DISEASE PREDICTION AND DETECTION (icare)

**IT PROJECT REPORT**

***Submitted in partial fulfillment of th*e *requirements for the award of the degree***

***Of***

**-BACHELOR OF TECHNOLOGY-**

***In***

**CSAI**

***By***

**Jhanvee Khola(03001172020)**

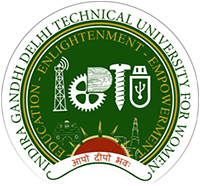
**Khushi Punia(05201172020)**

**Charu Singh(02301172020)**

**Princy Singhal(05101172020)**

***Guided by***

**Dr.Shweta Singhal**



**INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN**

**NEW DELHI – 110006**

**- AUGUST 2021 -**

**CERTIFICATE**

This is to certify that the project entitled **“DISEASE PREDICTION & DETECTION(icare)”** is being submitted at IGDTUW, Delhi for the award of **Bachelor of Technology** in **CSAI** degree. It contains the record of bonafide work carried out by **Jhanvee Khola, Khushi Punia, Charu Singh and Princy Singhal** under my supervision and guidance. It is further certified that the work presented here has reached the standard of B.Tech and to the best of my knowledge has not been submitted anywhere else for the award of any other degree or diploma.

**Dr.Shweta Singhal**

**UNDERTAKING REGARDING ANTI-PLAGIARISM**

We,hereby, declare that the material/ content presented in the report are free from plagiarism and is properly cited and written in my own words. In case, plagiarism is detected at any stage, We would be solely responsible for it.

**ACKNOWLEDGEMENT**

We are grateful to our respectable teacher, Dr.Shweta Singhal, whose insightful leadership and knowledge benefited us to complete this project successfully. Thank you so much for your continuous support and presence whenever needed.

We would also like to thank Mrs.Garima Jaiswal for her advice and contribution to the project and the preparation of this report.

Last but not the least, We would like to thank everyone who is involved in the project directly or indirectly.

**Princy Singhal**

**Khushi Punia**

**Jhanvee Khola**

**Charu Singh**

**DECLARATION**

We, solemnly declare that the project report, **DISEASE PREDICTION AND DETECTION,** is based on my own work carried out during the course of our study under the supervision of **DR.SHWETA SINGHAL**. We assert the statements made and conclusions drawn are an outcome of my research work. We further certify that:

1. The work contained in the report is original and has been done by us under the supervision of our supervisor.
2. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.
3. We have followed the guidelines provided by the university in writing the report.
4. Whenever we have used materials (text, data, theoretical analysis/equations, codes/program, figures, tables, pictures, text etc.) from other sources, we have given due credit to them in the report and have also given their details in the references.

**Our team**

**ABSTRACT**

|  |
| --- |
| * **Problem statement-**   Due to modern lifestyle, diseases are increasing at an alarming rate. Our lifestyle, work type and food habits are the various factors which influence our health causing heart diseases and other health issues like Hypertension, Diabetes etc. In emerging countries like India, the urban population faces the wrath of it with 5.8 million deaths per year.   * **Base of the project**   Data science and AI systems are used for prediction and detection of lethal diseases which will in turn help to take early actions against them.  Data mining technique is one of the most leading research areas in healthcare due to the high importance of valuable data. It has various applications in the healthcare field through a variety of algorithms that we have utilized for disease prediction.  Through Collected data and Medical data mining we wish to integrate it to form a hospital information system. This new technology enables the innovation of trends and predictive patterns in data along with the creation and testing of hypotheses by generating insight-provoking visualizations.   * **Design of the project** * Searching for the diseases and Collecting datasets:   The datasets for training the machine learning models is collected from authentic sources and is selected so that it has the appropriate attributes with accurate patient data.  We collected data regarding patient information for diseases such as diabetes, coronary heart disease and breast cancer as input through a website.   * Data Pre processing: Data pre processing is a data mining technique which is used to transform the raw data in a useful and efficient format. These data are loaded and checked and missing values are converted to null. True and false value is replaced to 1 and 0, respectively. Data having string values are converted to numeric form. * Data Visualization: The data is plotted graphically to have a better understanding of the data.  Algorithms: The concept of data mining helps the end users to extract useful information from large databases. Machine learning algorithms are useful in prediction of data using sample datasets provided. It can also be helpful in providing vital statistics, real-time data, and advanced analytics in terms of the patient’s disease, lab test results, blood pressure, family history, clinical trial data, and more to doctors.  Algorithms used in the project:   * Naive Bayes classification algorithm * Linear Regression * Random forest algorithm * Decision Tree classifier * Support Vector Machine(SVM) * Performance Analysis * **The Overview**   A website is made to take input from users regrading various factors that can be contributing to cause that disease.  Already existing dataset is trained using machine learning algorithms to predict whether a person has chances to contracting the disease. Exploratory data analysis is carried out in order to develop the machine learning algorithms. Finally the results would be shown. Finally the results are shown on the website along with some blogs and health suggestions. |

**INDEX**

Certificate ……………………………………………………………………….... 1

Undertaking regarding anti plagiarism ……………………………………...…..... 2

Acknowledgement ………………………………………………………....……... 3

Declaration ……………………………………………………………………....... 4

Summary/Abstract...............………………………………………………………5

1. **Introduction**
   1. Literature review …………………………………………………..... #
   2. Scope of the project.……………………………. #
   3. Design …………………………………………………... #
2. **Conclusion** ………………………………………………………………… #
3. **Future**
4. **Scope** ………………………………………………………...……...#
5. **Bibliography** ………………………..........................................................#
6. **Contribution** ………………………..........................................................#

**INTRODUCTION**

# **Literature Survey/ Work done so far**

# Research Papers and Surveys done so far-

* ”Disease prediction using principal of component analysis” in  2016 International Conference on Global Trends in Signal Processing,
* ”Disease Prediction in Data Mining Technique – A Survey” by S.Vijiyarani and S.Sudha
* A Survey on Disease Prediction by Machine Learning over Big Data from Healthcare Communities

International Journals-

* ”International Journal of Computer Applications & Information Technology”
* [2017 International Conference On Smart Technologies For Smart Nation (SmartTechCo](https://ieeexplore.ieee.org/xpl/conhome/8356464/proceeding)n)
* IOSR Journal of Engineering (IOSRJEN)

# Applications made-

* A smartphone app called Healthians can provide smart reports from user inputs and helps you reach required treatment goals.
* Smart Health Care App: An Android App to detect disease based on symptoms.
* A smartphone application mfit predicts whether you are likely to have a disease like PCOD etc.
* **Scope of the work**
* The implementation of artificial intelligence provides benefits in early detection by being able to pinpoint any risk alerts a patient may have. This alert will allow the practitioners to get patients to a MRI/CT scan sooner for a disease evaluation.
* It has become important to discover hidden patterns and relationships from medical databases. In classical clinical diagnosis, it requires lots of tests which could complicate the disease prediction. Hence the data mining techniques can help medical expertise to make the decision about the disease using computer aided decision support systems.
* In a study, the early detection alert provided 87.6% accuracy in a diagnosis and prognosis evaluation. That said, the practitioners will be able to implement treatment sooner and predict whether the patient will have a higher possibility of future disease similar to that. Previously, it was challenging for healthcare professionals to collect and analyse the huge volume of data for effective predictions and treatments since there were no technologies or tools available. Now, with machine learning, it’s been relatively easy, as big data technologies are mature enough for wide-scale adoption.
* For example, machine learning can be used in 48-hour post stroke patients gaining a perdition accuracy of 70% whether the patient may have another stroke or not. Many people's lives are cut short due to cancer. However, due to the age of big data we are able to combat this malicious disease.
* Also, patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhalation of harmful gases, intake of contaminated food, pickles and drugs. So, patient datasets can be used to evaluate prediction algorithms in an effort to reduce burden on doctors.
* **Explanation of various aspects**

The main focus is on using different algorithms and combination of several target attributes for different types of disease prediction using data mining i.e. The process of identifying commercially useful patterns or relationships in databases or other computer repositories through the use of advanced statistical tools.It is defined as sifting through very large amounts of data for useful information.

Various datasets are used to predict/detect whether a patient is likely to get a disease based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient and this data is then acted upon with important techniques like categorization and preprocessing to extract necessary information.

***Here are the main steps involved in the making of the prediction and detection system :-***

1. Data collection: Datasets with numerous attributes and patient data, containing valuable information regarding a disease, are collected through authentic sources. These datasets give a number of variables along with a target condition of having or not having a particular disease. The data is investigated using ML explainability tools and techniques.

2.  Data Analysis and Cleaning: This is in general looking at the data to

figure out what's going on. Inspect the data: Check whether there is

any missing data, irrelevant data and do a clean up.

1. Data Visualization: Plotting various functions in the dataset to visualise and understand the data better.
2. Feature selection: Selecting necessary features for classification or constructing new attributes from the given set of attributes to help the mining process.
3. Searching for trends, relations & correlations: Finding common trends or relations between different attributes. Combining statistical analysis, ML & database technology to extract hidden patterns and relationships from

Databases.

1. Drawing final inference: Using the extracted and filtered information, predict whether the patient can be identified to be having the mentioned disease or not.

1. Performance Analysis: Analysing and then enhancing the accuracy of the trained model.

**CONCLUSION**

Summarizing the report here, we can say that machine learning algorithms can go a long way in achieving great heights in healthcare sector.

It is the need of the hour that we finally use Data visualization and Data mining techniques to effectively view large datasets and thus draw conclusion from it. Various attributes can be summed up to predict the likeliness to be suffering from a particular disease.

Many steps like disease searching, data collection, data cleaning and analysis, data visualisations, searching correlations, data pre processing through various algorithms have been explained in above pages.

All the codes were written in python language(using libraries like pandas, matplotlib, sklearn) in jupyter notebook or collaboratory.

Since each disease have different set of parameters and effected organs, data collected had specific attributes as mentioned:

1.**LUNGCANCER**:Smokingstatus,yellowfingers,anxiety,peer pressure,chronicdisease,fatigue,allergy,cough,alchoholconsumption,shortness of breath, chest pain, difficulty in breathing

2.**FETALHEALTH:**chestpain,restingbloodpressure,cholesterol,fastingbloodsugar,electrographicresults,maximumheartrate achieved,angina induced while exercise,previous peak,number of major vessels,Thal rate

3.**BREAST CANCER**: details about each nucleus, severity of pain, feeling of soreness or lump

4.**STROKE**: Hypertension, work type, residence type, average glucose level, BMI, smoking status

5.**LIVER**: Total Bilirubin, conjugated bilirubin, alkaline phosphatase,Alanine Aminotransferase, total proteins, Albumin and globulin.

6.**HEART ATTACK:** chest pain severity, resting BP, cholesterol, Fasting BP, maximum heart rate achieved, resting electrographic results, exercise induced results, exercise induced angina, previous peak

7.**TUMOR**:Radius,texture,perimeter,area,concavity,smoothness,compactness

8.**COVID 19**: Chest pain severity, BP, cholesterol ,fasting BP, maximum heart rate achieved, electrographic results, induced angina, previous peak

Based on these ,user will input the values and we train and test the model by different algorithms(like Nave bayes, Logistic regression, Decision tree classifier, random forest, K neighbors) and chances of contracting the disease is predicted. The model with highest accuracy is selected and implemented.

***We have created a website named ‘icare’ which has the following features:***

* Set of questionnaires where the user can input the details about the symptoms and get their chances of having it
* Health blogs by experts
* Diet and routine advices for the patients higher on risks
* Doctor’s recommendations on the same

**FUTURE SCOPE**

We see endless possibilities in trying to improve our healthcare with machine learning and we ought to do the same. Our project can also be extended for providing complete care to the patient visiting our website(‘icare’).

* After the patient finds himself or herself vulnerable to a disease, elaborative plan on hospital management, diagnostics and necessary dietary changes along with exercising schedule would be well handed to the person.
* We would like to ensure the treatment goes hassle free by providing the best doctors in the area or a particular hospital keeping in mind the person’s economic condition.
* Since healthcare and agriculture needs the most attention ,the model can be expanded so that it can use image recognition to detect plant disease in crop and thus help agriculture.
* We wish to link professional help directly to our websites(by providing toll free numbers) where you can get your initial consulations free or low on cost from the doctors who wish to collaborate

**BIBLIOGRAPHY**

**Many resources and the work already done in this field helped us a lot to shape our project. Many useful insights from various research papers, surveys and international journals were collected as mentioned below.**

## *Research Papers and Surveys -*

* ”Disease prediction using principal of component analysis” in  2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC).
* ”Disease Prediction in Data Mining Technique – A Survey” by S.Vijiyarani and S.Sudha

* A Survey on Disease Prediction by Machine Learning over Big Data from Healthcare Communities

International Journals-

* ”International Journal of Computer Applications & Information Technology”
* [2017 International Conference On Smart Technologies For Smart Nation (SmartTechCo](https://ieeexplore.ieee.org/xpl/conhome/8356464/proceeding)n)
* IOSR Journal of Engineering (IOSRJEN)
* International Journal of Computer Trends and Technology (IJCTT)
* International Research Journal of Engineering and Technology (IRJET))

*For collecting Datasets for training our machine learning model:*

* [*https://www.kaggle.com/fedesoriano/stroke-prediction-dataset*](https://www.kaggle.com/fedesoriano/stroke-prediction-dataset)
* [*https://www.kaggle.com/uciml/breast-cancer-wisconsin-data*](https://www.kaggle.com/uciml/breast-cancer-wisconsin-data)
* [*https://www.kaggle.com/imkrkannan/lung-cancer-dataset-by-staceyinrobert*](https://www.kaggle.com/imkrkannan/lung-cancer-dataset-by-staceyinrobert)
* [*https://www.kaggle.com/uciml/indian-liver-patient-records*](https://www.kaggle.com/uciml/indian-liver-patient-records)
* [*https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset*](https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset)
* [*https://www.kaggle.com/imkrkannan/lung-cancer-dataset-by-staceyinrobert*](https://www.kaggle.com/imkrkannan/lung-cancer-dataset-by-staceyinrobert)
* [*https://www.kaggle.com/uciml/indian-liver-patient-records*](https://www.kaggle.com/uciml/indian-liver-patient-records)
* [*https://www.kaggle.com/samsonqian/tumors*](https://www.kaggle.com/samsonqian/tumors)

*For learning sklearn,machine learning algorithms and python libraries:*

* [*https://scikit-learn.org/stable/*](https://scikit-learn.org/stable/)
* [*https://github.com/scikit-learn/scikit-learn*](https://github.com/scikit-learn/scikit-learn)
* [*https://pandas.pydata.org/*](https://pandas.pydata.org/)
* [*https://www.youtube.com/watch?v=ZyhVh-qRZPA*](https://www.youtube.com/watch?v=ZyhVh-qRZPA)
* [*https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html*](https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html)
* [*https://www.tutorialspoint.com/machine\_learning\_with\_python/machine\_learning\_with\_python\_classification\_algorithms\_logistic\_regression.htm#:~:text=Logistic%20regression%20is%20a%20supervised,be%20only%20two%20possible%20classes.&text=Mathematically%2C%20a%20logistic%20regression%20model,as%20a%20function%20of%20X*](https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic_regression.htm#:~:text=Logistic%20regression%20is%20a%20supervised,be%20only%20two%20possible%20classes.&text=Mathematically%2C%20a%20logistic%20regression%20model,as%20a%20function%20of%20X)*.*