Predictive Analysis of Chronic Diseases based on Dietary Habits Identification using Machine Learning Techniques

Aniketh Mahesh Rao
School of Computing
National College of Ireland
Dublin, Ireland
x22166343@student.ncirl.ie

***Abstract*—Co-morbidities like Diabetes, Obesity, and cardiovascular diseases are leading causes of morbidity mortality. In today's day and age where an individual is always on the go, fast food is preferred over a healthy meal as it is cheaper and easily available. But these habits have a lot of impact on one's physical and mental health than we can ever imagine. Unhealthy eating habits coupled with a lack of physical activity are bound to lead to co-morbidities. A meal containing high fat, low fibre, and excess salt or sugar has been seen to cause a spike in blood sugar levels, and high cholesterol, and increased blood pressure. An increase in cholesterol circulating in blood vessels gets accumulated in the lumen of vessels causing narrowing of it and eventually forming a blockage or thrombus also known as arteriosclerosis. This could ultimately result in various cardiovascular diseases like coronary artery disease, cerebrovascular disease, peripheral artery disease, or aortic atherosclerosis. Hyperglycaemia leads to the formation of sugar-coated LDL which sticks around longer in the arteries thus heightening the probability of succumbing to these illnesses. Making lifestyle changes like at least 150 minutes per week of moderate-intensity physical activity, and improving eating habits along with dietary changes can bring about drastic changes as compared to just relying on medication. Regular exercise has been shown to make hemodynamic, morphological, metabolic, and psychological changes in an individual's body. Along with an increase in muscular strength, endurance, and overall improvement in the dynamic performance of the patient. Hence an improved and better quality of life.**

***Index Terms*—Machine Learning, Prediction, CRISP-DM, Chronic Disease**

I. Introduction

We live in a fast-paced and continuously changing world which evolves day by day but, in this hassle we tend to forget the importance of healthy lifestyle. Dietary habits, physical exercises and meditation are few of the daily power-dose which leads to wholesome lifestyle. As per studies, majority of Chronic diseases are often caused by unhealthy lifestyle along with genetic and environmental reasons.

Unhealthy lifestyle immensely focuses on the Dietary habits of a person which leads to growth of Chronic diseases which is a long-lasting and generally slow to progress illness which can treated by modifying the eating habits and changing the lifestyle. In this paper, Machine Learning algorithms are being applied to identify the dietary habits and physical activities of a person and predicting the probability of three major Chronic diseases i.e, Obesity, Cardiovascular disease-Cholesterol and Diabetes. Using different datasets we will try to predict the changes of one getting any Chronic diseases due to lack of active lifestyle.

In this study, we aim to unravel the intricate relationship between an individual's dietary habits and physical activity, and the likelihood of developing chronic diseases. Our approach involves comprehensive analysis of lifestyle patterns to determine the correlation with different types of chronic diseases. Specifically, we aim to identify unhealthy dietary habits that could potentially lead to chronic diseases and estimate the probability of their occurrence. Through this research, we seek to provide insights that can help individuals make informed choices regarding their lifestyle and prevent the onset of chronic diseases.

II. Related Work

A. Classification and Prediction on the Effects of Nutritional Intake on Overweight/Obesity using Deep Learning Model

The objective of the research was to classify and predict the link between nutritional intake and overweight/obesity risk by comparing a Deep Neural Network (DNN) model with other machine learning techniques such as logistic regression and decision tree. Obesity is a major public health concern, and the study aimed to identify the relationship between nutritional intake and obesity by applying prediction models. The research team used the DNN model along with other machine learning techniques, such as the k-nearest neighbor's algorithm (K-NN) and random-forests decision tree (RF) algorithms, to predict the association between nutritional intake and obesity. [1] [2]

B. Predictions of Diabetes and Diet Recommendation System for Diabetic Patients using Machine Learning Techniques

The aim of this study is to develop a diet recommendation system (DRS) using machine learning techniques for the diagnosis of diabetes and suggesting appropriate diets for diabetic patients. The selection of the proper diet for diabetic patients is based on thorough data analysis [3]. The research uses different machine learning algorithms like Decision tree,

Random Forest and Naive Bayes, along with precision, recall, F measure, and accuracy research methods to find the accuracy and predict the diet recommendation system. The data is pre-processed before applying machine learning algorithms to ensure accurate results in recommending the dietary system [4] [5].

C. A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques

This study focuses on Cardio-Vascular Disease (CVD) which affects the heart and blood vessels. The research uses machine learning techniques and data mining methodologies to predict the occurrence of this disease. The research suggests that Decision Tree and Naive Bayes models perform better since they use categorical data for prediction and classification. After applying various machine learning algorithms, it was found that Support Vector Machine (SVM) had the highest accuracy and Logistic Regression had the lowest [6]. SVM and Decision Tree were identified as the models with the highest accuracy for predicting CVD [7]. The study finds that Decision Tree is an effective method for processing large amounts of data for classification purposes [8].

III. Description of Datasets

A. Dataset 1 - Cardiovascular Diseases

The dataset related to cardiovascular diseases consists of over 70,000 records, and it includes various attributes such as weight, height, smoking, and cholesterol levels. This information helps in identifying the dietary habits that contribute to the development of high cholesterol and various cardiovascular diseases that can lead to severe health issues if not treated properly. [9].

B. Dataset 2 - Diabetes

The Diabetes dataset is a vast repository of health-related information comprising over 250,000 data points of individuals. The data includes vital metrics such as BMI and smoking habits, as well as other crucial aspects required to predict and understand the dietary habits that may lead to the onset of Diabetes. By analyzing this rich and comprehensive data, we can unravel the intricate relationships between various lifestyle factors and chronic diseases such as Diabetes, thus paving the way for more informed and effective interventions and prevention strategies. [10].

C. Dataset 3 - Obesity

The Obesity dataset is an invaluable resource that provides comprehensive information on over 80,000 data points related to an individual's lifestyle choices, such as dietary habits, physical activity levels, and other related factors that contribute to the development of obesity.Through careful examination of the data, we can gain a clear understanding of the factors that contribute to obesity, and in turn, develop effective strategies for prevention and treatment. This dataset represents a significant step forward in our understanding of obesity and provides a wealth of information that can be used to improve public health and well-being. [11].

IV. Methodology

Sedentary lifestyle and unhealthy eating habits are few of the things which an individual can control but that is neglected all the time, having said that, diseases caused due to this bad habits may not accept an persons life in the early stages but surely and definitely affects in the long run and which can lead to more severe disease such as Chronic diseases. As per studies, Chronic diseases are long-lasting and can be fatal if not identified at an early stage. This kind of diseases can be treated and managed but lifestyle changes and proper diet are few of the key aspects which are needed and with proper medication.

To achieve this one should identify if a person has any kind of Chronic disease or are there any symptoms which lead to the possibility of having Chronic diseases. In this paper, we are predicting the probability of one having any Chronic disease based on eating habits, unhealthy lifestyle and lack of physical activity. We are targeting three most common Chronic diseases which are Cardiovascular, Diabetes and Obesity.

We have taken three datasets related to the Chronic diseases which we are addressing, which involves all variables such as smoking, high blood pressure, weight, age and many more such attributes which will help us to predict the possibility of any person having Chronic diseases in later stages of their life.

In this study, we are running 5 Machine Learning models for each dataset and according to the output generated by these models we will predict which model is the best fit among all remaining 4 models. At the final stage after running all the models for all 3 databases, we can predict which 3 models are the best fit or our cause for the 3 databases.

A. Dataset 1 - Cardiovascular Diseases

In this dataset, data was loaded into the Jupyter notebook and cleaning of the data was done which included removal of duplicate values, after cleaning and pre-processing, we checked the data and the attributes present in this dataset. We selected the target variable and performed the prediction by splitting the data into 80 percent and 20 percent configuration, which means 80 percent for training and 20 percent for testing the prediction. Then, we plotted the correlation heat map of variables present in the data set. We checked the Body Mass Index distribution and analysed the data in which 50 percent of people lie in overweight zone.

Logistic Regression was performed on the above dataset and we got an accuracy of 71.75 percent for training data and 71.63 for testing data. As we ran all the remaining models we got to know that the best model for this dataset is the logistic Regression model as can be seen in the below image.

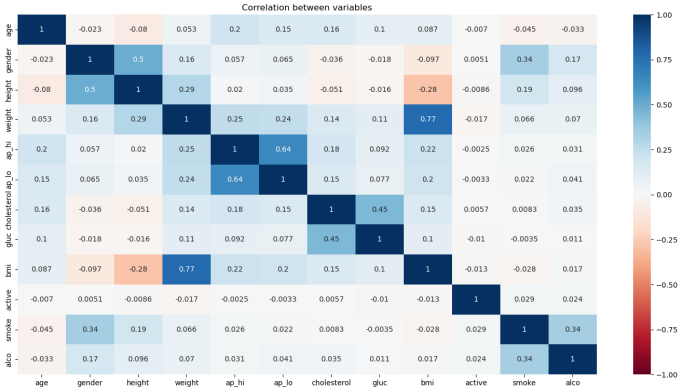

Fig. 1. Correlation Heat Map

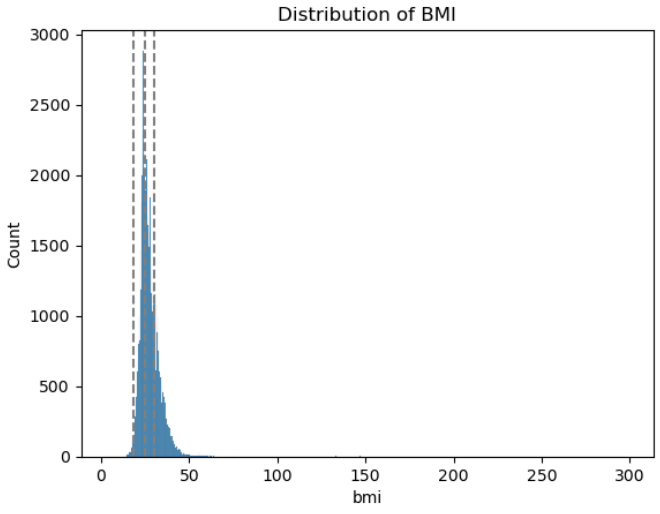

Fig. 2. Body Mass Index(BMI) Distribution

Out[66]:

	train_accuracy	train_precision	train_recall	train_f1	train_roc_auc	test_accuracy	test_precision	test_recall	test_f1	test_roc_auc	model
0	0.717537	0.737382	0.669332	0.701711	0.717190	0.716315	0.730344	0.666074	0.696730	0.715256	Logistic Regression
2	0.999746	0.999890	0.999598	0.999744	0.999745	0.710879	0.706816	0.698963	0.702868	0.710628	Random Forest
4	0.702388	0.751248	0.598657	0.666328	0.701642	0.703196	0.743981	0.599704	0.664096	0.701015	Naive Bayes
1	0.999783	1.000000	0.999562	0.999781	0.999781	0.637240	0.627614	0.635704	0.631633	0.637208	Decision Tree
3	0.514225	0.844523	0.026175	0.050775	0.510713	0.521200	0.830275	0.026815	0.051952	0.510782	Perceptron

Fig. 3. Final Result Dataset 1

Similarly, for all the remaining dataset we ran five models and for the dataset 2 - Diabetes we also got Logistic Regression as the best fit model with accuracy 84.59 percent for training data and 84.56 for the test data.

Out[62]:

	train_accuracy	train_precision	train_recall	train_f1	train_roc_auc	test_accuracy	test_precision	test_recall	test_f1	test_roc_auc	model
0	0.845879	0.529695	0.162377	0.248559	0.567765	0.845573	0.557490	0.169602	0.260081	0.571978	Logistic Regression
2	0.993442	0.994300	0.993745	0.978784	0.981358	0.842341	0.518496	0.207168	0.296049	0.585258	Random Forest
3	0.843134	0.527187	0.007000	0.013816	0.502915	0.839936	0.490000	0.006035	0.011924	0.502419	Perceptron
1	0.993486	0.998628	0.959822	0.978841	0.979788	0.778244	0.320701	0.344993	0.332404	0.602888	Decision Tree
4	0.765576	0.347718	0.563249	0.429987	0.683250	0.770104	0.361880	0.571992	0.443299	0.689919	Naive Bayes

Fig. 4. Final Result Dataset 2

Similarly, for dataset 3 - Obesity we got Gaussian Naive Bayes as the best fit model with accuracy as 63.65 percent for training dataset and 64.31 percent for the testing data.

Out[77]:

	train_accuracy	train_precision	train_recall	train_f1	train_roc_auc	test_accuracy	test_precision	test_recall	test_f1	test_roc_auc	model
4	0.636475	0.524105	0.164596	0.250516	0.538577	0.643103	0.522339	0.163812	0.249407	0.539414	Naive Bayes
0	0.630855	0.000000	0.000000	0.000000	0.499975	0.638031	0.000000	0.000000	0.000000	0.500000	Logistic Regression
3	0.630870	0.000000	0.000000	0.000000	0.499988	0.638031	0.000000	0.000000	0.000000	0.500000	Perceptron
2	0.999984	0.999958	1.000000	0.999979	0.999988	0.631457	0.484761	0.288877	0.362020	0.557343	Random Forest
1	0.999984	1.000000	0.999958	0.999979	0.999979	0.554630	0.387765	0.398028	0.392830	0.520751	Decision Tree

Fig. 5. Final Result Dataset 3

V. Conclusion

As the results suggests, healthy lifestyle, proper balanced diet are the key factors with which we can avoid the initial phase of Chronic Diseases, as predicted physical activity is very important and a good diet because improper lifestyle can definitely affect the person and there is an high probability of causing Chronic Diseases. As predicted lack of physical activity causes a high risk of Obesity and Cardiovascular diseases along with that eating junk foods bad dietary habit can lead to Cholesterol which eventually lead to Diabetes. To avoid such scenarios it is recommended that having a healthy lifestyle and a balanced appropriate diet can reduce the early signs of Chronic Disease.

References

[1] Hyerim Kim, Dong Hoon Lim, and Yoona Kim. Classification and prediction on the effects of nutritional intake on overweight/obesity, dyslipidemia, hypertension and type 2 diabetes mellitus using deep learning model: 4–7th korea national health and nutrition examination survey. *International Journal of Environmental Research and Public Health*, 18(11):5597, 2021.

[2] Fabio Mendoza Palechor and Alexis de la Hoz Manotas. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in brief*, 25:104344, 2019.

[3] Divya Jain and Vijendra Singh. Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, 19(3):179–189, 2018.

[4] Salliah Shafi Bhat and Gufran Ahmad Ansari. Predictions of diabetes and diet recommendation system for diabetic patients using machine learning techniques. In *2021 2nd International Conference for Emerging Technology (INCET)*, pages 1–5. IEEE, 2021.

[5] Xianwen Shang, Yanping Li, Haiquan Xu, Qian Zhang, Ailing Liu, Songming Du, Hongwei Guo, and Guansheng Ma. Leading dietary determinants identified using machine learning techniques and a healthy diet score for changes in cardiometabolic risk factors in children: a longitudinal analysis. *Nutrition journal*, 19(1):1–16, 2020.

[6] Joseph Rigdon and Sanjay Basu. Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the usa using nationally randomly sampled data. *BMJ open*, 9(11):e032703, 2019.

[7] Jingshu Liu, Zachariah Zhang, and Narges Razavian. Deep ehr: Chronic disease prediction using medical notes. In *Machine Learning for Healthcare Conference*, pages 440–464. PMLR, 2018.

[8] M Swathy and K Saruladha. A comparative study of classification and prediction of cardio-vascular diseases (cvd) using machine learning and deep learning techniques. *ICT Express*, 8(1):109–116, 2022.

[9] Cardiovascular Disease dataset, 1 2019.

[10] Diabetes Health Indicators Dataset, 11 2021.

[11] Nutrition, Physical Activity, and Obesity, 2 2023.