Early Lifestyle Disease Prediction

Batch Number: CSD13

Roll Number	Student Name
FAIZ KHAN	20201CSD0115
JEEVITH JOSEPH	20201CSD0111
TEJAS GOWDA	20201CSD0072

Under the Supervision of,

Dr. V Chandra Sekar
School of Computer Science & Engineering
Presidency University

Introduction

- In an era where data and technology are transforming every aspect of our lives, healthcare is no exception. With the rising prevalence of lifestyle-related diseases, such as diabetes, heart disease, and obesity, the need for innovative and cost-effective preventive healthcare solutions has never been more pressing. Early detection and intervention are key to reducing the human and economic toll of these diseases. This project sets out to explore a novel approach to healthcare that harnesses the power of detailed demographic and vital statistics to predict the likelihood of lifestyle diseases, enabling timely interventions and substantial cost savings in the long run.
- Lifestyle diseases are characterized by their strong association with lifestyle choices, such as diet, exercise, and stress management. It is well-established that early identification of risk factors and lifestyle modifications can significantly reduce the incidence of these diseases. The traditional healthcare system, however, is often geared towards treating diseases once they manifest, incurring substantial costs for patients, providers, and society as a whole.
- 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. In this project, we will delve into the methods and technologies necessary to harness the potential of data-driven predictive modeling in healthcare. By integrating technology, data, and medical expertise, we aim to pave the way for a healthcare system that not only extends lives but also makes them healthier and more affordable. This proactive approach holds the promise of not only enhancing the quality of life for individuals but also reducing the economic burden of lifestyle diseases on healthcare systems worldwide.

Lifestyle Diseases

- Type 2 Diabetes - Blood sugar level, family history, diet habits.
- Hypertension (High Blood Pressure) - Blood pressure measurements, salt intake, exercise habits.
- Cardiovascular Disease (Heart Disease) - Cholesterol levels, smoking habits, exercise frequency.
- Stroke - Blood pressure, smoking status, alcohol consumption.
- Obesity - Body Mass Index (BMI), waist circumference, daily calorie intake.
- Metabolic Syndrome - Blood pressure, blood sugar level, BMI, waist circumference.
- High Cholesterol - Total cholesterol, LDL and HDL levels, dietary habits.
- Coronary Artery Disease - Family history of heart disease, cholesterol levels, smoking history.
- Osteoporosis - Bone density, dietary calcium intake, physical activity.
- Chronic Kidney Disease - Blood pressure, urine protein levels, diabetes status.
- Chronic Obstructive Pulmonary Disease (COPD) - Smoking history, shortness of breath.
- Asthma - Wheezing, coughing, family history of asthma.
- Liver Disease (Non-Alcoholic Fatty Liver Disease) - Liver enzymes, body weight, alcohol consumption.
- Gastroesophageal Reflux Disease (GERD) - Frequency of heartburn, diet habits.
- Rheumatoid Arthritis - Joint pain, stiffness, family history of arthritis.
- Depression - Mood changes, sleep disturbances, loss of interest.
- Anxiety Disorders - Worrying, restlessness, panic attacks.
- Sleep Apnea - Loud snoring, excessive daytime sleepiness, neck circumference.
- Colon Cancer - Family history of colon cancer, rectal bleeding, changes in bowel habits.
- Breast Cancer - Family history, breast lump, breast pain or tenderness.

Literature Review

Our literature review synthesizes findings from various studies related to lifestyle disease prediction using machine learning. We draw insights from the following key authors and papers:

1. Sharma and Majumdar (2009) have shed light on the emergence of "Occupational lifestyle diseases," emphasizing the importance of early detection. They explore this issue in the *Indian Journal of Occupational and Environmental Medicine* [1].

2. The "DNA Test Cost in India" website [2] provides valuable information on DNA testing costs, which plays a crucial role in our data collection process.

3. Suzuki et al. (2005) delve into the "Effect of changes on body weight and lifestyle in nonalcoholic fatty liver disease," as published in the *Journal of Hepatology* [3]. Their research highlights the intricate relationship between lifestyle factors and diseases.

4. Pattekari and Parveen (2012) present a "Prediction system for heart disease using Naïve Bayes" in the *International Journal of Advanced Computer and Mathematical Sciences* [4]. Their approach leverages machine learning algorithms for disease prediction.

5. Anand and Shakti (2015) explore the "Prediction of diabetes based on personal lifestyle indicators" in the *Next generation computing technologies* conference proceedings [5]. Their work underscores the significance of lifestyle indicators in predicting diabetes.

9. Hossain et al. (2018) present "PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques" in *Procedia Computer Science* [9]. Their approach focuses on obesity prediction and showcases the power of data mining.

10. Sayali Ambekar and Dr. Rashmi Phalnikar (2018) contribute significantly with their paper on "Disease prediction by using machine learning" in the *International Journal of Computer Engineering and Applications* [10]. This paper forms the basis of our project, emphasizing the importance of machine learning in disease prediction.

We have integrated these valuable insights into our project, ensuring a robust and informed approach to lifestyle disease prediction. These sources provide diverse perspectives, helping us build a comprehensive system for early prediction of lifestyle diseases.

Proposed Method : Developing a Website for Lifestyle Disease Prediction

- **Data Collection and Preparation:**
  - Gather a comprehensive dataset that includes demographic information and vital statistics of individuals, along with their lifestyle disease status (e.g., diabetes, heart disease).
  - 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, cholesterol levels, and lifestyle habits (e.g., smoking, physical activity, diet).
  - Perform feature engineering to create new variables or transform existing ones, if necessary, 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.
- **Website Integration:**
  - Develop a user-friendly website that incorporates a user friendly interface for interacting 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:**
  - 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.

Objectives

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.

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.

- 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.

- Training and Testing Data:

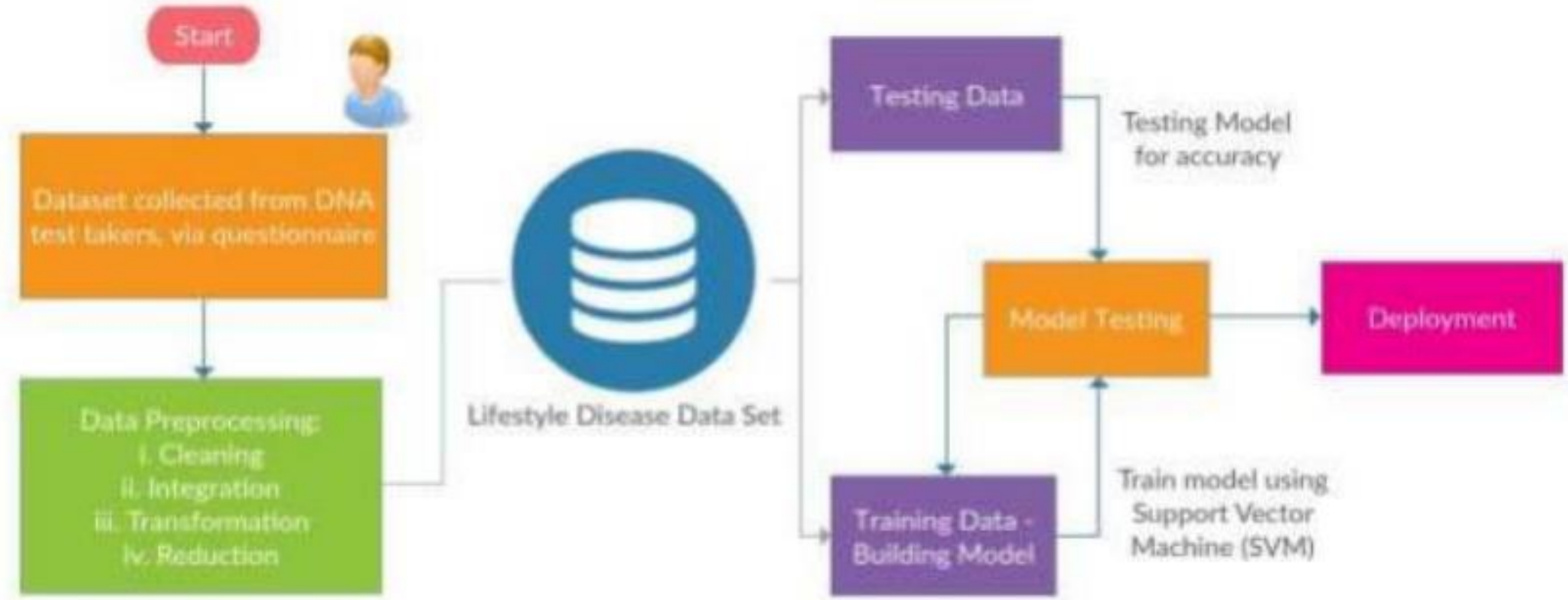
The proposed model needs to be trained and tested under various conditions by altering SVM parameters so that correctness can be obtained. From the collected data, 70:30 will be used to train and test the model, respectively. In case of necessity, there must be provisions to improvise on the algorithm 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)

EXAMPLE OF DATASETS

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
General_	Checkup	Exercise	Heart_Di	Skin_Car	Other_C	Depressi	Diabetes	Arthritis	Sex	Age_Cat	Height_(	Weight_(	BMI	Smoking_	Alcohol_	Fruit_Cor	Green_V	FriedPotato_Consumption		
Poor	Within the past	No	No	No	No	No	No	Yes	Female	70-74	150	32.66	14.54	Yes	0	30	16	12		
Very Goc	Within the past	No	Yes	No	No	No	Yes	No	Female	70-74	165	77.11	28.29	No	0	30	0	4		
Very Goc	Within the past	Yes	No	No	No	No	Yes	No	Female	60-64	163	88.45	33.47	No	4	12	3	16		
Poor	Within the past	Yes	Yes	No	No	No	Yes	No	Male	75-79	180	93.44	28.73	No	0	30	30	8		
Good	Within the past	No	No	No	No	No	No	No	Male	80+	191	88.45	24.37	Yes	0	8	4	0		
Good	Within the past	No	No	No	No	Yes	No	Yes	Male	60-64	183	154.22	46.11	No	0	12	12	12		
Fair	Within the past	Yes	Yes	No	No	No	No	Yes	Male	60-64	175	69.85	22.74	Yes	0	16	8	0		
Good	Within the past	Yes	No	No	No	No	No	Yes	Female	65-69	165	108.86	39.94	Yes	3	30	8	8		
Fair	Within the past	No	No	No	No	Yes	No	No	Female	65-69	163	72.57	27.46	Yes	0	12	12	4		
Fair	Within the past	No	No	No	No	No	Yes	Yes	Female	70-74	163	91.63	34.67	No	0	12	12	1		
Fair	Within the past	Yes	Yes	No	No	No	No	Yes	Female	75-79	160	74.84	29.23	No	0	30	20	2		
Fair	Within the past	No	Yes	Yes	No	No	Yes	No	Male	75-79	175	73.48	23.92	No	0	2	8	30		
Very Goc	Within the past	No	No	No	No	Yes	No	No	Female	50-54	168	83.91	29.86	No	8	8	0	2		
Fair	Within the past	No	No	Yes	No	No	No	No	Male	65-69	178	113.4	35.87	Yes	4	2	3	4		
Excellent	Within the past	Yes	No	No	No	No	No	No	Female	70-74	152	52.16	22.46	No	0	30	4	0		
Fair	Within the past	No	No	No	No	No	Yes	Yes	Female	70-74	163	116.12	43.94	No	0	8	8	4		
Good	Within the past	No	No	No	No	No	No	No	Male	80+	183	99.79	29.84	No	0	1	4	20		
Very Goc	Within the past	Yes	No	No	No	No	No	Yes	Male	80+	168	81.65	29.05	No	30	30	12	8		
Fair	Within the past	Yes	No	No	No	No	Yes	No	Male	45-49	178	104.33	33	Yes	2	16	12	8		
Good	Within the past	No	No	Yes	Yes	No	Yes	Yes	Female	70-74	163	79.38	30.04	No	0	12	8	4		

EXAMPLE OF DATASETS

Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Categor	Blood Pre	Heart Rate	Daily Step	Sleep Disorder
Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
Male	28	Sales Representativ	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
Male	28	Sales Representativ	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
Male	28	Software Engineer	5.9	4	30	8	Obese	140/90	85	3000	Insomnia
Male	29	Teacher	6.3	6	40	7	Obese	140/90	82	3500	Insomnia
Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
Male	29	Doctor	6.1	6	30	8	Normal	120/80	70	8000	None
Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
Male	29	Doctor	6.1	6	30	8	Normal	120/80	70	8000	None
Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None

Timeline of Project

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.

Expected Outcomes

- **Accurate Lifestyle Disease Predictions:** The project will provide users with accurate predictions regarding their risk of developing specific lifestyle diseases based on their input data, helping individuals make informed decisions about their health.
- **Improved Health Awareness:** Users will have access to educational resources that enhance their understanding of lifestyle diseases, risk factors, and preventive measures, promoting health literacy.
- **Enhanced User Engagement:** Through a user-centric approach and an intuitive user interface, the project will attract and engage a diverse user base, encouraging regular utilization of the platform.
- **Cost Savings in Healthcare:** By enabling early prediction and prevention of lifestyle diseases, the project can contribute to cost savings in healthcare by reducing the need for expensive treatments and hospitalizations.
- **Contribution to Research:** The project will contribute to the field of predictive healthcare and preventive medicine, providing a valuable tool for healthcare professionals and researchers.

Conclusion

- In an era marked by remarkable advancements in technology and an increasing focus on preventive healthcare, the development of a lifestyle disease prediction website represents a significant stride toward accessible, personalized, and data-driven health management. This project has culminated in the creation of a dynamic platform that empowers individuals to take proactive measures for their well-being and make informed decisions about their health.
- The positive impact of this project extends far beyond technology. It represents a step toward healthier lives, more informed choices, and a future where healthcare is not just curative but profoundly preventive. The commitment to improving public health, cost savings, and enhancing user satisfaction underscores the significance of this endeavor. With each prediction, each piece of health education, and each user empowered, we are collectively working toward a healthier, more informed, and more resilient society.
- In closing, this lifestyle disease prediction website is not just a project; it is a testament to the power of technology to improve lives, reduce the burden of disease, and empower individuals to embrace healthier, more proactive lifestyles. It is a step into the future of healthcare, where prevention is paramount, and knowledge is a potent tool.

References

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

- https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?select=dataset.csv
- https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction
- https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning
- https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset
- https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset
- https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression

Thank You