**A**

**Synopsis**

**on**

**Disease Prediction Using Machine learning**

**Submitted by**

**Group. No.**

**AMAN RAGHAVA- 2100290140020**

**SHUBHANGI SINGH- 2100290140133**



**Submitted to**

**Dr. ARUN TRIPATHI**

**Department of Computer Applications,**

**KIET Group of Institutions,**

**Delhi-NCR, Ghaziabad**

INTRODUCTION

The load of cardiovascular diseases is rapidly increasing all

over the world from the past few years. Even if these diseases

has found as the most important source of death, it has been

announced as the most manageable and avoidable disease [1].

Mainly, blockage in arteries causes heart stroke. It occurs when

heart does not pump the blood around the body efficiently.

Having high blood pressure is also one of the main causes of

getting a heart disease. A survey says that, in 2011 to 2014, the

commonness of hypertension in the world was about 35%,

which is also a cause of heart disease. Similarly, there are many

more reasons for getting a heart disease such as obesity, not

taking in proper nutrition, increased cholesterol and lack of

physical activity. So, prevention is very necessary. For

prevention, awareness of heart diseases is important. Around

47% of people dies outside the hospital and it shows that they

don’t act on early warning signs.

Nowadays, lifespan of a human being is reduced because of

heart diseases. So, World Health Organization (WHO)

developed targets for prevention of non-communicable diseases

(NCDs) in 2013, in which, 25% of relative reduction is from

cardiovascular diseases and it is being ensured that at least 50%

of patients with cardiovascular diseases have access to relevant

drugs and medical counselling by 2025 [2]. Around 17.9 million

people died just because of cardiovascular diseases in 2016,

which is 31% of deaths around the world.

A major challenge in heart diseases is its detection [3]. It is

difficult to predict that a person has a heart disease or not. There

are instruments available which can predict heart diseases but

either they are expensive or are not efficient to calculate the

chance of heart disease in human [4]. A survey of World Health

Organization (WHO) says that medical professionals are able

to predict just 67% of heart disease, so there is a vast scope of

research in this field [5]. In case of India, access to good doctors

and hospitals in rural areas is very low. A 2016 WHO report

says that, just 58% of the doctors have medical degree in urban

areas and 19% in rural areas.

In USA, someone has a heart attack every 40 seconds, that is,

more than one person dies in USA due to heart attack. Apart

from this, Turkmenistan have the highest rate of deaths till

2012, with 712 deaths per 100,000 people. Whereas,

Kazakhstan have the second highest rate of deaths due to heart

diseases. India holds 56th position in this series [6]. Study also

shows that, at ages 30-69 years, 1.3 million cardiovascular

deaths, 0.9 million (68.4%) were caused by coronary heart

disease and 0.4 million (28.0 %) by stroke

Heart diseases are a major challenge in medical science,

Machine Learning could be a good choice for predicting any

heart disease in humans [7]. Heart diseases can be predicted

using Neural Network, Decision Tree, KNN, etc. Later in this

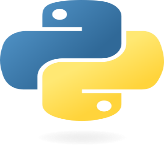
paper, we will see that how Logistic Regression is used to find

the accuracy for heart disease. It also shows that how ML will

help in our future for heart disease.

Technologies / Software Requirements

* Python (3.7 version)
* Flask
* Cross Validation (CV)
* Pandas
* Numpy
* HTML
* CSS
* Scikit-Learn
* Machine Learning



Python (3.7 version)

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore, reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this very simple approach.

Flask

Flask is a web application framework written in Python. It was developed by Armin Ronacher, who led a team of international Python enthusiasts called Poocco. Flask is based on the Werkzeg WSGI toolkit and the Jinja2 template engine. Both are Pocco projects.

Cross Validation (CV)

Suppose you train a model on a given dataset using any specific algorithm. You tried to find the accuracy of the trained model using the same training data and found the accuracy to be 95% or maybe even 100%. Is your model ready for prediction? The answer is no. Why? Because your model has trained itself on the given data, i.e. it knows the data and it has generalized over it very well. But when you try and predict over a new set of data, it’s most likely to give you very bad accuracy, because it has never seen the data before and thus it fails to generalizes well over it. This is the problem of overfitting. To tackle such problem, Cross-validation comes into the picture. Cross-validation is a resampling technique with a basic idea of dividing the training dataset into two parts i.e. train and test. On one part(train) you try to train the model and on the second part(test) i.e. the data which is unseen for the model, you make the prediction and check how well your model works on it. If the model works with good accuracy on your test data, it means that the model has not overfitted the training data and can be trusted with the prediction, whereas if it performs with bad accuracy then our model is not to be trusted and we need to tweak our algorithm.

Pandas

**Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named [Numpy](https://www.activestate.com/products/python/python-packages/), which provides support for multi-dimensional arrays.**

Numpy

NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python.

HTML &CSS

[HTML](https://www.w3.org/html/) (the Hypertext Markup Language) and [CSS](https://www.w3.org/Style/CSS/) (Cascading Style Sheets) are two of the core technologies for building Web pages. HTML provides the structure of the page, CSS the (visual and aural) layout, for a variety of devices. Along with [graphics](https://www.w3.org/standards/webdesign/graphics) and [scripting](https://www.w3.org/standards/webdesign/script), HTML and CSS are the basis of building Web pages and Web Applications

Scikit-Learn

**Scikit-learn** (formerly **scikits.learn** and also known as **sklearn**) is a [free software](https://en.wikipedia.org/wiki/Free_software) [machine learning](https://en.wikipedia.org/wiki/Machine_learning) [library](https://en.wikipedia.org/wiki/Library_(computing)) for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) [programming language](https://en.wikipedia.org/wiki/Programming_language).[[3]](https://en.wikipedia.org/wiki/Scikit-learn#cite_note-jmlr-3) It features various [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and [clustering](https://en.wikipedia.org/wiki/Cluster_analysis) algorithms including [support-vector machines](https://en.wikipedia.org/wiki/Support_vector_machine), [random forests](https://en.wikipedia.org/wiki/Random_forests), [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting), [*k*-means](https://en.wikipedia.org/wiki/K-means_clustering) and [DBSCAN](https://en.wikipedia.org/wiki/DBSCAN), and is designed to interoperate with the Python numerical and scientific libraries [NumPy](https://en.wikipedia.org/wiki/NumPy) and [SciPy](https://en.wikipedia.org/wiki/SciPy). Scikit-learn is a [NumFOCUS](https://en.wikipedia.org/w/index.php?title=NumFOCUS&action=edit&redlink=1" \o "NumFOCUS (page does not exist)) fiscally sponsored project.

Machine Learning

The term machine learning was first coined in the 1950s when [Artificial Intelligence pioneer Arthur Samuel](https://en.wikipedia.org/wiki/Arthur_Samuel) built the first self-learning system for playing checkers. He noticed that the more the system played, the better it performed.

Fueled by advances in statistics and computer science, as well as [better datasets](https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/) and the [growth of neural networks](https://eos.org/agu-news/the-rise-of-machine-learning), machine learning has truly taken off in recent years.

Today, whether you realize it or not, [machine learning is everywhere](https://monkeylearn.com/blog/machine-learning-applications/) ‒ automated translation, image recognition, voice search technology, self-driving cars, and beyond.

In this guide, we’ll explain how machine learning works and how you can use it in your business. We’ll also introduce you to [machine learning tools](https://monkeylearn.com/blog/machine-learning-tools/) and show you how to get started with [no-code machine learning](https://monkeylearn.com/blog/no-code-machine-learning/).

MODULE DESCRIPTION

The disease prediction model involves three modules:

 (a) data cleaning,

(b) feature extraction,

(c) classification.

Reports / Outputs

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease.

CONCLUSION

The result for this prediction system  displays a convenient user interface consisting  of details like name, symptoms and the  algorithm that we use to predict as a button and  the results will be predicted based on the  implemented algorithm.

It also displays the accuracy percentage  on which algorithm has the best accuracy so  based on the accuracy of the decision tree,  random forest and naive bayes algorithm  random forest has the better accuracy  percentage of 0.96. It is a best suited algorithm  for this model.

SERVE TO SOCIETY

Nowadays, humans face various diseases due to the current environmental condition and their living habits. The identification and prediction of such diseases at their earlier stages are much important, so as to prevent the extremity of it. It is difficult for doctors to manually identify the diseases accurately most of the time. The goal of this paper is to identify and predict the patients with more common chronic illnesses. This could be achieved by using a cutting-edge machine learning technique to ensure that this categorization reliably identifies persons with chronic diseases. The prediction of diseases is also a challenging task. Hence, data mining plays a critical role in disease prediction. The proposed system offers a broad disease prognosis based on patient’s symptoms by using the machine learning algorithms such as convolutional neural network (CNN) for automatic feature extraction and disease prediction and K-nearest neighbor (KNN) for distance calculation to find the exact match in the data set and the final disease prediction outcome. A collection of disease symptoms has been performed for the preparation of the data set along with the person’s living habits, and details related to doctor consultations are taken into account in this general disease prediction. Finally, a comparative study of the proposed system with various algorithms such as Naïve Bayes, decision tree, and logistic regression has been demonstrated in this paper.

Gantt Chart

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TASK | WEEK  1-2 | WEEK  3-4 | WEEK  5-6 | WEEK  7-8 | WEEK  9-10 | WEEK  11-12 |
| Requeiement &Feasibility Check |  |  |  |  |  |  |
| Designing |  |  |  |  |  |  |
| Coding |  |  |  |  |  |  |
| Testing |  |  |  |  |  |  |

THANK

YOU