

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

health= pd.read_csv("C:/Users/sravan/data.csv")
health.head()

health.info()

health.isnull().sum() #checking for misisng values
health.Value.value_counts() #checking the values present in each columns

health.corr()

health.Indicator.value_counts()

health.Year.value_counts()

health["Race/ Ethnicity"].value_counts()

health["Indicator Category"].value_counts()
health.Place.value_counts()

health["Value"]= health["Value"].fillna(health["Value"].mean())
for column in ["Source","BCHC Requested Methodology"]:
    health[column].fillna(health[column].mode()[0], inplace= True)
```

```
health= health.drop(columns=["Methods","Notes"])
health.head()

health.isnull().sum()

# Data Visualization
sns.countplot(y= health["Indicator Category"])
plt.show()

sns.countplot(y= health["Year"])
plt.show()

sns.countplot(y= health["Race/ Ethnicity"])
#plt.figure(figsize=(100,100))
plt.show()


groupvalues=health.groupby('Indicator Category').sum().reset_index()
groupvalues

plt.figure(figsize=(15,10))
sns.set(style="darkgrid")
g = sns.barplot(groupvalues["Indicator Category"],groupvalues['Value'])
for index, row in groupvalues.iterrows():
    g.text(row.name,row.Value, round(row.Value,2), color='black', ha="center")
    g.set_xticklabels(g.get_xticklabels(),rotation= 90, fontsize= 18)
    g.set_xlabel("Indicator Category", fontsize=18)
plt.show()
```

```
sns.countplot(health["Gender"])
```

```
plt.show()


plt.figure(figsize = (10,5))
labels = 'Male', 'Female', 'Both'
sizes = np.array([12.4, 17.9, 69.6])
colors = ['yellowgreen', 'violet', 'yellow']


p, tx, autotexts = plt.pie(sizes, labels=labels, colors=colors,
            autopct="", shadow=True)


for i, a in enumerate(autotexts):
    a.set_text("{}".format(sizes[i]))


plt.axis('equal')
plt.show()
plt.figure(figsize=(40,10))
sns.set(style="darkgrid")
groupvalues1=health.groupby('Place').sum().reset_index()
g = sns.barplot(groupvalues1['Place'],groupvalues1['Value'])
for index, row in groupvalues.iterrows():
    g.text(row.name,row.Value, round(row.Value,2), color='black', ha="center")
    g.set_xticklabels(g.get_xticklabels(), rotation= 90, fontsize= 25)
    g.set_xlabel("Place")
plt.show()
health['Place'].value_counts()
health['State']=health['Place'].apply(lambda x: x.split(",")).str[1]
plt.figure(figsize=(15,10))
cp=sns.countplot(x=health['State'],data=health,order = health['State'].value_counts().index)
cp.set_xticklabels(cp.get_xticklabels(),rotation=90,fontsize=18)
cp.set_xlabel('State',fontsize=15)
```

```
cp.set_ylabel('Count',fontsize=10)
```

```
plt.show()


plt.figure(figsize = (15,10))
health.State.value_counts().plot(kind="pie")
plt.show()


sns.set(style="darkgrid")
ax = sns.countplot(y='Indicator Category', hue="Gender",data=health)
plt.show()


plt.figure(figsize=(40,10))
sns.set(style="darkgrid")
groupvalues2=health.groupby('Indicator').sum().reset_index()
g = sns.barplot(groupvalues2['Indicator'],groupvalues2['Value'])
for index, row in groupvalues.iterrows():
    g.text(row.name,row.Value, round(row.Value,2), color='black', ha="center")
    g.set_xticklabels(g.get_xticklabels(),rotation= 90, fontsize= 25)
    g.set_xlabel("Indicator")


plt.show()


plt.figure(figsize=(25,12))
cp=sns.countplot(x=health['Indicator'],data=health,hue=health['State'],order = 
health['Indicator'].value_counts().index)
cp.set_xticklabels(cp.get_xticklabels(),rotation=90,fontsize=18)
cp.set_xlabel('Indicator',fontsize=15)
cp.set_ylabel('Count',fontsize=18)
```

```
health_dummies= pd.get_dummies(health)
X= health_dummies.drop(columns= "Value")
Y= health_dummies["Value"]
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score
train_x, val_x, train_y, val_y= train_test_split(X,Y, test_size= 0.4, random_state= 100)
lr= LinearRegression()
lr.fit(train_x,train_y)
train_pred= lr.predict(train_x)
test_pred= lr.predict(val_x)

print("train_r2:", r2_score(train_pred,train_y))
```

5. SCREENSHOTS

5.1 DISEASE CATEGORIZATION

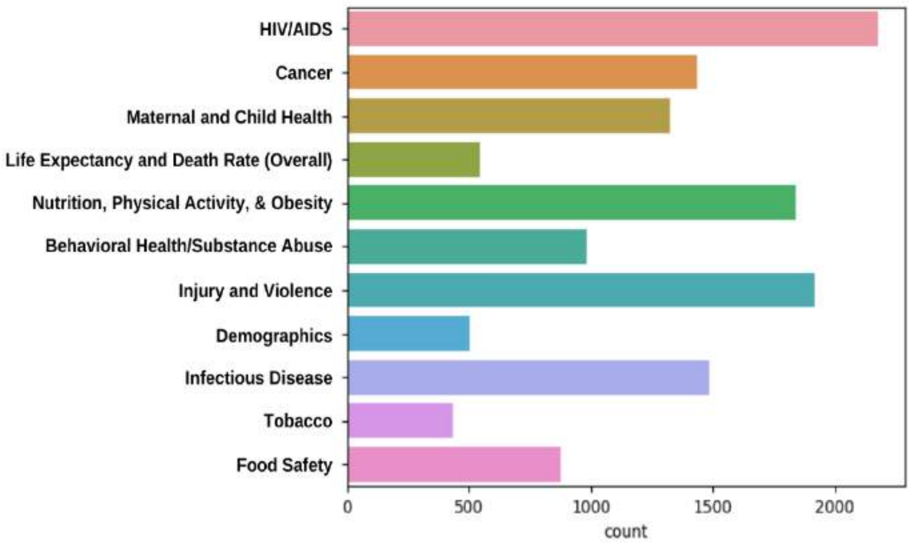


Fig-7.2 : Disease Categorization

Screenshot 5.1: Disease categorization of City Health Prediction Model Using Random Forest Classification Method

5.2 DATA VISUALIZATION

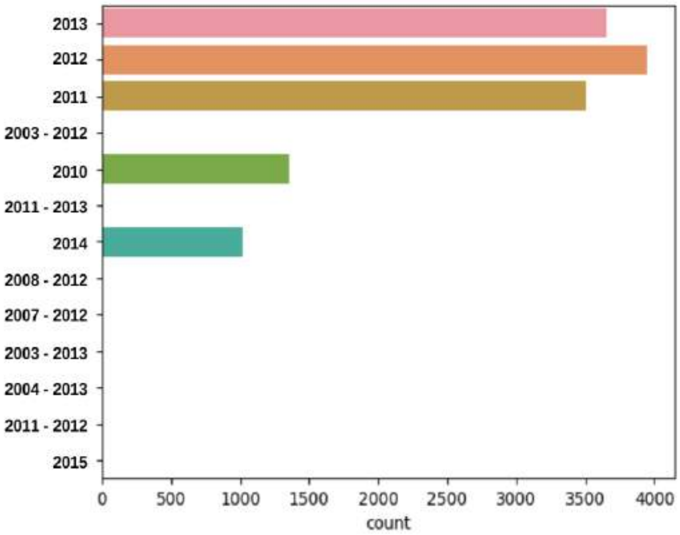


Fig-7.3 : Visualization of data based on the years where people effected most

Screenshot 5.2: Visualization of data based on the years where people effected most

5.3 RACE/ETHNICITY

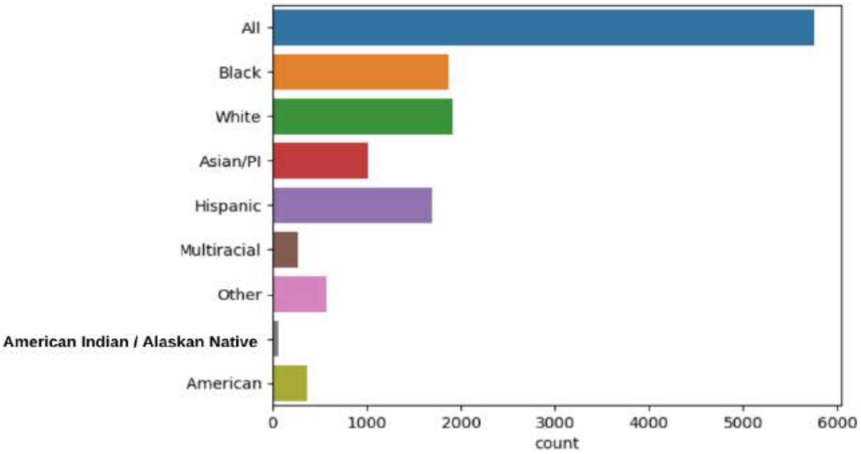


Fig-7.4 : Race/ Ethnicity

Screenshot 5.3: Race/Ethnicity

5.4 PEOPLE EFFECTED ACCORDING TO DISEASES

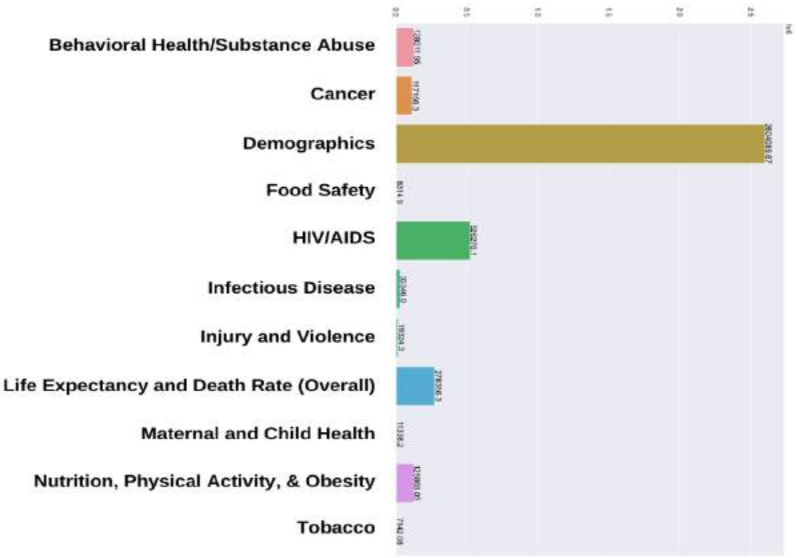


Fig-7.5 : People effected according to diseases

Screenshot 5.4: People effected according to diseases

5.5 MALE AND FEMALE DATA VISUALIZATION

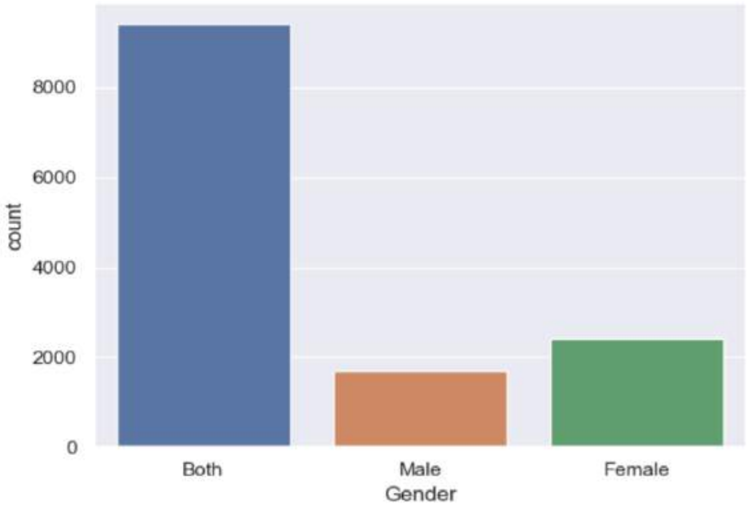


Fig-7.6 : Male and Female data visualization

Screenshot 5.2: Male and Female data visualization

5.6 MALE AND FEMALE DATA PIE VISUALIZATION

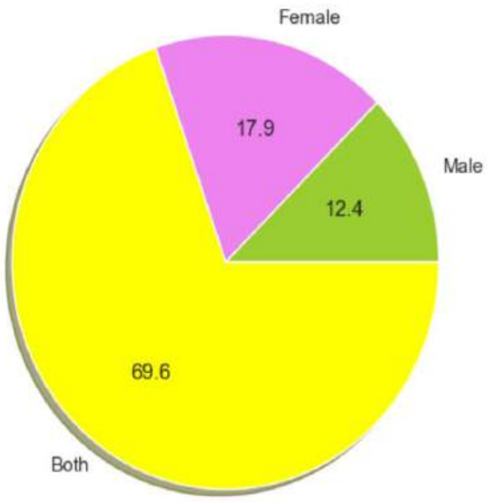


Fig-7.7 : Male and Female data pie visualization

Screenshot 5.6: Male and Female data pie visualization

5.7 STATES EFFECTED BASED ON THEIR RESULTS

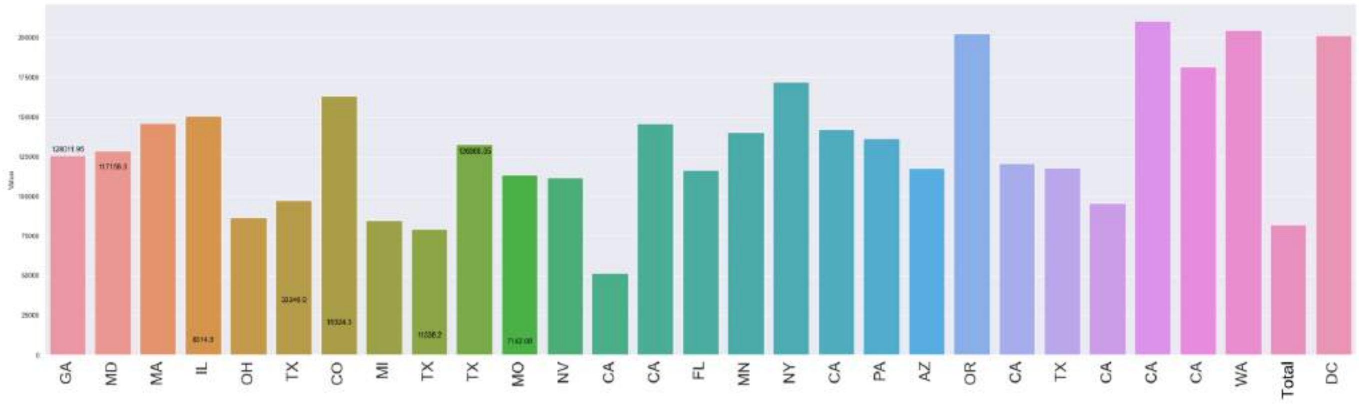


Fig-7.8 : State effected based on their results

Screenshot 5.2: States effected based on result

5.8 STATES EFFECTED BASED ON THEIR POPULATION

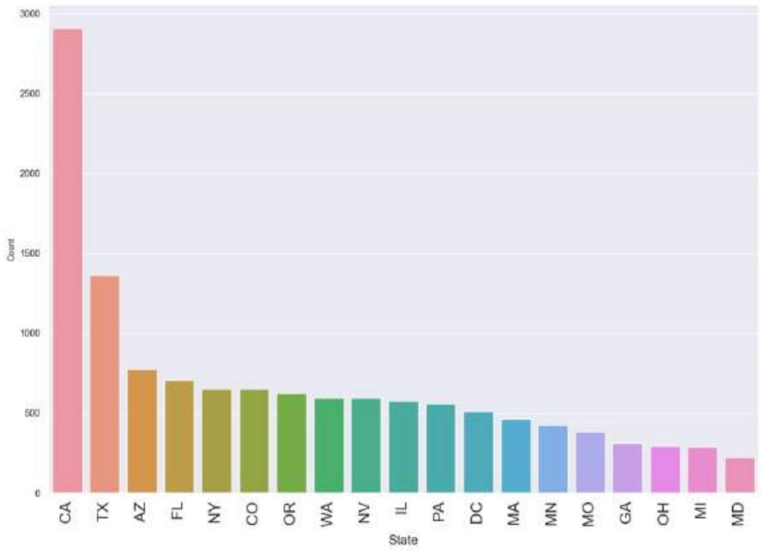


Fig-7.9 : States effected based on their population

Screenshot 5.8: States effected based on their population

5.9 DATA VISUALIZATION BASED ON PEOPLE EFFECTED THE MOST

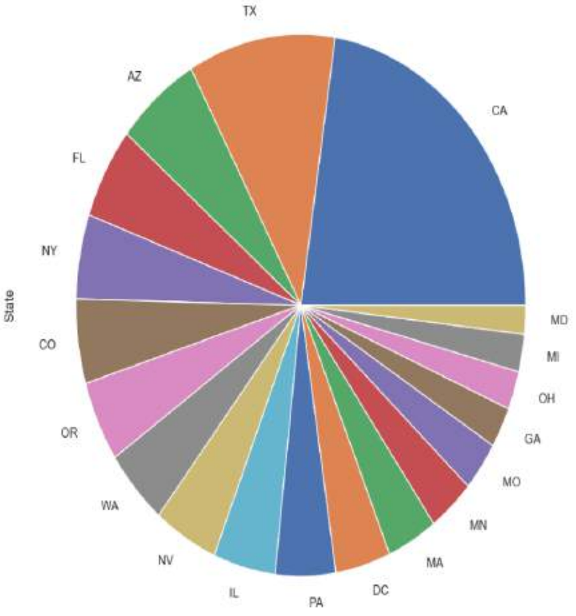


Fig-8 : Visualization of data state wise based on people effected the most

Screenshot 5.9: Visualization of data state wise based on people effected the most

5.10 DISEASE CATEGORIZATION BASED ON GENDER

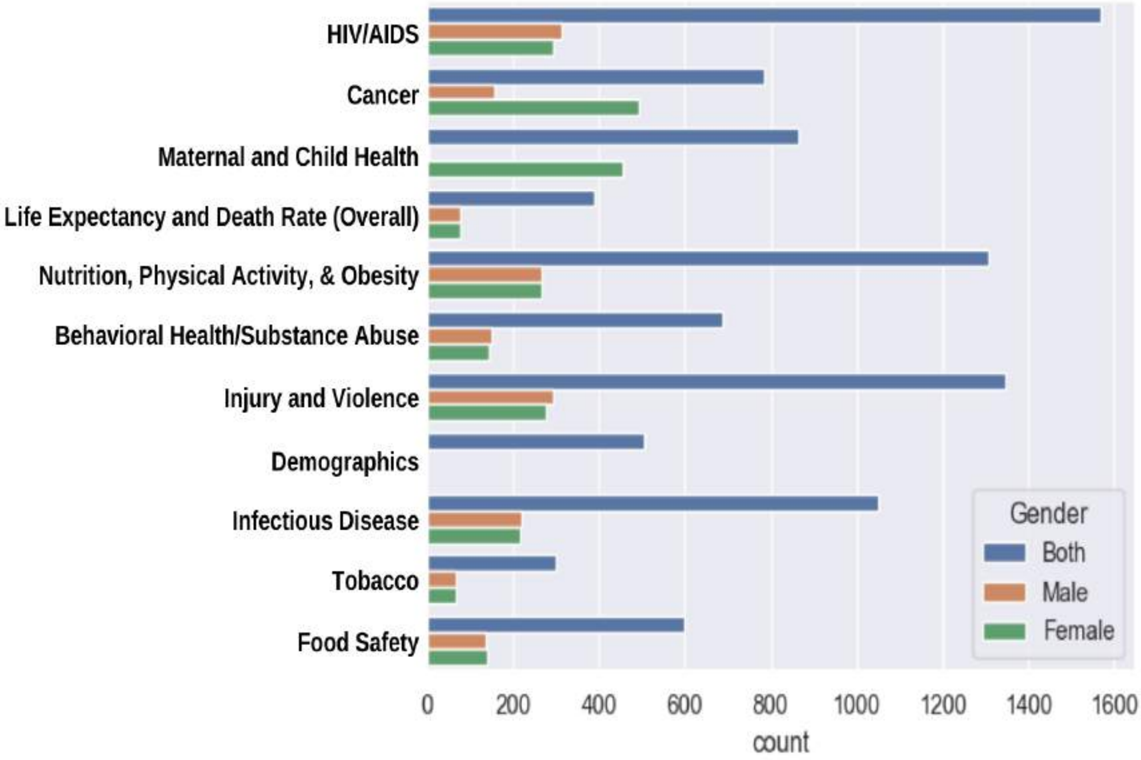


Fig-8.1 : Disease Categorization based on gender

Screenshot 5.10: Disease Categorization based on gender

5.11 DATA REPRESENTATION BASED ON VARIANTS

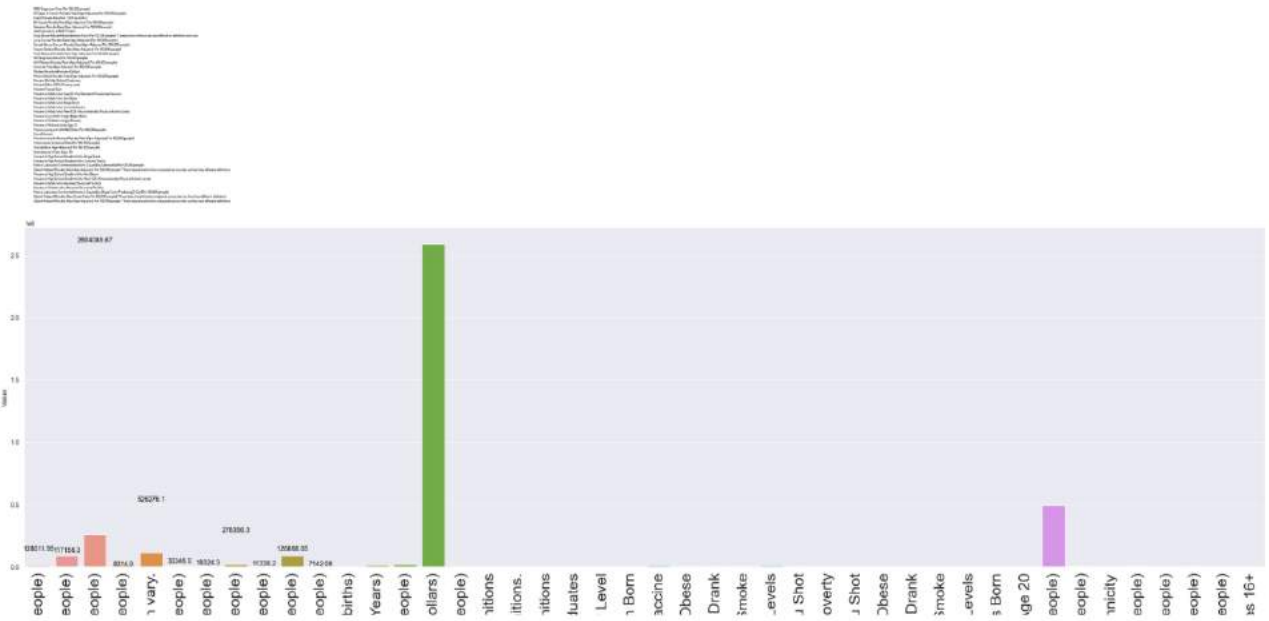


Fig-8.2 : Data representation based on variants..

Screenshot 5.10: Data representation based on variants

6. TESTING

6. TESTING

6.1 INTRODUCTION TO TESTING

Testing is a procedure, which uncovers blunders in the program. Programming testing is a basic component of programming quality affirmation and speaks to a definitive audit of determination, outline and coding. The expanding perceivability of programming as a framework component and chaperon costs related with a product disappointment are propelling variables for we arranged, through testing. Testing is the way toward executing a program with the plan of finding a mistake. The plan of tests for programming and other built items can be as trying as the underlying outline of the item itself It is the significant quality measure utilized amid programming improvement. Amid testing, the program is executed with an arrangement of experiments and the yield of the program for the experiments is assessed to decide whether the program is executing as it is relied upon to perform.

6.2 TESTING STRATEGIES

A technique for programming testing coordinates the outline of programming experiments into an all around arranged arrangement of steps that outcome in fruitful improvement of the product. The procedure gives a guide that portrays the means to be taken, when, and how much exertion, time, and assets will be required. The procedure joins test arranging, experiment configuration, test execution, and test outcome gathering and assessment. The procedure gives direction to the specialist and an arrangement of points of reference for the chief. Due to time weights, advance must be quantifiable and issues must surface as ahead of schedule as would be prudent

Keeping in mind the end goal to ensure that the framework does not have blunders, the distinctive levels of testing techniques that are connected at varying periods of programming improvement are:

6.2.1 UNIT TESTING

Unit Testing is done on singular modules as they are finished and turned out to be executable. It is restricted just to the planner's prerequisites. It centers testing around the capacity or programming module. It Concentrates on the interior preparing rationale and information structures. It is rearranged when a module is composed with high union

- Reduces the quantity of experiments
- Allows mistakes to be all the more effectively anticipated and revealed**Black**

6.2.2 BOX TESTING

It is otherwise called Functional testing. A product testing strategy whereby the inward workings of the thing being tried are not known by the analyzer. For instance, in a discovery test on a product outline the analyzer just knows the information sources and what the normal results ought to be and not how the program touches base at those yields. The analyzer does not ever inspect the programming code and does not require any further learning of the program other than its determinations. In this system some experiments are produced as information conditions that completely execute every single practical prerequisite for the program. This testing has been utilizations to discover mistakes in the accompanying classifications:

- Incorrect or missing capacities
- Interface blunders
- Errors in information structure or outside database get to
- Performance blunders
- Initialization and end blunders.

In this testing just the yield is checked for rightness.

6.2.3 WHITE BOX TESTING

It is otherwise called Glass box, Structural, Clear box and Open box testing . A product testing procedure whereby express learning of the inner workings of the thing being tried are utilized to choose the test information. Not at all like discovery testing, white box testing utilizes particular learning of programming code to inspect yields. The test is precise just if the analyzer comprehends what the program should do. He or she would then be able to check whether the program veers from its expected objective. White box testing does not represent blunders caused by oversight, and all obvious code should likewise be discernable. For an entire programming examination, both white box and discovery tests are required.

In this the experiments are produced on the rationale of every module by drawing stream diagrams of that module and sensible choices are tried on every one of the cases. It has been utilizations to produce the experiments in the accompanying cases:

- Guarantee that every single free way have been Executed.
- Execute every single intelligent choice on their actual and false Sides.

6.2.4 INTEGRATION TESTING

Coordination testing guarantees that product and subsystems cooperate an entirety. It tests the interface of the considerable number of modules to ensure that the modules carry on legitimately when coordinated together. It is characterized as a deliberate procedure for developing the product engineering. In the meantime reconciliation is happening, lead tests to reveal blunders related with interfaces. Its Objective is to take unit tried modules and assemble a program structure in view of the recommended outline

Two Approaches of Integration Testing

- Non-incremental Integration Testing
- Incremental Integration Testing

6.2.5 SYSTEM TESTING

Framework testing includes in-house testing of the whole framework before conveyance to the client. Its point is to fulfill the client the framework meets all necessities of the customer's determinations. This testing assesses working of framework from client perspective, with the assistance of particular report. It doesn't require any inward learning of framework like plan or structure of code.

It contains utilitarian and non-useful zones of utilization/item. Framework Testing is known as a super arrangement of a wide range of testing as all the significant sorts of testing are shrouded in it. In spite of the fact that attention on sorts of testing may differ on the premise of item, association procedures, course of events and necessities. Framework Testing is the start of genuine testing where you test an item all in all and not a module/highlight.

6.2.6 ACCEPTANCE TESTING

Acknowledgment testing, a testing method performed to decide if the product framework has met the prerequisite particulars. The principle motivation behind this test is to assess the framework's consistence with the business necessities and check in the event that it is has met the required criteria for conveyance to end clients. It is a pre-conveyance testing in which whole framework is tried at customer's site on genuine information to discover blunders. The acknowledgment test bodies of evidence are executed against the test information or utilizing an acknowledgment test content and afterward the outcomes are contrasted and the normal ones.

The acknowledgment test exercises are completed in stages. Right off the bat, the essential tests are executed, and if the test outcomes are palatable then the execution of more intricate situations are done.

6.3 TEST APPROACH

A Test approach is the test system usage of a venture, characterizes how testing would be done. The decision of test methodologies or test technique is a standout amongst the most intense factor in the achievement of the test exertion and the precision of the test designs and gauges.

Testing should be possible in two ways

- Bottom up approach
- Top down approach

Bottom up Approach

Testing can be performed beginning from littlest and most reduced level modules and continuing each one in turn. In this approach testing is directed from sub module to primary module, if the fundamental module is not built up a transitory program called DRIVERS is utilized to recreate the principle module. At the point when base level modules are tried consideration swings to those on the following level that utilization the lower level ones they are tried exclusively and afterward connected with the already inspected bring down level modules

Top down Approach

In this approach testing is directed from fundamental module to sub module. in the event that the sub module is not built up an impermanent program called STUB is utilized for mimic the sub module. This sort of testing begins from upper level modules. Since the nitty gritty exercises more often than not performed in the lower level schedules are not given stubs are composed. A stub is a module shell called by upper level module and that when achieved legitimately will restore a message to the calling module demonstrating that appropriate association happened.

6.4 VALIDATION

The way toward assessing programming amid the improvement procedure or toward the finish of the advancement procedure to decide if it fulfills determined business prerequisites. Approval Testing guarantees that the item really addresses the customer's issues. It can likewise be characterized as to exhibit that the item satisfies its proposed utilize when sent on proper condition.

The framework has been tried and actualized effectively and along these lines guaranteed that every one of the prerequisites as recorded in the product necessities determination are totally satisfied.

6.5 TEST CASES

Experiments include an arrangement of steps, conditions and sources of info that can be utilized while performing testing undertakings. The principle expectation of this action is to guarantee whether a product passes or bombs as far as usefulness and different perspectives. The way toward creating experiments can likewise help discover issues in the prerequisites or plan of an application. Experiment goes about as the beginning stage for the test execution, and in the wake of applying an arrangement of information esteems, the application has a conclusive result and leaves the framework at some end point or otherwise called execution post condition.

Table 6.5.1 Test Cases

Test case ID	Test case name	Purpose	Input	Output
1	Disease categorization	To check the disease categories	User data	Categorization of diseases
2	Health Prediction	To predict the health	User data	Data visualization based on years where people effected most
3	Health Prediction	To predict the health	User data	People effected according to disease Male and female data visualization, Male and female pie data visualization State effected based on their results, States effected based on their population Visualization of data state wise based on people effected the most Disease categorization based on gender Data representation based on variants

7. CONCLUSION

7. CONCLUSION & FUTURE SCOPE

Machine learning (ML) techniques are crucial in different business fields. Healthcare field facing more problems and it is becoming more expensive. Several ML techniques are used to rectify them. This paper presents various ML techniques for prediction of various diseases like heart disease, breast cancer, diabetic disease and thyroid disease. From the earlier study, it is recognized that naive Bayes provides 86% of accuracy for the diagnosis of heart disease. SVM gives 96.40% of accuracy for the breast cancer diagnosis, and CART provides 79% of accuracy for the detection of diabetic disease. In future, we are trying to improve the accuracy of breast cancer prediction by using different machine learning algorithms.

8. BIBILOGRAPHY

8. BIBILOGRAPHY

8.1 REFERENCES

[1] Bouwens J. : Embracing Change: The healthcare industry focuses on new growth drivers and leadership requirements

[2] Pianin E.: US Health Care Costs Surge to 17 Percent of GDP. The Financial Times

[3] Brown B. : Top 7 Healthcare Trends and Challenges from Our Financial Expert

[4] Nambiar R., Sethi A., Bhardwaj R., Vargheese R. : A look at challenges and opportunities of Big Data analytics in healthcare. IEEE Big Data Conference 2013

[5]: Kayyali B., Knott D., Kuiken S. : The Big Data Revolution in US Health Care: Accelerating value and innovation. McKinsey & Company

[6] SAS Institute: Big Data: What it is and why it matters? [7] Bhardwaj R., Adhiraaj Sethi A., Nambiar R. : Big data in genomics: An overview. IEE Big Data Conference 2014

[8] Applod K. : Five big industry changes to watch in 2016

[9] Rickert J. : Patient-Centered Care: What It Means And How To Get There

[10] Daveport T. : Industrial-Strength Analytics with Machine Learning. The Wall Street Journal

[11] SAS Institute: Machine Learning: What it is and what it matters

[12] Wikipedia: Machine Learning

[13] Maddux D. : The Human Condition in Structured and Unstructured Data. Acumen Physician Solutions

[14] Brownlee J. : What is Machine Learning: A Tour of Authoritative Definitions and a Handy One-Liner You Can Use. www.machinelearningmastery.com

[15] Dolley S. : Big Data Solution to Harnessing Unstructured Data in Healthcare.

[16] Hauskrecht M., Visweswaran S., Cooper G., Clermont G.: Clinical Alerting of Unusual Care that Is Based on Machine Learning from Past EMR Data

[17] Page D. : Challenges in Machine Learning from Electronic Health Records. MLHC 2015

[18] Dolley S. : Big Data Solution to Harnessing Unstructured Data in Healthcare. www.cloudera.com

[19] Shah Lab Website: https://shahlab.stanford.edu/

[20] CB Insights: From Virtual Nurses to Drug Discovery: 106 Artificial Intelligence Startups in Healthcare

[21] BenevolentAI website: www.benevolentai.com

[22] AliveCor Website: www.alivecor.com.com

[23] Ginger.io Website: www.ginger.io

[24] Berg Health Website: https://www.berghealth.com

[25] Krol A. : Berg and the Pursuit of the Body's Hidden Drugs

[26] Enlitic Website: www.enlitic.com

[27] Fast Company Staff: The World's Top 10 Most Innovative Companies in Health Care

[28] Kohn M. : Real World Data and Clinical Decision Support

[29] Taghizadeh G. : Top 5 Companies Revolutionizing Healthcare with Machine Learning

[30] Huang SH., LePendu P., Iyer SV., Tai-Seale M., Carrell

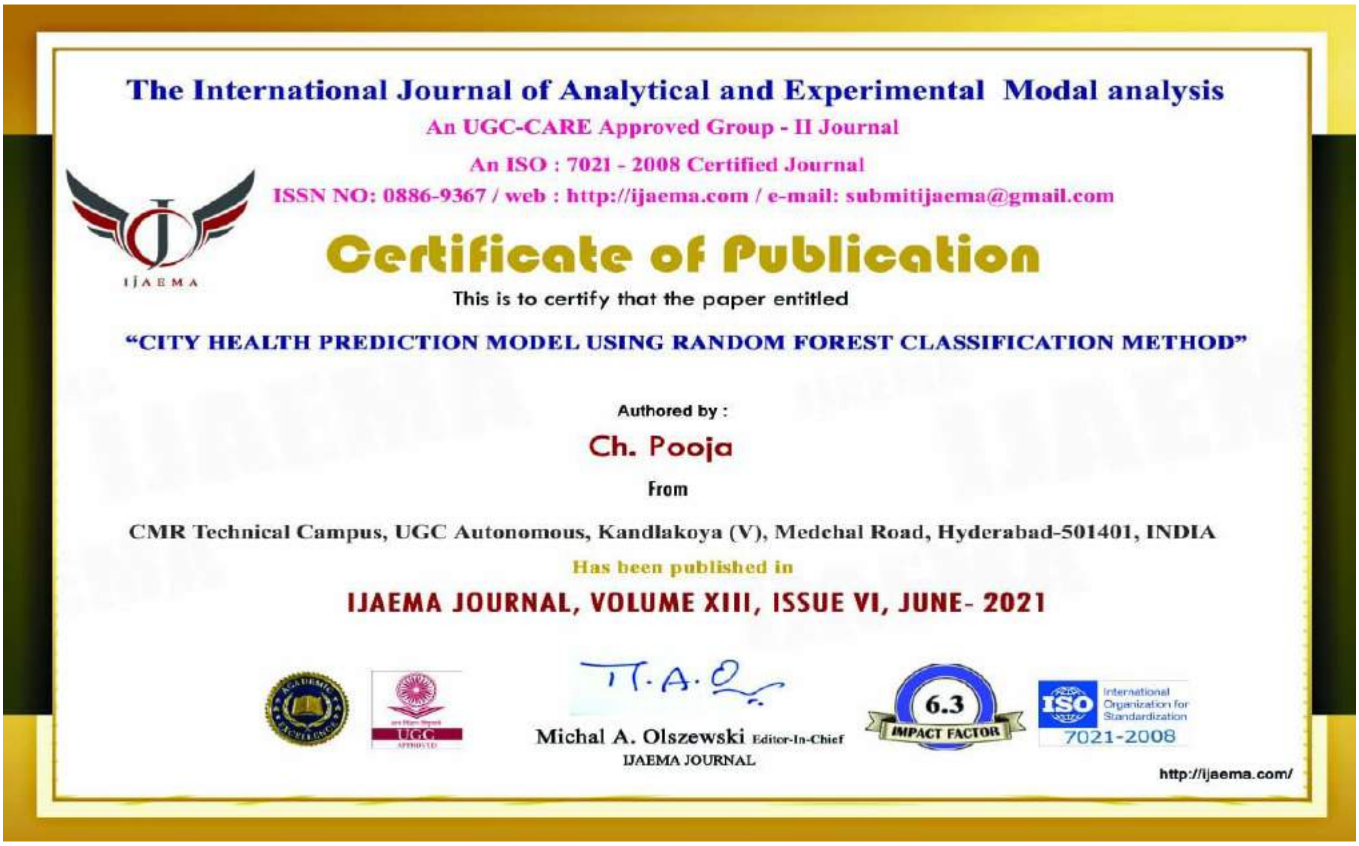
D., Shah NH. : Toward personalizing treatment for depression:

predicting diagnosis and severity.

[31] MIT Technology Review: 50 Smartest Companies 2016

GitHub

https://github.com/snehithkumar-d/City-Health-Prediction-Model-Using-Random-Forest-Classification-Method.git



IJAEMA

The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - II Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

Certificate of Publication

This is to certify that the paper entitled

"CITY HEALTH PREDICTION MODEL USING RANDOM FOREST CLASSIFICATION METHOD"

Authored by :

Ch. Pooja

From

CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Has been published in

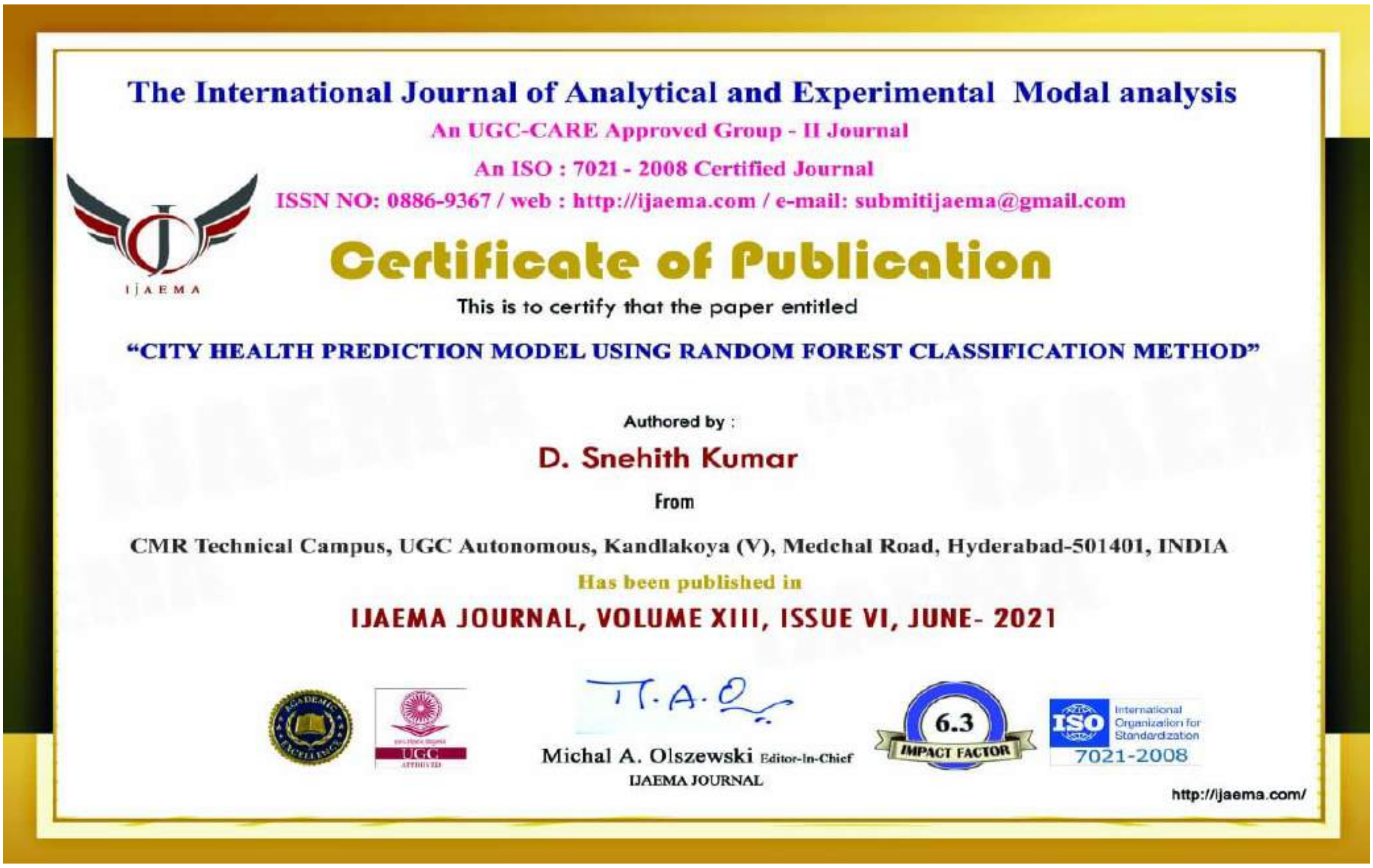
IJAEMA JOURNAL, VOLUME XIII, ISSUE VI, JUNE- 2021

Michal A. Olszewski Editor-In-Chief
IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/



The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - II Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

IJAEMA

Certificate of Publication

This is to certify that the paper entitled

"CITY HEALTH PREDICTION MODEL USING RANDOM FOREST CLASSIFICATION METHOD"

Authored by :

D. Snehith Kumar

From

CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Has been published in

IJAEMA JOURNAL, VOLUME XIII, ISSUE VI, JUNE- 2021

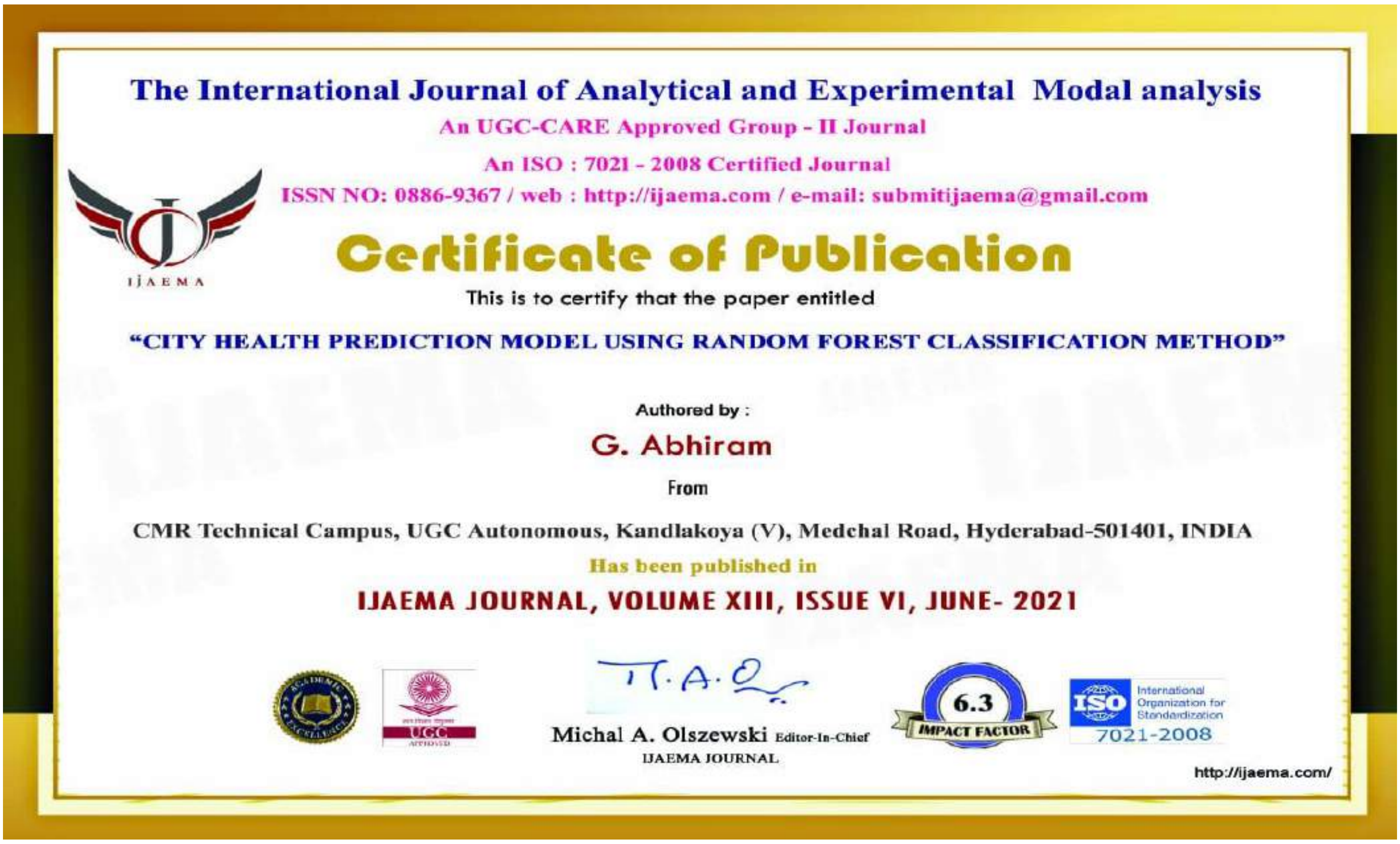
UGC

Michal A. Olszewski Editor-In-Chief
IJAEMA JOURNAL

6.3
IMPACT FACTOR

ISO International Organization for Standardization
7021-2008

http://ijaema.com/



The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - II Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

IJAEMA

Certificate of Publication

This is to certify that the paper entitled

"CITY HEALTH PREDICTION MODEL USING RANDOM FOREST CLASSIFICATION METHOD"

Authored by :

G. Abhiram

From

CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Has been published in

IJAEMA JOURNAL, VOLUME XIII, ISSUE VI, JUNE- 2021

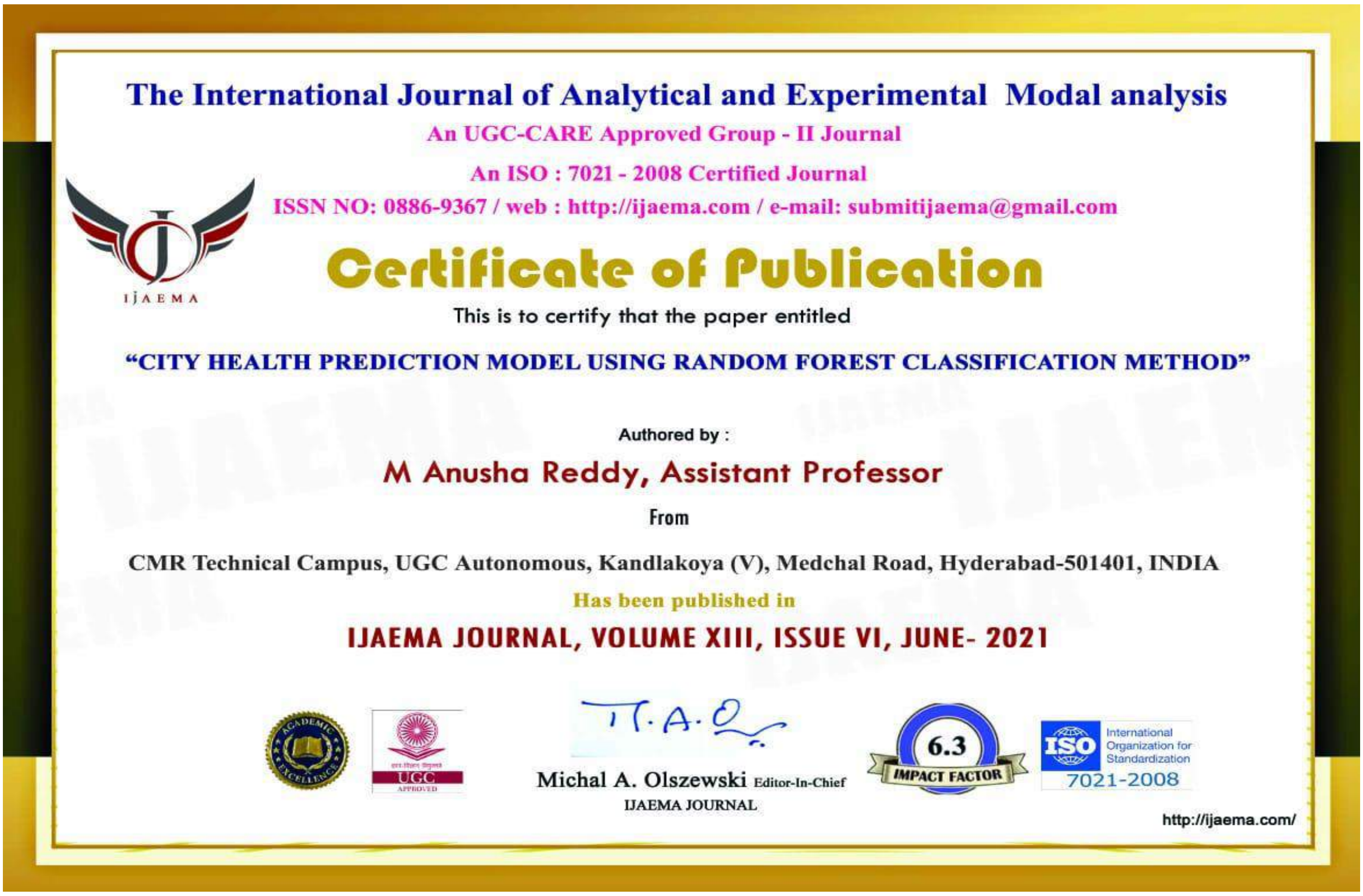
UGC APPROVED

Michal A. Olszewski Editor-In-Chief
IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/



The International Journal of Analytical and Experimental Modal analysis

An UGC-CARE Approved Group - II Journal

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0886-9367 / web : http://ijaema.com / e-mail: submitijaema@gmail.com

IJAEMA

Certificate of Publication

This is to certify that the paper entitled

"CITY HEALTH PREDICTION MODEL USING RANDOM FOREST CLASSIFICATION METHOD"

Authored by :

M Anusha Reddy, Assistant Professor

From

CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Has been published in

IJAEMA JOURNAL, VOLUME XIII, ISSUE VI, JUNE- 2021

ACADEMIC EXCELLENCE

UGC APPROVED

Michal A. Olszewski Editor-In-Chief
IJAEMA JOURNAL

6.3 IMPACT FACTOR

ISO International Organization for Standardization 7021-2008

http://ijaema.com/

CITY HEALTH PREDICTION MODEL USING RANDOM FOREST CLASSIFICATION METHOD

Ch Pooja
UG Scholar, Computer Engineering
Department, CMR Technical Campus,
UGC Autonomous, Kandlakoya (V),
Medchal Road, Hyderabad-501401, INDIA

G Abhiram
UG Scholar, Computer Engineering
Department, CMR Technical Campus,
UGC Autonomous, Kandlakoya (V),
Medchal Road, Hyderabad-501401, INDIA

D Snehith Kumar
UG Scholar, Computer Engineering
Department, CMR Technical Campus,
UGC Autonomous, Kandlakoya (V),
Medchal Road, Hyderabad-501401, INDIA

M Anusha Reddy
Assistant Professor, CMR Technical
Campus, UGC Autonomous, Kandlakoya
(V), Medchal Road, Hyderabad-501401,
INDIA

Abstract:

City Health Office in Indonesia is creating a health report every year, describing the condition of the city public health. The report is used as the source of determining the city health index. The construction of a city health development index is important to produce an objective formula. In this study, the classification method Random Forest is used to developing a proper model for prediction and analysis of the health index of a city. The goal of this work is to find a prediction model to make a more accurate prediction and reducing errors in dealing with the city health index. The performance of the model is evaluated by using three parameters: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The research shows that the model of Random Forest with a 15 percent data test by using 200 decision trees gives the best results.

Keywords—*City Health, Public Health, Random Forest, Classification, Mean Absolute Error.*

I Introduction

Healthcare is one of the fastest growing sectors today and is currently in the core of a complete global overhaul and transformation. Russell Reynolds and Associates cites that global healthcare costs, currently estimated at $6 trillion to $7 trillion, are projected to reach more than $12 trillion within just seven years. This trend is also exemplified domestically, in the United States. The total spending on healthcare in the United States increased up to 5.3 percent and has topped $3 trillion nationwide. Additionally, healthcare spending in the United States represents 17 percent of the total gross domestic product (GDP); our healthcare costs are rising at rates close to double of our economic growth rate. In addition to a rise in the amount consumers are spending on healthcare, the federal government has been forced to pay more and more for healthcare as costs become too high for patients to afford. The amount of money the federal government has allocated for healthcare spending has increased by 11.7 percent in 2014 to an incredible $844 billion in 2015. This rise in federal funding represents the significant disparity between the cost of healthcare and the financial burden on consumers. Given this rapid growth in costs, a number of actions must be taken to ensure the costs of healthcare do not further spiral out of control. The need for patient-physician communication, follow-up appointments, and the availability of specialists have also become painfully apparent. Innovation and technological solutions may be the solution to fix the issues with our modern-day healthcare system. These innovations range from swallowable microchips that alert doctors when medication has been taken to large scale data analysis to determine which medications are most effective. However, recently, machine learning has been identified as having major technological application in the healthcare realm. While such technologies will probably never completely replace physicians, they can transform the healthcare sector, benefiting both patients and providers.

The field of medicine has taken significant strides in its advancement; the development of vaccinations, antibiotics, and even the concept of sterilization have disrupted the industry and caused a cascade-like effect on all patients and doctors involved in healthcare. Needless to say, the human population has progressed a great degree from past medical care. The healthcare industry is made up of preventive, diagnostic, remedial, and therapeutic sectors. Each of these sectors work together to provide a comprehensive, holistic experience for the modern-day patients.

II. LITERATURE SURVEY

BenevolentAI: BenevolentAI was founded in 2013 by Ken Mulvany founder of Proximagen. BenevolentBio is focused on applying technology in the bioscience industries. The initial focus has been on human health – generating new ideas that have the potential to improve the lives of millions and deliver better medicines to patients faster in currently overlooked areas such as orphan disease and rare cancers. BenevolentTech is developing an advanced artificial intelligence platform that helps scientists make new discoveries and redefines how scientists gain access to, and use, all the data available to drive innovation. The technology is built upon a deep judgement system that learns and reasons from the interaction between human reasoning and data. Butterfly Network: Butterfly Network is transforming diagnostic and therapeutic imaging with devices, deep learning, and the cloud. Butterfly Network operates at the intersection of engineering and medicine by bringing together world-class talent in computer science, physics, mechanical engineering, electrical engineering and medicine.

Digital Reasoning Systems: Digital Reasoning is a global leader in using artificial intelligence to understand human communications. Its cognitive computing platform, Synthesys, automates key tasks and uncovers transformative insights across vast amounts of human communications for many of the world’s most elite companies, organizations, and agencies.

Flatiron Health: Flatiron Health is a health care technology company and operator of the OncologyCloud platform. Integrating across the entire clinical data spectrum, Flatiron Health allows cancer care providers and life science companies to gain deep business and clinical intelligence through its web-based platform.

H2O.ai: H2O is a provider of an open source based predictive analytics platform for data scientists and application developers who need scalable and fast machine learning for smart business applications. These applications include smart home appliances, self-driving cars, personalized digital content, smart assistants, and others. Pathway Genomics: Pathway Genomics, founded in 2008, provides physicians and their patients with accurate genetic information to improve or maintain health and wellness. The company's mobile health applications merge artificial intelligence and deep learning with personal genetic information that provides personalized health and wellness guidance.

WellTok: WellTok combines knowledge of the healthcare industry and social networking technology in its CafeWell.com channel to achieve levels of consumer engagement for healthcare population managers through Social Health Management. WellTok's software/Internet products focus on providing a complete, integrated.

III. PROPOSED METHODOLOGY

A model aimed at predicting the adoption of technology in the health care sector has had a tremendously positive impact on medical processes along with the practices in which health care professionals engage. Many attempts have been by researchers in the past for predicting flight delays using Machine Learning, Deep Learning and Big Data approaches. Kalliguddi(author) constructed regression models like Decision Tree Regressor, Random Forest regressor on flight data for predicting both departure and arrival delays. The main issues is to find the error rate in terms od predictions and reducting the error factor in the model. Predictive algorithms enable computers to recognize patterns in data and draw deductions from those patterns that show the likelihood of particular events occurring in the future. This kind of algorithm is used in many types of activities, ranging from detection of credit card fraud and the optimization of search engines to stock market analysis and speech recognition. Health forecasting can be translated into effective interventions with individual patients, the analytic tools will be useless. So healthcare organizations must develop the infrastructure and the culture required to turn the data into action. That infrastructure must provide the ability to generate timely reports and use automation tools to apply intervention strategies across a patient population. Improving efficiencies for operational management of health care business operations. Accuracy of diagnosis and treatment in personal medicine Increased insights to enhance cohort treatment The current interest in predictive modeling is part of a larger trend to employ business and clinical intelligence (B&CI) applications in healthcare. For optimal use in chronic disease management, predictive analytics should be applied to longitudinal rather than episodic data.

This requires getting patients involved. For example, patients might be asked to fill out online functional status surveys at regular intervals.

In select healthcare settings, remote monitoring data may also be routinely available.

Project Planning:

Step1: Data set collection.

Step2: Feature Extraction

Step3: Data processing and corelation between the features

Step4: apply regression.

Step5: Predict the results and evlation of Error factors.

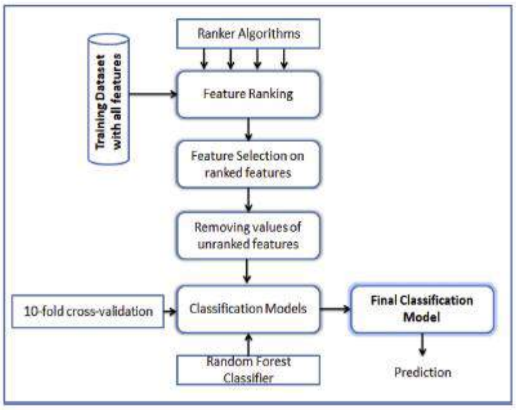

Ranker Algorithms
Training Dataset with all features
Feature Ranking
Feature Selection on ranked features
Removing values of unranked features
10-fold cross-validation
Classification Models
Final Classification Model
Random Forest Classifier
Prediction

Fig 1: Project Architecture

Data and methodology:

The data is obtained from Copernicus website which is a free database. It has satellite images from all over the world. It updates regularly and images from a certain date can be obtained. The image obtained is then further processed to get the desired data. Ground Truth data is referred to as the data that is collected on field by observation. In this step, data is collected by visiting some fields that have known objects. These fields may include vegetation fields, buildings, roads, rivers etc.

Processing:

After stacking the image is now ready to be processed. The processing takes place in a software called ENVI. First the ground truth Regions Of Interest are loaded on the stacked image. The image having Ground truth ROIs loaded over it is shown in figure-1. When the Ground truth ROIs appear on the image, a mask is made in order to select only the desired image from the empty background. After making mask, the desired machine learning algorithms are run on the image. In this project we have used both supervised and unsupervised algorithms. The supervised algorithms include neural network and support vector machine while unsupervised algorithm includes K-means clustering. These particular algorithms are selected because they are the most popular and widely used algorithms that are present for use in the ENVI software. Secondly it was aimed to compare one of the most popular algorithm from unsupervised classification to two most popular supervised classification algorithms

Post processing:

After the data is processed and image is classified then begins the phase of post processing. In post processing we make confusion matrix. The confusion matrix gives the accuracy and statistics of the result.

IV RESULT ANALYSIS

After preprocessing and feature extraction of our dataset, 80% of the dataset was selected for training and 20% of the dataset was selected for testing. For error calculation, we are using scikit-learn metrics.

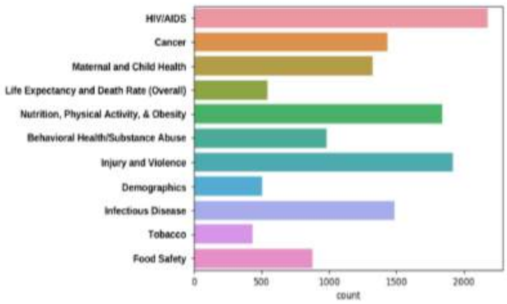


Fig 2: Disease Categorization

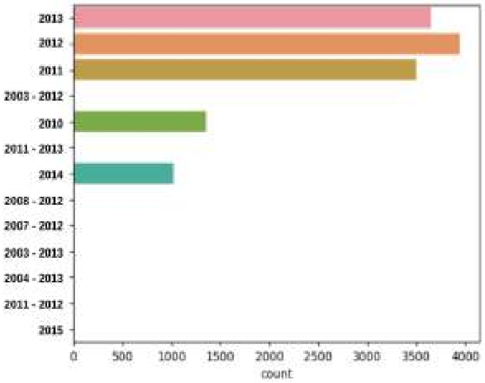


Fig 3: Visualization of data based on the years where people effected most

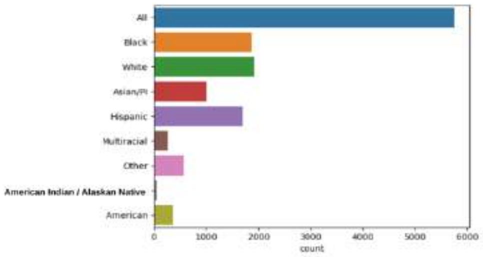


Fig 4 : Race/Ethnicity

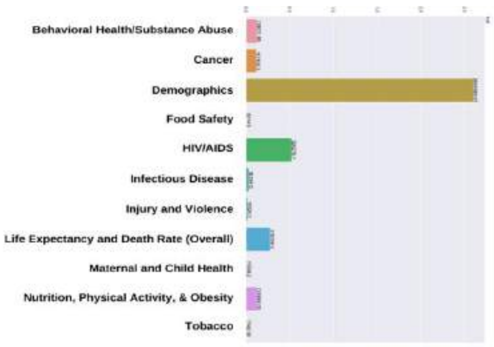


Fig 5 : People effected according to diseases

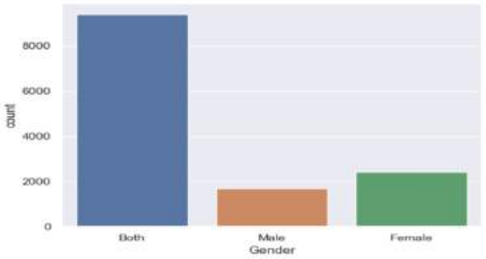


Fig 6 : Male and Female data visualization

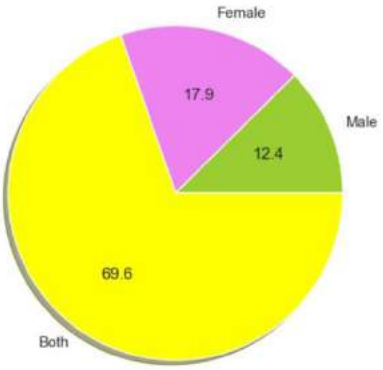


Fig 7 :Male and Female data pie visualization

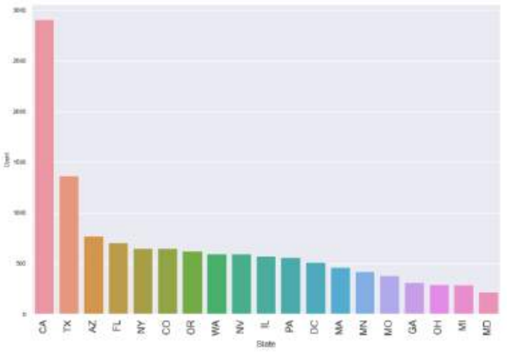


Fig 8 :States effected based on their population

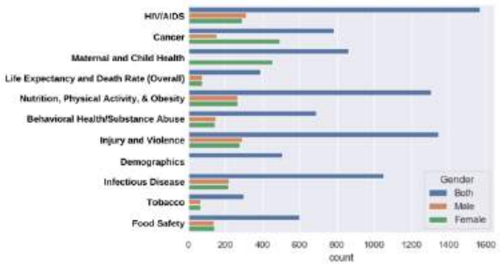


Fig 9 : Disease Categorization based on gender

V CONCLUSION

Machine learning (ML) techniques are crucial in different business fields. Healthcare field facing more problems and it is becoming more expensive. Several ML techniques are used to rectify them. This paper presents various ML techniques for prediction of various diseases like heart disease, breast cancer, diabetic disease and thyroid disease. From the earlier study, it is recognized that naive Bayes provides 86% of accuracy for the diagnosis of heart disease. SVM gives 96.40% of accuracy for the breast cancer diagnosis, and CART provides 79% of accuracy for the detection of diabetic disease. In future, we are trying to improve the accuracy of breast cancer prediction by using different machine learning algorithms.

VI REFERENCES

[1] Bouwens J. : Embracing Change: The healthcare industry focuses on new growth drivers and leadership requirements

[2] Pianin E.: US Health Care Costs Surge to 17 Percent of GDP. The Financial Times

[3] Brown B. : Top 7 Healthcare Trends and Challenges from Our Financial Expert

[4] Nambiar R., Sethi A., Bhardwaj R., Vargheese R. : A look at challenges and opportunities of Big Data analytics in healthcare. IEEE Big Data Conference 2013

[5]: Prasadu Peddi (2019), "AN EFFICIENT ANALYSIS OF STOCKS DATA USING MapReduce", ISSN: 1320-0682, Vol 6, issue 1, pp:22-34.

[6] SAS Institute: Big Data: What it is and why it matters?

[7] Bhardwaj R., Adhiraaj Sethi A., Nambiar R. : Big data in genomics: An overview. IEE Big Data Conference 2014

[8] Prasadu Peddi (2020), "Public auditing mechanism to verify data integrity in cloud storage", International Journal of Emerging Trends in Engineering Research, vol 8, Issue 9, Pages 5220–5225.

[9] Rickert J. : Patient-Centered Care: What It Means And How To Get There

[10] Daveport T. : Industrial-Strength Analytics with Machine Learning. The Wall Street Journal

[11] SAS Institute: Machine Learning: What it is and what it matters

[12] Wikipedia: Machine Learning

[13] Maddux D. : The Human Condition in Structured and Unstructured Data. Acumen Physician Solutions

[14] Brownlee J. : What is Machine Learning: A Tour of Authoritative Definitions and a Handy One-Liner You Can Use. www.machinelearningmastery.com.

[15] Prasadu Peddi (2017) "Design of Simulators for Job Group Resource Allocation Scheduling In Grid and Cloud Computing Environments", ISSN: 2319-8753 volume 6 issue 8 pp: 17805-17811.