GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY

Private University Estd. in Karnataka State by Act No. 41 of 2013

BANGALORE

PRESIDENCY GROUP OVER 40 YEARS OF ACADEMIC WISDOM

A Project Report

On

“Early Lifestyle Disease Prediction”

Batch Details

Sl. No.	Roll Number	Student Name
1	20201CSD0115	Faiz Khan
2	20201CSD0111	Jeevith Joseph
3	20201CSD0072	Tejas Gowda
4		
5		

School of Computer Science,

Presidency University, Bengaluru.

Under the guidance of,

Dr. V Chandra Sekar

School of Computer Science,

Presidency University, Bengaluru

CONTENTS

1. INTRODUCTION

This review delves into the innovative project titled "Early Lifestyle Disease Prediction." This project, with its focus on leveraging data and technology to address the increasing incidence of lifestyle-related diseases like diabetes and heart disease, represents a significant step in the realm of healthcare transformation.

Without a doubt, the prevalence of lifestyle-related diseases is on the rise, necessitating early detection and intervention to mitigate their impact. This initiative seeks to employ data-driven predictive modeling to tackle these health challenges proactively. It is not only about reducing the personal burdens these diseases place on individuals but also about lessening the economic strain they impose on healthcare systems.

- Challenges in Healthcare:

Today, people do not have time for the regular checkup. They are so busy with their work that they rarely have time for their own health. But the thing is that they can do the analysis if the appropriate application can provide them the overall health status of the person. This is because they need not to give the time separately for this, rather they can just utilize the time, for example, the time of traveling, etc. The only thing required is the smartphones which almost everyone has with them in this century. So, this can be considered as the portable health checker which everyone can use easily just through a web application. In this, we have decided to give the sign-up page where the user can sign up using his name, id and password. Then further modules will have the diseases portion. Though the whole health check is a somewhat more difficult task, so initially we are adding some of the very common diseases data like heart disease, breast cancer, etc. Later on, we will keep adding more and more diseases.

- Background:

Healthcare During the Pandemic: Prediction Of the disease during the pandemic is critical for public health planning of efficient health care allocation and monitoring the effects of policy interventions. In another study, an algorithm is proposed to achieve local updates and global updates which are critical for the learning process. ML/AI is also used to solve wireless network problems. Cheat AI. represented how artificial neural networks can be used to solve various problems in wireless networks.

Diabetes and Pre-diabetes: Diabetes has been in society for a very long time. Diabetes is further dependent on an individual’s body, diet, and way of living. In another study, pre-diabetes is predicted using different applications on the Korean population.

ICU Challenges: Death in intensive care units (ICUs) worldwide and its recognition, particularly in the early stages of the disease, remains a medical challenge.

- Motivation:

Lifestyle Diseases: Lifestyle diseases are common among the population today not only in India but also in almost every country. Lifestyle diseases are caused because of the habits that we have on a day-to-day basis. The way one lives his life is the major cause of it. It includes heart disease, hypertension, etc., which all may hear of. In our life also, one also comes across at least

one person who is either suffering from such diseases or the diseases became the reason for his death. We also came across many such people who died because they were not aware of their disease and were left with no appropriate time for treatment. That is why we decided to develop a portal for showing early disease information and solutions or precautions.

- Problem Definition:

Platform Overview: In this platform, the people and doctors are joined and they are registered login logout, and also they have early knowledge of what is happening in their area, and the true information is given to this platform. About the disease which is a large amount of spearing in this area first of all the people see and their area which diseases are spread and also check their symptoms, also in the bottom, and also give what the precaution do for the disease not to come, also if you have any other disease then also have a chat box to say your problem. There was the best doctor, and they convey you to what to do for these diseases. Doctors joined this platform as a sub-admin. They solve the people's posted questions and also add current disease information and safety tips and what precautions we do for these diseases and also know about these doctors who give this information. Admin doctor verifies this doctor's added information and verifies the doctor's status and manages the site settings.

- Scope:

Project Impact: This project will help the user to overcome the lifestyle diseases which are in themselves a big threat to humans, will reduce the unawareness about the diseases and will help people to remain healthy which is of utmost importance in today's fast-growing world. It will also change the lifestyle of people for the better. It will also give clarity about the health of a person or his current status.

- Future Prospects:

In future mostly AI and ML is going to be implemented everywhere. Everyone will get so busy with their work that they will not get enough time to visit the doctor, especially when they get serious. But they will ignore the minor and common disease which will eventually become more serious in the future. Like in tuberculosis, a person starts to cough, and only with preventive measures, they can be submerged at that time. So this app will become more and more common in the future. Without visiting doctors, people will get to know about their symptoms.

2. LITERATURE REVIEW

After a survey of this topic, we are developing this project using a machine learning concept, which is Data Mining. Another concept, like Deep Learning, is a technique for analyzing a huge amount of data in different aspects to discover useful information or knowledge discovery. It combines the concepts of artificial intelligence, statistics, probability, machine learning, deep learning, and database system technology. The processes of data collection, selection, cleaning, handling missing values, transformation, mining, evaluation of patterns, and knowledge visualization are involved in the Deep Learning process.

Title: Disease prediction by using machine learning [10]

Authors: Sayali Ambhekar and Dr. Rashmi Phalinkar

Year: 2018

The system that's used to prognosticate the conditions from the symptoms, which are given by the cases or any stoner. The system processes the symptoms handed by the stoner as input and gives the affair as the probability of the complaint. Naïve Bayes classifier is used in the vaticination of the complaint, which is a supervised machine learning algorithm. The probability of the complaint is calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early complaint discovery and case care.

Title: Prediction of diabetes based on personal lifestyle indicators [5]

Authors: Anand and Shakti

Year: 2015

Diabetes Mellitus or Diabetes has been portrayed as worse than Cancer and HIV (Human Immunodeficiency Virus). It develops when there are high blood sugar situations over a prolonged period. Recently, it has been quoted as a trouble factor for developing Alzheimer and a leading cause of blindness and order failure. Prevention of the complaint is a hot topic for discussion in the healthcare community. Multitudinous ways have been discovered to find the causes of diabetes and cure it. This disquisition paper is a discussion on establishing a relationship between diabetes trouble likely to be developed from a person's quotidian life conditioning, such as his/her eating habits, sleeping habits, physical exertion, along with other pointers such as BMI (Body Mass Index), waist circumference, etc.

The data is increasing exponentially, as in the case of the health sector. It is also a major data-producing sector, which is not only heterogeneous but also valuable as it stores sensitive health information about the person, which can even cost the life of a person. The majority of the methods are used to predict, prevent, and manage diseases appropriately and efficiently. Medical diagnosis is subjective and important in other aspects and depends on the data available, and in this case, the data entered by the user. Healthcare-related Deep Learning is a difficult field, as some minor changes may lead to a huge difference in the predictions and will further affect the output. It explores hidden patterns, which further helps in conveying and extracting knowledge in a database to predict diseases that a person may suffer from.

We will use both the core models of Deep Learning, i.e., descriptive as well as predictive in big

data. In the case of descriptive data analysis, it uses user data to identify patterns in the data and analyze the relationship between various variables and samples. Descriptive models are prior association rules, data clustering, summarizing, and visualization. These models are generally developed by using the complete dataset, but we will try to reduce the number of variables or samples required to predict the output, which increases its performance as well as efficiency. In the case of predictive data analysis, it uses historical data and current data for predicting the probabilities of future lifestyle diseases or used for diagnosing and curing the diseases as well.

Decision trees, artificial neural network (ANN), random forecasting, and regression (linear, logistic, and ridge) are the commonly used predictive data models.

System Analysis

System Analysis is the process of gathering and interpreting facts, diagnosing problems, and using the facts to improve the system. The System analysis chapter will show the overall system analysis of the concept, description of the system, meaning of the system. System analysis is the study of sets of interacting entities, including computer system analysis. The development of computer-based information systems includes the system analysis phase, which produces or enhances the data model, which itself is to create or enhance a database. There are a number of different approaches to system analysis. The analysis is the process used to analyze Portal for Early Prediction of Lifestyle Diseases.

Advantages of Existing Methods:

- Predictive Modeling in Healthcare: Existing methods often utilize predictive modeling in healthcare to assess an individual's risk of developing lifestyle-related diseases. These models leverage historical patient data, including demographic and vital statistics, which can provide valuable insights into disease likelihood. This approach offers the advantage of leveraging data for informed decision-making.

- Demographic and Vital Statistics Integration: Prior research has shown that incorporating demographic and vital statistics, such as age, gender, BMI, blood pressure, and cholesterol levels, enhances the accuracy of predictions. This integration ensures that the predictions are based on comprehensive information, which can be beneficial for preventive healthcare.

- Cost Reduction through Early Intervention: Several studies have underscored the cost-saving potential of early disease prediction. By identifying individuals at risk, healthcare providers can implement preventive measures, reducing the need for expensive treatments and hospitalizations. This not only benefits individuals but also eases the economic burden on healthcare systems.

- Advancements in Machine Learning Algorithms: With the advancement of machine learning techniques, existing methods are becoming more accurate and efficient in predicting lifestyle-related diseases. Algorithms such as logistic regression, decision trees, random forests, and support vector machines are increasingly reliable in identifying at-risk individuals.

Limitations (Disadvantages) of Existing Methods:

- Data Quality Challenges :Existing methods often grapple with challenges related to data

quality. Inaccurate or incomplete data can compromise the effectiveness of predictive models, leading to unreliable predictions.

- Model Interpretability : Some existing predictive models may lack interpretability. This means that it can be challenging for healthcare professionals and individuals to understand how the model arrived at a particular prediction, potentially reducing trust in the results.

- Continuous Model Refinement : To maintain the accuracy of predictive models, ongoing refinement and updates are necessary. The need for continuous model improvement poses a challenge in terms of resource allocation and maintenance.

- Privacy Concerns : The integration of sensitive medical and personal data in existing methods raises privacy concerns. Ensuring that user data is protected and compliant with relevant regulations is a constant challenge.

While these are general advantages and limitations related to existing methods for predicting lifestyle-related diseases, it's essential to note that the specifics of the project's methods may differ. Further details and methodological considerations will be explored in subsequent sections.

3. OBJECTIVES

Enhance Predictive Model Accuracy:

The primary objective of this project is to improve the accuracy of predictive models used for early lifestyle disease prediction. By leveraging the insights gathered from existing methods and addressing their limitations, we aim to create a more precise and reliable predictive model.

Ensure Model Interpretability and Transparency:

A critical goal is to enhance the interpretability and transparency of the predictive model. This objective stems from the limitations of existing methods, where complex models may lack user-friendliness. We aim to develop a model that not only provides accurate predictions but also ensures that healthcare professionals and individuals can easily comprehend and trust the results.

Personalized Disease Prediction:

Provide personalized disease prediction for individuals based on their unique combination of genetic and lifestyle data.
Deliver actionable insights and recommendations to individuals, suggesting lifestyle modifications and preventive measures.

Optimize Data Privacy and Security:

Building on the privacy concerns highlighted in the literature survey, our project aims to prioritize data privacy and security. We intend to implement stringent measures for anonymizing and encrypting sensitive information, thereby safeguarding user data and ensuring compliance with relevant regulations.
Promote User Engagement and Education:

Beyond the technical aspects, we aim to create a user-centric platform that promotes engagement and education. This objective aligns with the advantages of existing methods related to enhancing health awareness. The project will provide educational resources to users, enhancing their understanding of lifestyle diseases, risk factors, and preventive measures.
These objectives directly address the research gaps identified in the literature survey and guide our efforts toward the development of an innovative and user-friendly solution for early lifestyle disease prediction.

Overall Objectives:

- Empower Personalized Healthcare:
  Develop a platform that empowers individuals to take control of their health
  by providing personalized predictions and actionable insights regarding lifestyle diseases.
- Cost-Effective Prevention: Create a cost-effective healthcare solution that reduces the financial burden on
  individuals and healthcare systems by preventing disease onset and costly treatments.
- Data-Driven Healthcare: Leverage the power of data analytics and machine learning to pioneer an innovative
  approach to healthcare that emphasizes early prediction and prevention.

Specific Objectives :
• Comprehensive Data Acquisition: Acquire a comprehensive dataset, including diverse demographic information
and vital statistics, to underpin accurate disease prediction.
• Optimized Feature Selection: Identify and select the most influential features to enhance the precision of lifestyle
disease predictions.
• User-Centric Interface: Develop a user-centric website with a user friendly interface that enables effortless data
input and real-time prediction retrieval.
• Intuitive User Experience: Implement natural language processing (NLP) to ensure a user-friendly and intuitive
experience while interacting with the website.

The objective of this project is as follows:
-Easily predict diseases possibilities.
-Show diseases information.
-Clear people's doubts about any disease.
-Create a user friendly web application.

METHODOLOGY :

Data Collection:
Data will be collected from hospitals with the consent of patients who have completed their DNA test. Hospitals will provide test results and other essential factors necessary to develop the proposed system. The dataset shall contain the following patient attributes:

1. Unhealthy eating habits (1–5)
2. Lack of physical activity (1–5)
3. Obesity (yes/no)
4. Stress and anxiety (1–5)
5. Poor sleep (1–5)
6. Smoking (daily, sometimes, or never)
7. Alcoholism (daily, sometimes, or never)
8. Family history of lifestyle disease (yes/no)
9. Gender (male/female) (Grading: 1=excellent, 2=good, 3=average, 4=bad, 5=very bad)
10. Age
11. Body Mass Index (BMI)
12. Waist circumference
13. Blood pressure (systolic and diastolic)
14. Cholesterol levels (HDL and LDL)
15. Physical activity level (sedentary, low, moderate, high)
16. Diet composition (balanced, high sugar, high fat, etc.)
17. Hours of daily screen time
18. Frequency of fast food consumption
19. Sleep duration (hours per night)
20. Stress triggers (work-related, personal, financial, etc.)
21. Mental health assessment (depression, anxiety, etc.)
22. Blood sugar levels (fasting and post-meal)
23. Heart rate at rest
24. History of cardiovascular diseases (yes/no)
25. History of respiratory diseases (yes/no)
26. History of metabolic diseases (yes/no)
27. Alcohol consumption patterns (beer, wine, spirits)
28. History of cancer (yes/no)
29. Medication usage (including prescription and over-the-counter drugs)
30. Dental health and oral hygiene (e.g., frequency of dental check-ups, oral hygiene practices)

These attributes can provide comprehensive data for the assessment of lifestyle diseases and their risk factors.

The aforementioned attributes have confirmed to be linked to the majority of lifestyle diseases. As per [5], lifestyle disease diabetes can take place due to insalubrious eating habits, genetics, and absence of physical activity. Note that the attributes will be input variables for the simulated system.

The abovementioned attributes are daily records averaged over a week. For the simulated system to work efficiently, data was assembled from people who suffer a lifestyle disease and from those who fail to clear DNA tests. Lifestyle diseases should be epitomized correspondingly so that the simulated system is unbiased toward a certain disease.

Data Preprocessing:
Data preparation requires approximately 80% of time. Once data is gathered, it needs to be preprocessed, cleaned, constructed, and formatted in a style that SVM comprehends and is able to work with. Data mining tools should be used to analyze collected real-time data. There is a possibility that real-time data might hold misplaced values; they need to be replaced with a median. Herein, data has to be comprehensively reconnoitred, and patterns or similarities in data need to be recognized.

Data preprocessing includes the following steps:

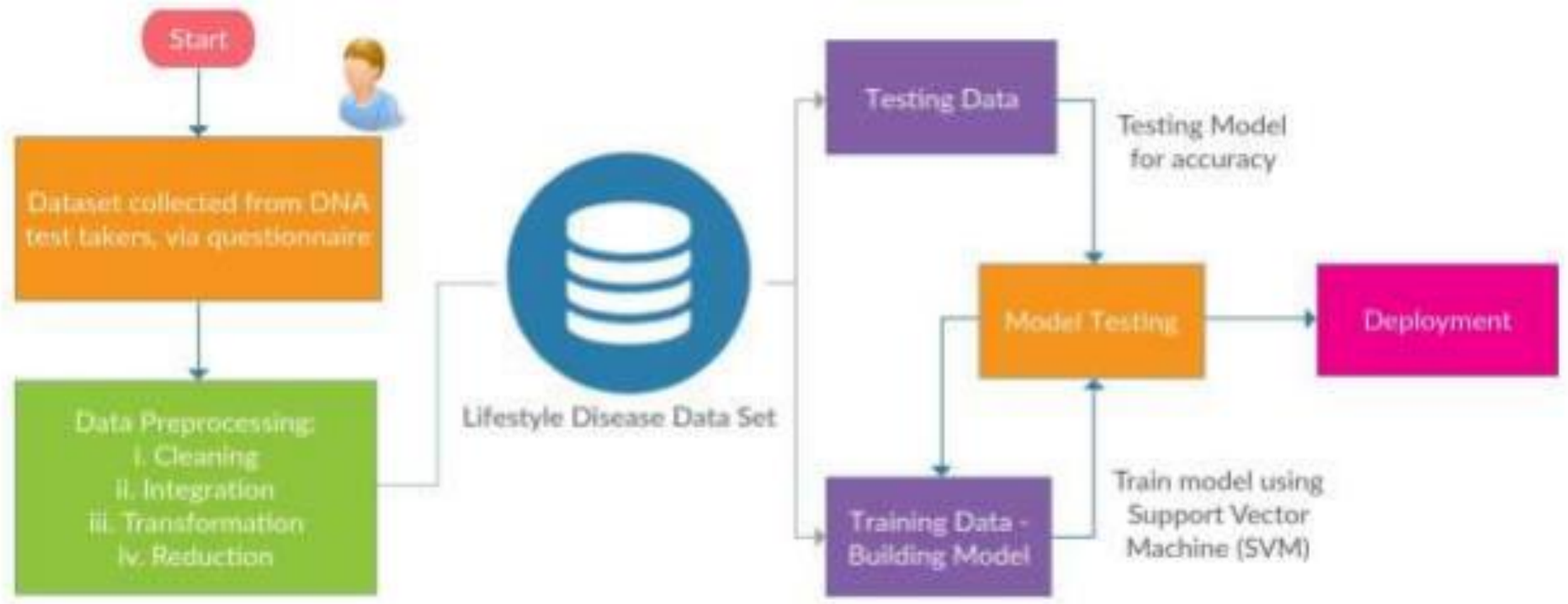
Data Integration: It is the process of combining significant data from different sources. Data will be obtained from the questionnaire that patients fill out after their DNA test.
Data Transformation: It is the process of applying different algorithms to integrated data for significant outputs using data analysis.
Data Reduction: Data filtering takes place, and only necessary data is considered for analysis. Only required fields will be taken.
Data Cleaning: It is the process of treating noisy data, inconsistency, and missing values in data. Spurious records are removed to make the dataset clean.
After data is preprocessed, it is added to the lifestyle disease dataset, which can be used for training and testing purposes.

Training and Testing Data:
The proposed model needs to be trained and tested under various conditions by altering SVM parameters so that correctness can be obtained. From the collected data, 70:30 will be used to train and test the model, respectively. In case of necessity, there must be provisions to improvise on the algorithm being used. Furthermore, the model must adapt to new changes made in the dataset as the dataset size will constantly increase. Data preprocessing would be needed to be performed on newly added data and include it in previously collected results. The model can then be retrained and checked for efficiency.

Working of the Model (Model Testing):
An individual who desires to know whether they are exposed to a lifestyle disease can make use of the model as a replacement for DNA tests. A questionnaire will be provided on a web application asking an individual to rate their eating habits, physical activity, anxiety, sleep, etc. Once an individual submits their questionnaire, the data collected via the questionnaire will act as an input to the model working behind the web application on a cloud service like Amazon Web Services or Microsoft Azure. The model will quickly respond with predicted results, which will be shown to the individual.

The obtained results should specify whether a person is susceptible to a disease or not. Also, it should display graphs, charts showing an individual the probability of them suffering a disease. Results should also advise an individual on medicines or exercises and motivate them to live a healthy lifestyle. Compared to DNA tests, the model will prove to be faster, cheaper, and easier to predict an individual's chances of suffering from a lifestyle disease. Moreover, there is a provision for an individual to change their input parameters and check for predictions.

Deployment:
Once the model is tested thoroughly, the web application will be deployed for users.


Start
Dataset collected from DNA test takers, via questionnaire
Data Preprocessing:
I. Cleaning
II. Integration
iii. Transformation
Iv. Reduction
Lifestyle Disease Data Set
Testing Data
Testing Model for accuracy
Model Testing
Deployment
Training Data - Building Model
Train model using Support Vector Machine (SVM)

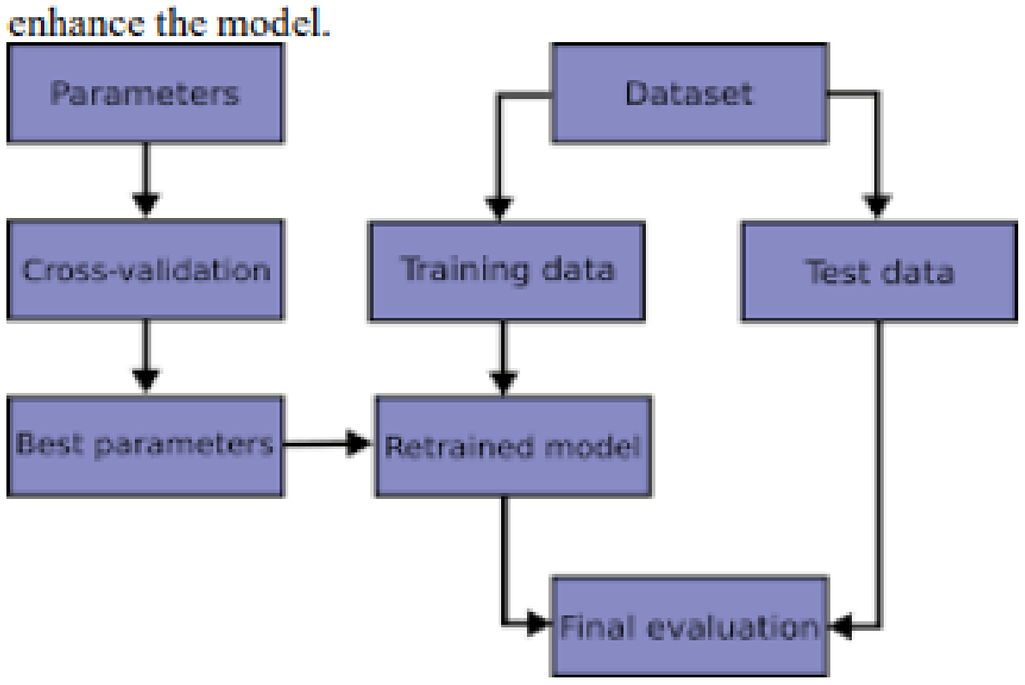
4. EXPERIMENTAL DETAILS

Solution to the problem:

Our system predicts the chances of diseases such as heart disease, diabetes, and cancer are major health challenges facing
societies worldwide. Lifestyle factors, such as diet, physical activity, stress levels, and sleep patterns, have a significant impact on
disease risk. In this study, we aimed to use machine learning algorithms to analyze individual lifestyle factors and predict the likelihood
of developing certain diseases.
Steps involve in data set processing :

- Import the libraries.
- Import the data set.
- Get the basic information about the data set.
- Get the correlation parameter for the variables.
- If required plot and analyse the parameters.
- Check for nulls in the data set.
- If any drop those rows.
- Check for any special characters in the columns.
- Just convert them in to nulls and drop those rows.
- Now need to select the features for the project
- Taking all the inputs and predicts the chances of disease as output.
- Now need to split the data set into sample and verification data set.
- And train the machine using the Randomforest.
- And finally using this trained model is routed to the flask web application which can create an application to work.
- To take the inputs from the user and display the result.

enhance the model.



Model Selection

- **Logistic Regression**: Logistic regression is a simple yet effective model for binary classification tasks. It's a good choice when you have a relatively small dataset and want to understand the relationship between features and disease risk.

- **K-Nearest Neighbors (KNN):** KNN is a non-parametric and instance-based model that can be used for both classification and regression tasks. It's suitable for situations where similar individuals are likely to have similar disease risks.

- **Random Forest**: Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy. It's robust, handles feature importance well, and can work with both numerical and categorical data.

- **Support Vector Machine (SVM)**: SVM is effective in separating classes and works well for binary classification. It's particularly useful when you have a moderate amount of data and want to find a clear boundary between disease and non-disease cases.

- **Neural Networks**: Deep learning models, such as feed-forward neural networks, can be used for disease prediction. They are capable of handling complex data patterns, but they may require larger datasets and more computational resources.

- **Naive Bayes**: Naive Bayes classifiers are simple probabilistic models that are particularly useful when dealing with text data or when you have a large number of features.

- **Decision Trees**: Decision trees are interpretable models that can be used for classification tasks. They are helpful in understanding the decision-making process.

- **Ensemble Method**s: Ensemble learning is a machine learning approach that combines predictions from multiple models to increase predictive performance . Ensemble techniques integrate various machine learning algorithms to make more accurate predictions than a single classifier. The use of ensemble models is intended to reduce the generalization error. This technique reduces model prediction error when the base models are diverse and independent. As outlined in figure 1 The approach relies on the collective output of individuals for generating forecasting. Although several base models exist, the ensemble model works and behaves as a single model

- **Bagging**: Bagging is aggregating the predictions of many decision trees that have been fit to different samples of the same dataset. Ensemble bagging is created by assembling a series of classifiers that repeatedly run a given algorithm on distinct versions of the training dataset. Bagging, also known as bootstrapping ,is the process of resampling data from the training set with the same cardinality as the starting set to reduce the classifier's variance and over-fitting . Compared to a single model, the final model should be less overfitted. A model with a high variance indicates that the outcome is sensitive to the practice data provided . As a result, even with more training data, the model may still perform poorly and may not even lower the variance of our model. The overall framework for bagging is presented in.

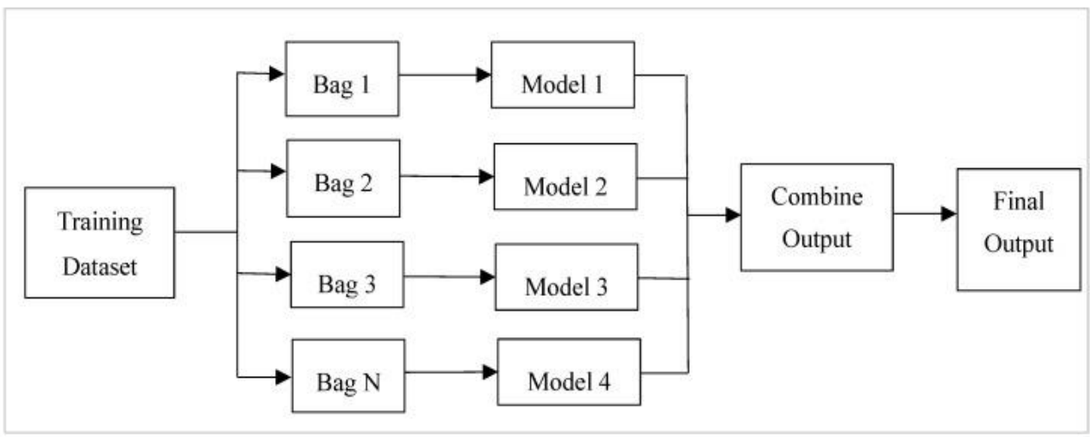


Figure 1 Steps followed in the bagging approach. Bags of data are formed from the input dataset, and models are used on all the bags. The output is combined from all the models.

- **Boosting:** Boosting algorithms use weighted averages to transform poor learners into strong ones. During boosting, the original dataset is partitioned into several subgroups. The subset

is used to train the classifier, which results in a sequence of models with modest performance . The elements that were incorrectly categorised by the prior model are used to build new subsets. The ensembling procedure then improves its performance by integrating the weak models using a cost function. It explained that, unlike bagging, each model functions independently before aggregating the inputs, with no model selection at the end. Boosting is a method of consecutively placing multiple weak pupils in a flexible manner. Intuitively, the new model focuses on the discoveries that have been shown to be the most difficult to match up until now, resulting in a good learner with less bias at the end of the process . Boosting can be used to solve regression and identification problems, such as bagging. illustrates the framework of the boosting approach.

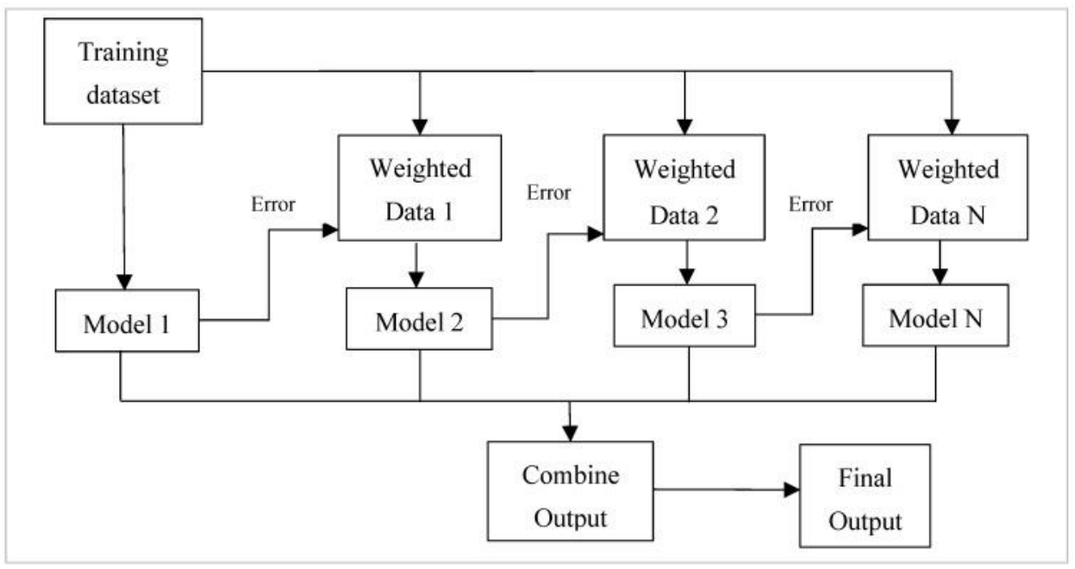


Figure 2

The framework used in the boosting approach. Different models are considered, and each model tries to compensate for the weakness of its predecessor by reducing the error.

Voting: The Voting Classifier ensemble approach is a strategy that aggregates predictions from numerous independent models (base estimators) to make a final prediction . It uses the “wisdom of the crowd” notion to create more accurate predictions by taking into account the aggregate judgement of numerous models rather than depending on a single model. In the Voting Classifier, there are two types of voting: hard voting, in which each model makes a prediction, and soft voting, in which each model forecasts the probability or confidence ratings for each class or label. The final prediction is made by summing the expected probabilities across all models and choosing the class with the highest average probability . Weighted voting allows multiple models to have different influences on the final forecast, which can be assigned manually or learned automatically based on the performance of the individual models. Because of this diversity, different models can affect the final prediction differently .

By combining the strengths of several models, the Voting Classifier increases overall performance and robustness, especially when distinct models have diverse properties and generate independent predictions. The Voting Classifier can overcome biases or limits in a single model and produce more accurate and trustworthy predictions by using the collective decision-making of numerous models. Overall, Voting Classifier is a versatile ensemble strategy that can be applied to a variety of machine learning applications by using the capabilities of different models to make more accurate predictions . The Voting Classifier is a versatile ensemble approach in machine learning that provides a number of benefits. By integrating various models with diverse strengths and weaknesses, it enhances accuracy, robustness, and model diversity while minimising bias and variance in predictions.

The Voting Classifier’s ensemble nature improves model stability by decreasing overfitting and increasing model variety . It offers various voting procedures like as hard voting, soft voting, and weighted voting, allowing for customisation based on the tasks and characteristics of specific

models. Furthermore, the Voting Classifier can improve interpretability by analysing the contributions of many models, assisting in understanding the underlying patterns and decision-making process . Overall, the Voting Classifier is an effective tool for enhancing predictive performance in various machine learning tasks.

DESIGN PROCEDURE

The design procedure for the development of the "Early Lifestyle Disease Prediction" web-based application involves the following steps:

- Requirement Analysis:

Define the functional and non-functional requirements of the application.
Identify user expectations, data sources, and specific features required for lifestyle disease prediction.
System Architecture Design:

Determine the overall system architecture, including front-end and back-end components.
Choose appropriate web development frameworks, machine learning libraries, and database management systems.
Plan the structure of the user interface, data storage, and server infrastructure.
Data Collection and Preparation:

Gather a comprehensive dataset containing demographic information and vital statistics.
Ensure data quality, accuracy, and privacy compliance by anonymizing and securing sensitive information.
Split the dataset into training and testing sets for model development and evaluation.
Feature Selection and Engineering:

Identify relevant features from the dataset, such as age, gender, BMI, blood pressure, and lifestyle habits.
Perform feature engineering to create new variables or transform existing ones to improve predictive accuracy.

- Machine Learning Model Development:

Implement machine learning algorithms suitable for classification tasks, such as logistic regression, decision trees, random forests, or support vector machines.
Train the models on the training dataset using the selected features.
Evaluate model performance using metrics like accuracy, precision, recall, and F1-score, and select the most appropriate model.

- Web Application Development:

Develop a user-friendly web interface for interaction with users.
Implement a natural language processing (NLP) module to understand user queries and requests.
Allow users to input their parameters, including age, gender, BMI, and other vital statistics.
Disease Prediction Module Integration:

Integrate the trained machine learning model into the website interface.
On user input, pass the parameters to the model, which will then predict the likelihood of lifestyle diseases based on the provided data.

Provide users with a clear and interpretable prediction, indicating their risk of developing specific diseases.

Testing and Validation:

Conduct extensive testing to ensure the application's functionality, security, and performance.
Validate the accuracy of disease predictions through real-world data testing.

Deployment:

Deploy the web-based application on the chosen server infrastructure.
Monitor and maintain the application for optimal performance.
The design procedure encompasses various stages, from requirements analysis to deployment, with a focus on delivering an accurate, user-friendly, and secure platform for early lifestyle disease prediction.

REQUIREMENTS :

Hardware Requirements :

- System : Any.
- RAM : 4 Gb or above.
- Hard Disk : 1Tb or above.
- Input Device : Keyboard or Mouse.
- Output Device : Monitor, PC or Laptop

Software Requirements:

- Operating System : Any.
- IDE : Visual Studio Code / Google Collab.
- Python
- Front End :
  - HTML (Hypertext Markup Language).
  - CSS (Cascading Style Sheets).
  - Java Script.
- Back End : Node Js.
- Library :
  - React Js.
- FrameWork : Express, Flask.
- UI FrameWork : React-TailWindCSS.
- Database : MongoDB.
- Project Managed : GitHub Git.

5. OUTCOMES

The "Early Lifestyle Disease Prediction" project is anticipated to yield several significant outcomes, contributing to healthcare improvement and the well-being of individuals. These outcomes include:

Accurate Lifestyle Disease Predictions:

The project aims to provide users with precise predictions regarding their risk of developing specific lifestyle diseases based on input data. These predictions empower individuals to make informed decisions about their health.
Enhanced Health Awareness:

Users will gain access to educational resources that improve their understanding of lifestyle diseases, risk factors, and preventive measures. This educational component promotes health literacy and empowers individuals to take proactive steps toward better health.
User Engagement and Satisfaction:

By adopting a user-centric approach and creating an intuitive interface, the project is designed to attract and engage a diverse user base. Encouraging regular utilization of the platform enhances user satisfaction and increases the likelihood of proactive health management.
Cost Savings in Healthcare:

By enabling early prediction and prevention of lifestyle diseases, the project has the potential to contribute to cost savings in healthcare. It reduces the need for expensive treatments and hospitalizations, benefiting both individuals and healthcare systems.
Contribution to Research:

The project's innovative approach to predictive healthcare and preventive medicine provides a valuable tool for healthcare professionals and researchers. It contributes to ongoing research in the field and offers a platform for further exploration.
The envisioned outcomes of this project extend beyond technology. They represent a significant step toward healthier lives, informed choices, and a healthcare future where prevention plays a paramount role. With each prediction, each piece of health education, and each empowered user, we collectively work toward a healthier, more informed, and more resilient society.

6. TIMELINE OF THE PROJECT/ PROJECT EXECUTION PLAN

Sl. No	Review	Date	Scheduled Task
1	Review - 0	09-10-23 to 13-10-23	Initial Project Planning and Proposal Submission.
2	Review - 1	06-11-23 to 10-11-23	Completion of Research and Data Collection and Preparation.
3	Review - 2	27-11-23 to 30-11-23	Completion of Machine Learning Model Development
4	Review - 3	26-12-23 to 30-12-23	Completion of Website Development and Deployment
5	Final	08-01-24 to 12-01-24	Project Submission and Presentation for Evaluation.

7. CONCLUSION

In closing, the "Early Lifestyle Disease Prediction" project stands as a beacon of progress in the realm of healthcare. It embodies the potential of technology to transform lives and reduce the burden of lifestyle-related diseases. The project's objectives, methodology, and expected outcomes collectively pave the way for accessible, personalized, and data-driven health management.

This study demonstrates the potential of machine learning for improving the accuracy and effectiveness of disease prediction, and for empowering individuals to take a proactive approach to their health and reduce their risk of developing certain diseases. Further additions to the model would include when an individual enters his/her details (i.e., input to the predictive model), the model would determine his/her identity based on several inputs, show an individual’s current status of his/her health contrary to a desired ideal health using graphs, let know lifestyle changes, provide balanced diet and doctor consultations, recommend exercises, etc.

The significance of this endeavor transcends technology. It signifies a critical stride toward proactive healthcare, where prevention takes precedence over treatment. The commitment to improving public health, reducing healthcare costs, and enhancing user satisfaction underscores the project's importance.

Each prediction generated, each piece of health education offered, and each individual empowered by this project contributes to a future where health is paramount, and knowledge is a potent tool. This "Early Lifestyle Disease Prediction" project is more than just a project; it's a testament to the transformative power of technology in improving lives, reducing disease burdens, and fostering healthier, more proactive lifestyles. It is a step into the future of healthcare, where prevention is the key, and information is a powerful ally.

REFERENCES

Here are the references formatted properly:

1. Sharma, M. and Majumdar, P.K. (2009). "Occupational lifestyle diseases: An emerging issue." *Indian Journal of Occupational and Environmental Medicine, 13*(3), 109–112.

2. DNA Test Cost in India. Available [Online] https://www.dnaforensics.in/dna-test-cost-in-india/ [Accessed on June 27, 2018].

3. Suzuki, A., Lindor, K., St Saver, J., Lymp, J., Mendes, F., Muto, A., Okada, T. and Angulo, P. (2005). "Effect of changes on body weight and lifestyle in nonalcoholic fatty liver disease." *Journal of Hepatology, 43*(6), 1060–1066.

4. Pattekari, S.A. and Parveen, A. (2012). "Prediction system for heart disease using Naïve Bayes." *International Journal of Advanced Computer and Mathematical Sciences, 3*(3), 290–294.

5. Anand, A. and Shakti, D. (2015). "Prediction of diabetes based on personal lifestyle indicators." In *Next generation computing technologies (NGCT), 2015 1st international conference on,* 673–676. IEEE.

6. Kanchan, B.D. and Kishor, M.M. (2016). "Study of machine learning algorithms for special disease prediction using principal of component analysis." In *Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016 International Conference on,* 5–10. IEEE.

7. Kazeminejad, A., Golbabaei, S. and Soltanian-Zadeh, H. (2017). "Graph theoretical metrics and machine learning for diagnosis of Parkinson‘s disease using rs-fMRI." In *Artificial Intelligence and Signal Processing Conference (AISP),* 134–139. IEEE.

8. Milgram, J., Cheriet, M. and Sabourin, R. (2006). "—One against one‖ or —one against all‖: Which one is better for handwriting recognition with SVMs?." *Tenth International Workshop on Frontiers in Handwriting Recognition,* La Baule (France), Suvisoft, 2006.

9. Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H. and Jahan, H. (2018). "PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques." *Procedia Computer Science, 132,* 1068–1076.

10. Sayali Ambekar and Dr. Rashmi Phalnikar (2018). "Disease prediction by using machine learning." *International Journal of Computer Engineering and Applications, vol. 12,* 1–6.

These references provide valuable insights and research findings that have contributed to the development and understanding of early lifestyle disease prediction. They have been instrumental in shaping the project and its objectives.