ABSTRACT

Diabetes is a chronic metabolic disorder that requires individualized management strategies to achieve optimal patient outcomes. The aim of this project is to develop a personalized health recommendation system based on an individual health collection and predict whether the individual has diabetes or not. By dividing individuals into groups based on their health profile, the system can provide tailored advice specific to the characteristics of each group.The methods used in this project include Label Encoding, Normalization/Standardization, K-Means Clustering, the Elbow Method, Silhouette Analysis, Descriptive Statistics, Rule-based Recommendations.

This project aims to develop a personalized health recommendation system for diabetes management by leveraging machine learning techniques to analyze individual health profiles and predict the presence of diabetes. The system employs Label Encoding to convert categorical variables into numerical values and Normalization/Standardization techniques to scale the data, ensuring balanced contributions from each feature. K-Means Clustering is used to group individuals based on their health profiles, with the optimal number of clusters determined using the Elbow Method and validated through Silhouette Analysis. Descriptive Statistics provide insights into the characteristics of each cluster, enabling the creation of rule-based recommendations tailored to the specific needs of each group. Additionally, a Random Forest classifier is used to predict diabetes based on the health data. The end goal is to implement a system that not only forecasts the likelihood of diabetes but also offers personalized health advice to manage and prevent the condition, thereby enhancing individualized healthcare outcomes. Key terms involved in this project include Elbow Method, K-Means Clustering, Label Encoding, Machine Learning, Random Forest, and Silhouette Analysis.

Index Terms- Elbow Method, K-Means Clustering, Label Encoding, Machine Learning, Random Forest, Silhouette Analysis.

ACKNOWLEDGEMENT

This project involves the collection and analysis of information from various sources and the efforts of many people beyond me.

I sincerely thank the **College Management** for all the infrastructure and facilities provided during my study in the institution.

I take this opportunity to acknowledge my profound gratitude to **Dr. Abdul Kareem**, Principal, MIT Kundapura for his support and encouragement provided during my study in the institution.

I sincerely thank **Mr. Melwin D'Souza**, Vice-Principal, MIT Kundapura for his suggestion and support for completion of the project work.

I would like to place on record my deep sense of gratitude to **Dr. Indra Vijay Singh**, Associative. Professor & HOD, Department of Artificial Intelligence and Machine Learning , MIT Kundapura for his inspiring guidance and insightful comments provided during the project work.

I express my gratitude to all the **teaching and non-teaching staff** of the Artificial Intelligence and Machine Learning department, MIT Kundapura for their support in the completion of this project.

I would like to thank all my **friends and classmates** who directly or indirectly helped me for the successful completion of the project.

Finally, I would like to thank my **parents** for their constant support and encouragement in hundreds of little ways that meant a lot for me.

K DILEEP (4MK21AI006)

SRAJAN K (4MK21AI015)

SWASTHIK (4MK21AI017)

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Diabetes, a prevalent multifaceted metabolic syndrome, poses significant challenges to global health. Diabetes, characterized by chronic hyperglycemia, results from the interplay of genetic, environmental, and lifestyle factors. In this research paper health indicator dataset is used which includes most of the features to predict and cluster them to give recommendations to the individual.

According to the World Health Organization (WHO 2023), diabetes affects 8.5% of adults aged 18 and above. In 2019, it was reported that diabetes became the direct cause of death of 1.5 million people. In low- and middle-income countries, diabetes rates increased by 13% between 2000 and 2019 (WHO 2023). It is estimated that, globally diabetes will affect 642 million people globally by 2040, or one out of every 10 people (Zou et al. 2018). Diabetes, as a chronic metabolic disease, requires special consideration because of its impact on people's well-being and the potential for severe adverse effects on one’s health(American Diabetes Association 2022).(Emmanuel Mbuya et al)

Symptoms of Diabetes

1) Frequent Urination
2) Increased thirst
3) Tired/Sleepiness
4) Weight loss
5) Blurred vision
6) Mood swings

Causes of diabetes

Many factors affect diabetes, and many of them are captured in your data set. Age is an important determinant, and insulin sensitivity and a hereditary decrease in pancreatic beta-cell function increase the risk of developing type 2 diabetes as individuals age Body weight (BMI) is another important factor, as higher BMI values are closely associated with it obolo . Gender differences also play a role; For example, mutations in women may affect insulin sensitivity, whereas men may exhibit patterns of fat distribution that contribute to diabetes risk Lifestyle factors such as smoking status and level of physical activity essential.

Chapter 2

LITERATURE REVIEW

Jyotirani et al[1] studied to develop a more accurate predictive model using multiple machine learning techniques on a dataset obtained from Kaggle, with 2000 data points each with 9 features. The target variable for predicting whether the patient has diabetes (1) or not (0). This study includes data preprocessing with methods including K-nearest neighbors (K-NN), logistic regression, decision tree, random forest and support vector machine (SVM), regression. They began by preprocessing the Pima Indians Diabetes dataset, dealing with missing values, and normalizing the data to ensure it was ready for analysis and then using PCA to reduce the size of the dataset on, and factors that retained significant differences were identified. The simplification of data structures at this stage made it easier to identify good initial centers for K-means clustering by reducing loss of information Using the central points from PCA, researchers assigned initial clusters the effects were good, which is generally sensitive to the choice of initial mean points.

Raj Keshav et al[2] , conducted a study on the development of scientific research. Rathore, Ashutosh D. Tathe, Tejaswini S. Pawar and Piyush S. Mahajan developed diabetes prediction algorithm Random forest classifier. Their research addresses a major global health challenge posed by diabetes, which, if not recognized and addressed early, can lead to serious complications affecting correlation analysis to understand feature relationships, and model training for each algorithm. The performance of each model is evaluated on training and testing accuracy, which shows that the decision tree algorithm achieves the highest accuracy with 100% training accuracy and 99% testing accuracy.

Yazan Jian et al[3] investigated the use of supervised machine learning algorithms to predict diabetes-related complications. The study used the Rashid Diabetes Research Center dataset in Ajman, UAE, which had record 884 types including 79 items were used. Prognostic challenges noted included metabolic changes, cholesterol, arthritis, asthma, diabetic foot, high blood pressure, obesity, and retinopathy The researchers used several necessary preprocessing steps to prevent missing values and data imbalance problems. Available feature selection methods were used to identify surface features for each complex. Machine learning algorithms including logistic regression, support vector machine, decision tree, random forest, AdaBoost, and XGBoost were used to generate binary classifiers for each challenge. These models were evaluated using iteration-level kfold cross-validation yields performance, including accuracy and F1-score production as the primary measure.

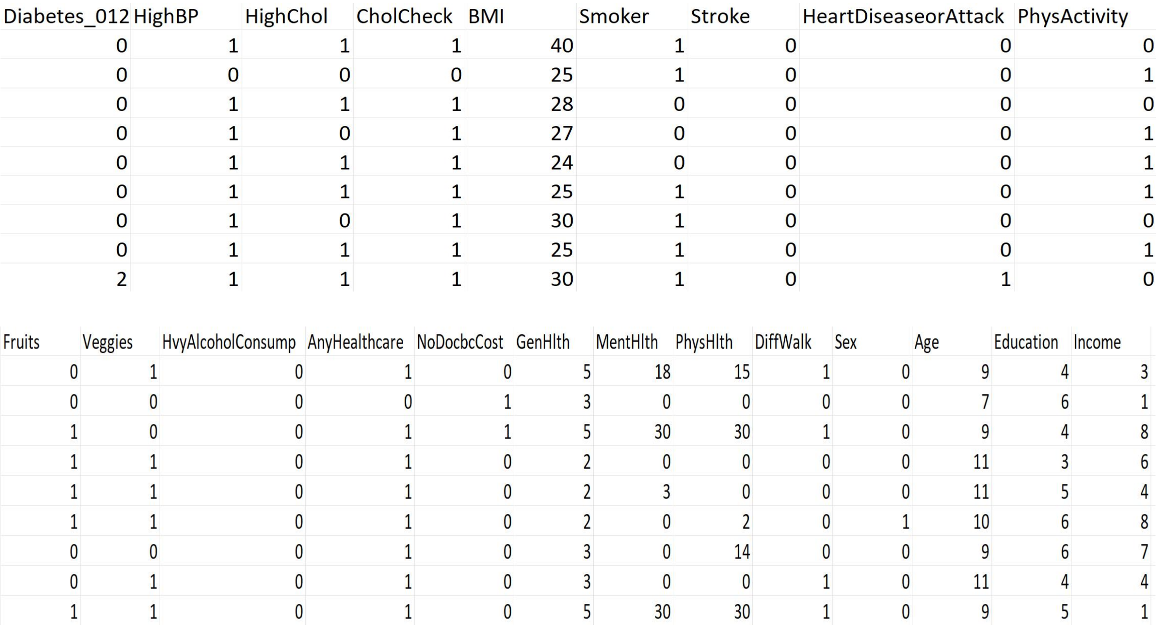
C. Zhu et al [4] conducted a study that aimed to increase the accuracy of diabetes prediction models by incorporating principal component analysis (PCA) and K-means clustering into logistic performance by developing hyperparameters Data vital organs. The researchers used machine learning to analyze medical data and predict the probability of prediabetes in individuals. Using a large dataset from Kaggle, the data were preprocessed, corrected for missing values, and normalized to ensure consistency. The main objective of the study is to implement an end-to-end diabetes prediction solution using a random forest system known to be difficult in the classification industry. The researchers divided the dataset into training and testing and also improved the model.

Chapter 3

METHODOLOGY

In this research paper, we aim to predict diabetes in individuals and provide suitable recommendations based on their health indicators. We have carried out data collection, data preprocessing and cleaning, training and testing of the model for prediction, and clustering the data to group individuals into clusters. Based on these clusters, we will provide appropriate recommendations.

The information in the dataset covers important parts of people's daily habits. It shows if people smoke, how much they exercise, and how much fruit and vegetables they eat. These details are important for understanding how these habits affect the chance of getting diabetes. The dataset also keeps track of people's health history, like if they have high blood pressure or high cholesterol, which are also risks for diabetes. It also notes if someone has had a stroke or heart disease, giving a more complete view of their heart health.



Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity
0	1	1	1	40	1	0	0	0
0	0	0	0	25	1	0	0	1
0	1	1	1	28	0	0	0	0
0	1	0	1	27	0	0	0	1
0	1	1	1	24	0	0	0	1
0	1	1	1	25	1	0	0	1
0	1	0	1	30	1	0	0	0
0	1	1	1	25	1	0	0	1
2	1	1	1	30	1	0	1	0

Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1	0	1	0	5	18	15	1	0	9	4	3
0	0	0	0	1	3	0	0	0	0	7	6	1
1	0	0	1	1	5	30	30	1	0	9	4	8
1	1	0	1	0	2	0	0	0	0	11	3	6
1	1	0	1	0	2	3	0	0	0	11	5	4
1	1	0	1	0	2	0	2	0	1	10	6	8
0	0	0	1	0	3	0	14	0	0	9	6	7
0	1	0	1	0	3	0	0	1	0	11	4	4
1	1	0	1	0	5	30	30	1	0	9	5	1

Figure 3.1 CSV FILE (Diabetes_health_indicators)

The dataset "diabetes_012_health_indicators_BRFSS2015.csv" used in this study is obtained from kaggle website.The dataset contains 21 features containing information on various health-related factors such as: Age, BMI (a measure of body fat based on height and weight), Sex, whether someone smokes, HighBP (if they have high blood pressure), HighChol (if they have high cholesterol), CholCheck (if they've had their cholesterol checked), Stroke (if they've had a stroke), HeartDiseaseorAttack (if they've had heart disease or a heart attack), PhysActivity (if they engage in physical activities), Fruits (if they eat fruits), Veggies (if they eat vegetables), HvyAlcoholConsump (if they consume a lot of alcohol), AnyHealthcare (if they have access to healthcare), NoDocbcCost (if they haven't seen a doctor due to cost), GenHlth (their overall health), MentHlth (their mental health), PhysHlth (their physical health), DiffWalk (if they have difficulty walking), Education (their education level), and Income (their income level).

The diabetes _ 012 _ health _ indicators _ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes.

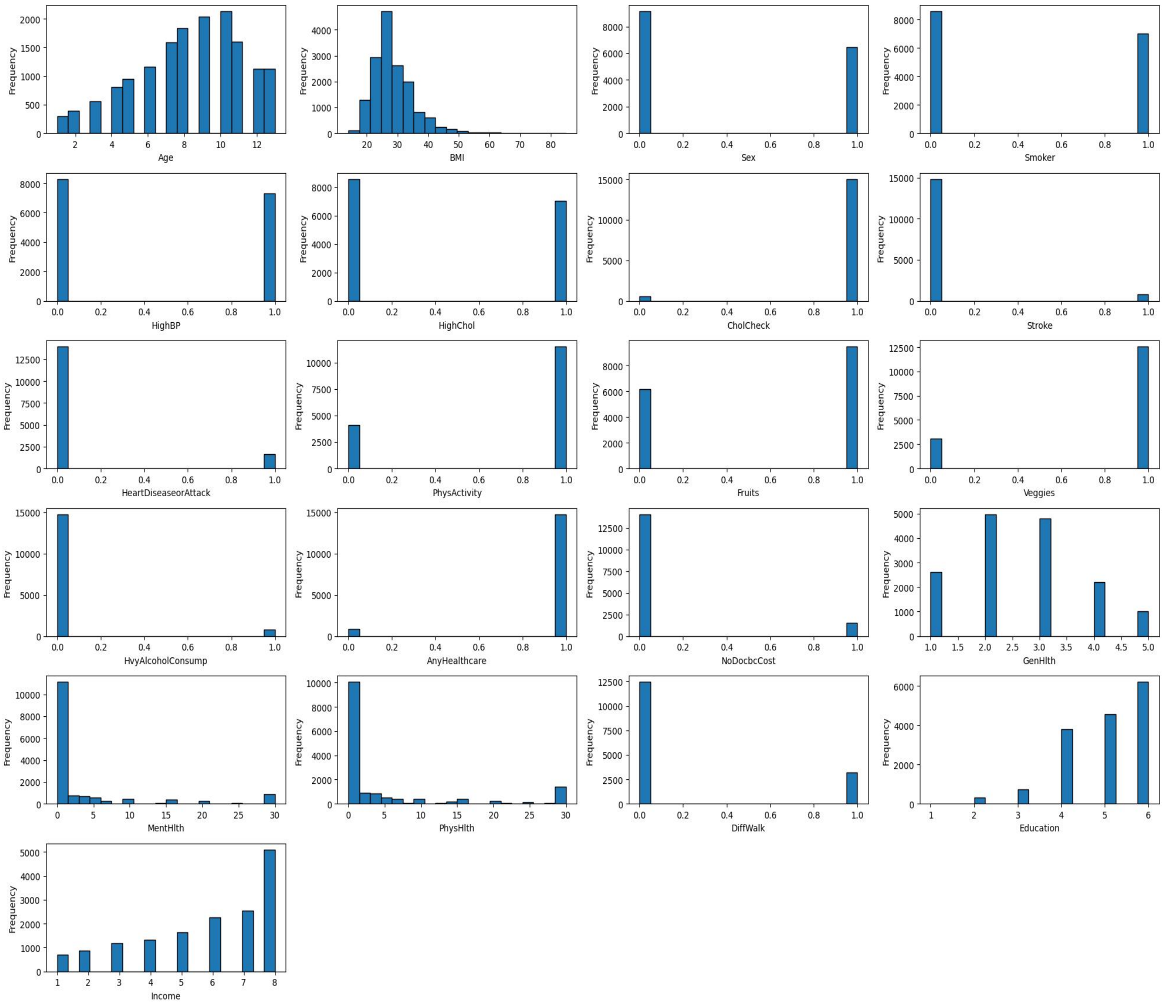


Figure 3.2 Distribution of various features in the dataset

S.No	Attribute	Description	Values
1	Age	Patient's age in years	Continuous value
2	BMI	Body Mass Index	Continuous value
3	Sex	Patient's sex	0: Female, 1: Male
4	Smoker	Whether the patient is a smoker	0: No, 1: Yes
5	HighBP	Whether the patient has high blood pressure	0: No, 1: Yes
6	HighChol	Whether the patient has high cholesterol	0: No, 1: Yes
7	CholCheck	Whether the patient has had their cholesterol checked	0: No,1:Yes
8	Stroke	Whether the patient has had a stroke	0: No, 1: Yes
9	HeartDiseaseorAttack	Whether the patient has had heart disease or a heart attack	0: No, 1: Yes
10	PhysActivity	Whether the patient engages in regular physical activity	0: No, 1: Yes
11	Fruits	Whether the patient regularly consumes fruits	0: No, 1: Yes
12	Veggies	Whether the patient regularly consumes vegetables	0: No, 1: Yes
13	HvyAlcoholConsump	Whether the patient engages in heavy alcohol consumption	0: No, 1: Yes
14	AnyHealthcare	Whether the patient has access to any healthcare services	0: No, 1: Yes
15	NoDocbcCost	Whether the patient has avoided visiting a doctor due to cost	0: No, 1: Yes
16	GenHlth	Patient's general health statusOrdinal scale (1-5)	Where 1 is excellent
17	MentHlth	Number of days the patient experienced poor mental health in the past month	Continuous value (0-30)
18	PhysHlth	Number of days the patient experienced poor physical health in the past month	Continuous value (0-30)
19	DiffWalk	Whether the patient has difficulty walking	0: No, 1: Yes
20	Education	Highest level of education attained by the patient	Ordinal scale (1-6)
21	Income	Patient's income level	Ordinal scale (1-8)
22	Diabetes_012	Diabetes status of the patient	0: No diabetes, 1: Prediabetes, 2: Diabetes

Table 3.1: Description of Features

Chapter 4

IMPLEMENTATION

Our personalized diabetes prediction and management system starts with the "Diabetes Health Indicators" dataset, which is comprehensive and includes numerous samples with detailed health and lifestyle information. This dataset features 21 attributes, including Age, BMI (Body Mass Index), Sex, Smoker status, High Blood Pressure (HighBP), High Cholesterol (HighChol), Cholesterol Check (CholCheck), Stroke history, Heart Disease or Heart Attack (HeartDiseaseorAttack), Physical Activity (PhysActivity), and consumption of Fruits and Vegetables (Fruits and Veggies). Other factors are Heavy Alcohol Consumption (HvyAlcoholConsump), access to Healthcare (AnyHealthcare), and whether Cost prevents seeing a Doctor (NoDocbcCost). Additionally, it includes General Health (GenHlth), Mental Health (MentHlth), Physical Health (PhysHlth), Difficulty Walking (DiffWalk), Education level, and Income. These features offer critical insights into an individual's health and lifestyle, essential for accurate diabetes prediction.The preprocessing phase is a critical step in transforming raw data into a format that is optimal for machine learning algorithms. The initial task involves encoding categorical variables and scaling continuous variables, ensuring they are on a comparable scale. This adjustment is crucial because it enhances the performance of machine learning models. Handling missing values is another important aspect of preprocessing; imputation techniques are used to fill in gaps, thereby creating a complete dataset ready for model training.

Once preprocessing is complete, the data is divided into training and testing sets, which are essential for model training and evaluation. The training set is employed to train a Random Forest Classifier. This machine learning algorithm is renowned for its accuracy and efficiency, especially when managing large datasets. The model is trained to predict whether an individual has diabetes based on 21 different features.After the model has been trained, its performance is assessed using the test set. Various metrics, including accuracy, classification report, and confusion matrix, are generated to evaluate its effectiveness. These metrics offer valuable insights into how well the model can correctly classify individuals as diabetic or non-diabetic. This evaluation is vital to ensuring the system's reliability and effectiveness in real-world applications.To boost the personalization capabilities of the system, K-Means Clustering is employed on the processed data. This unsupervised learning technique categorizes individuals with similar health profiles into distinct clusters. To identify the optimal number of clusters, the Elbow

Method and Silhouette Analysis are utilized. Each resulting cluster represents a group of individuals sharing similar health characteristics, enabling the system to provide tailored recommendations.Principal Component Analysis (PCA) serves as a tool to visualize clusters by reducing data dimensionality, making it possible to represent these clusters clearly in a two-dimensional space. This technique aids in comprehending how individuals are distributed across various health profiles and assessing the effectiveness of the clustering algorithm used.

Specific health recommendations are crafted for each cluster to help individuals manage their diabetes risk more effectively. These suggestions stem from common characteristics shared by people within each cluster. For instance, one group might benefit from increasing physical activity and improving diet quality, while another might need to focus on weight management and cutting down on alcohol consumption.To enhance user-friendliness, a function has been developed that takes individual health and lifestyle information to provide personalized diabetes predictions and recommendations. This function preprocesses the input data, employs a trained Random Forest model to predict diabetes risk, and utilizes the K-Means model to identify the appropriate cluster. Based on this cluster, tailored health recommendations are provided, offering actionable steps for managing diabetes risk.

Our advanced diabetes prediction and management system utilizes machine learning and clustering methods to deliver precise diabetes forecasts and customized health advice. By harnessing a rich dataset and employing strong algorithms such as the Random Forest Classifier for predictions and K-Means Clustering for tailored recommendations, the system supports individuals in effectively managing their diabetes risk, promoting improved health outcomes.

Chapter 5

RESULT

The performance of a personalized diabetes prediction and management system was assessed using a Random Forest Classifier, achieving an overall accuracy of about 82.7%. This means the model correctly predicted the diabetes status for roughly 83 out of every 100 cases. However, a closer inspection of the classification report reveals discrepancies in performance across different classes. The model showed high precision (0.85) and recall (0.96) for class 0 (no diabetes), indicating it effectively identifies individuals without diabetes. On the other hand, the model struggled with class 1 (prediabetes), achieving a precision and recall of 0.62, and class 2 (diabetes), where the precision was 0.51 and recall was 0.21. These figures suggest that, while the model is quite effective in predicting non-diabetic individuals, it faces significant challenges in accurately predicting prediabetes and diabetes. The confusion matrix further underscores these issues, showing that most misclassifications occur in class 2, with many individuals being incorrectly classified as non-diabetic. Despite the high overall accuracy, these findings indicate a need for further refinement to improve the model's sensitivity and specificity, particularly for identifying individuals at higher risk of diabetes.

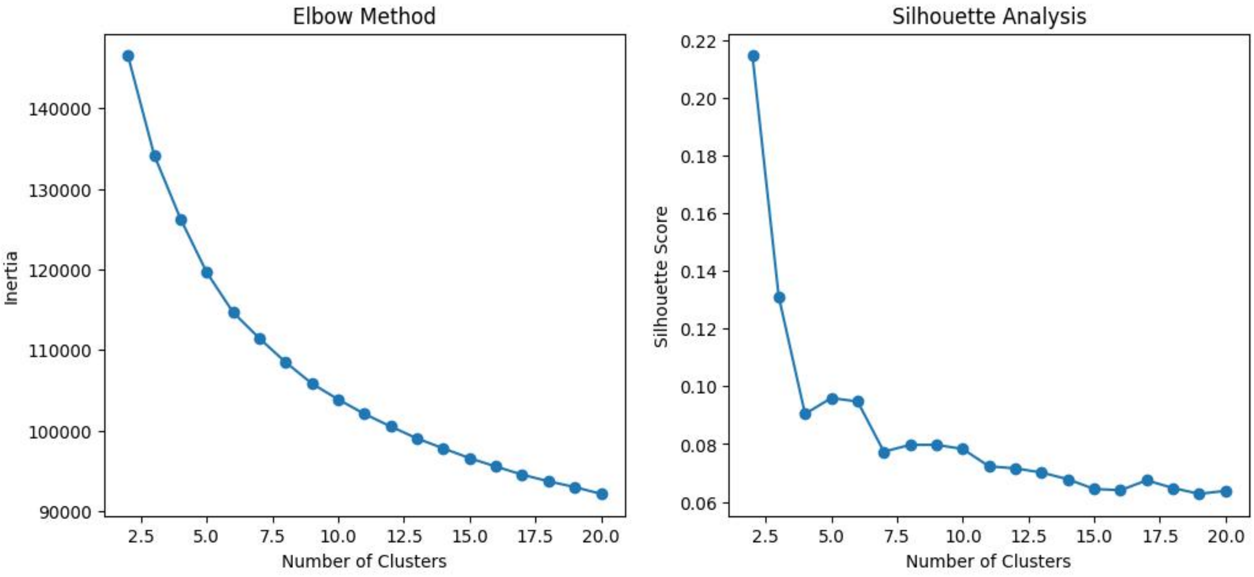
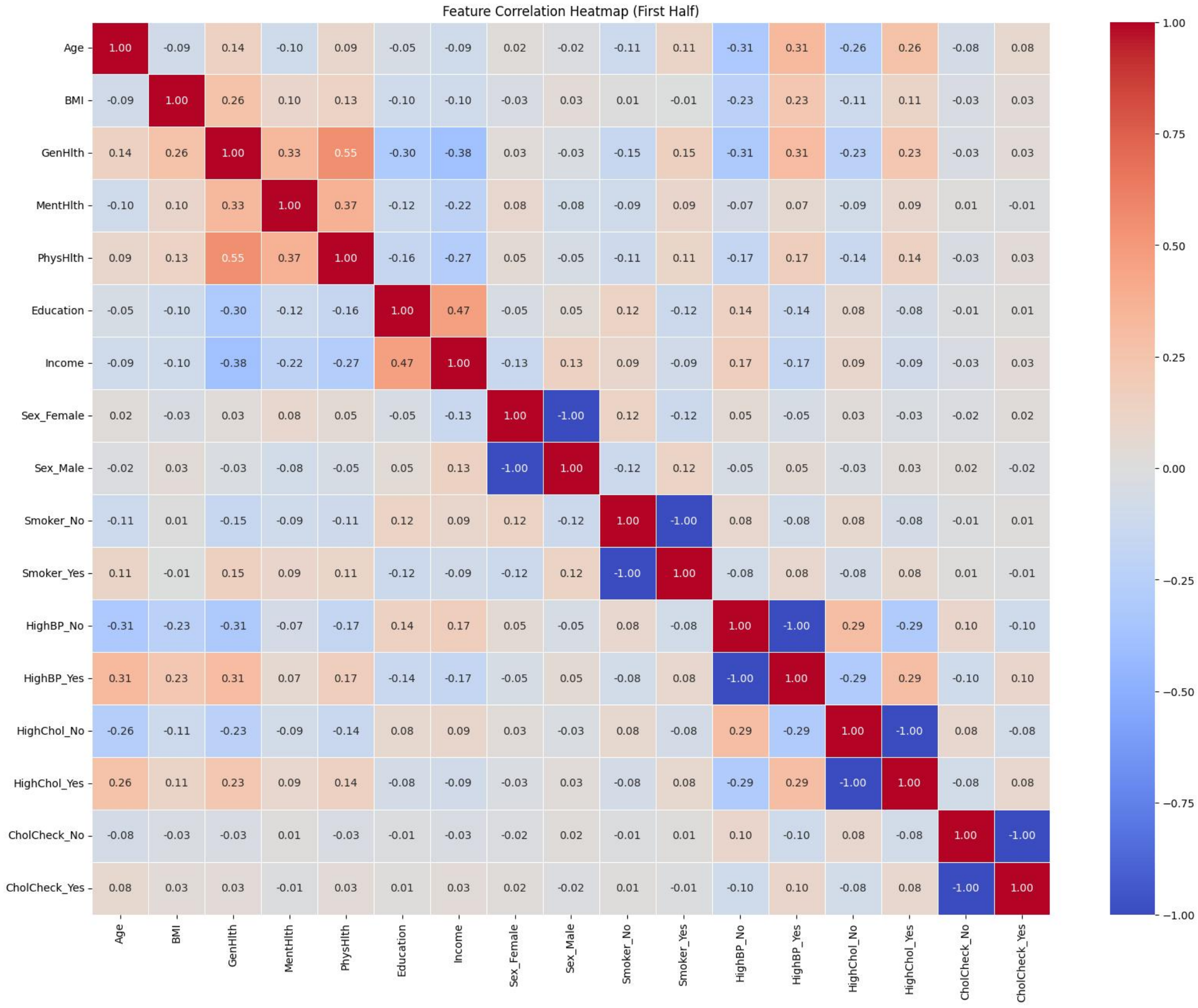


Figure 5.1: Optimal number of clusters determined using Elbow Method and Silhouette Analysis.

The Figure 5.1 represents the results of the Elbow Method and Silhouette Analysis, both utilized to determine the optimal number of clusters for K-Means clustering. The Elbow Method plot (left) shows the inertia (sum of squared distances) against the number of clusters, with a notable decrease in inertia initially that begins to level off around 4 clusters, suggesting this as the optimal number. This "elbow" point indicates that adding more clusters beyond this does not significantly reduce the inertia, making 4 clusters a reasonable choice. The Silhouette Analysis plot (right) measures how similar an object is to its own cluster compared to other clusters, with scores ranging from -1 to 1. The plot indicates the highest silhouette score around 2 clusters, but there is also a noticeable peak around 4 clusters, implying well-defined clusters. Combining these insights, both methods point towards 4 clusters as an optimal choice, balancing minimized inertia and well-defined clusters, thereby guiding the clustering approach for personalized health recommendations in the diabetes management system.



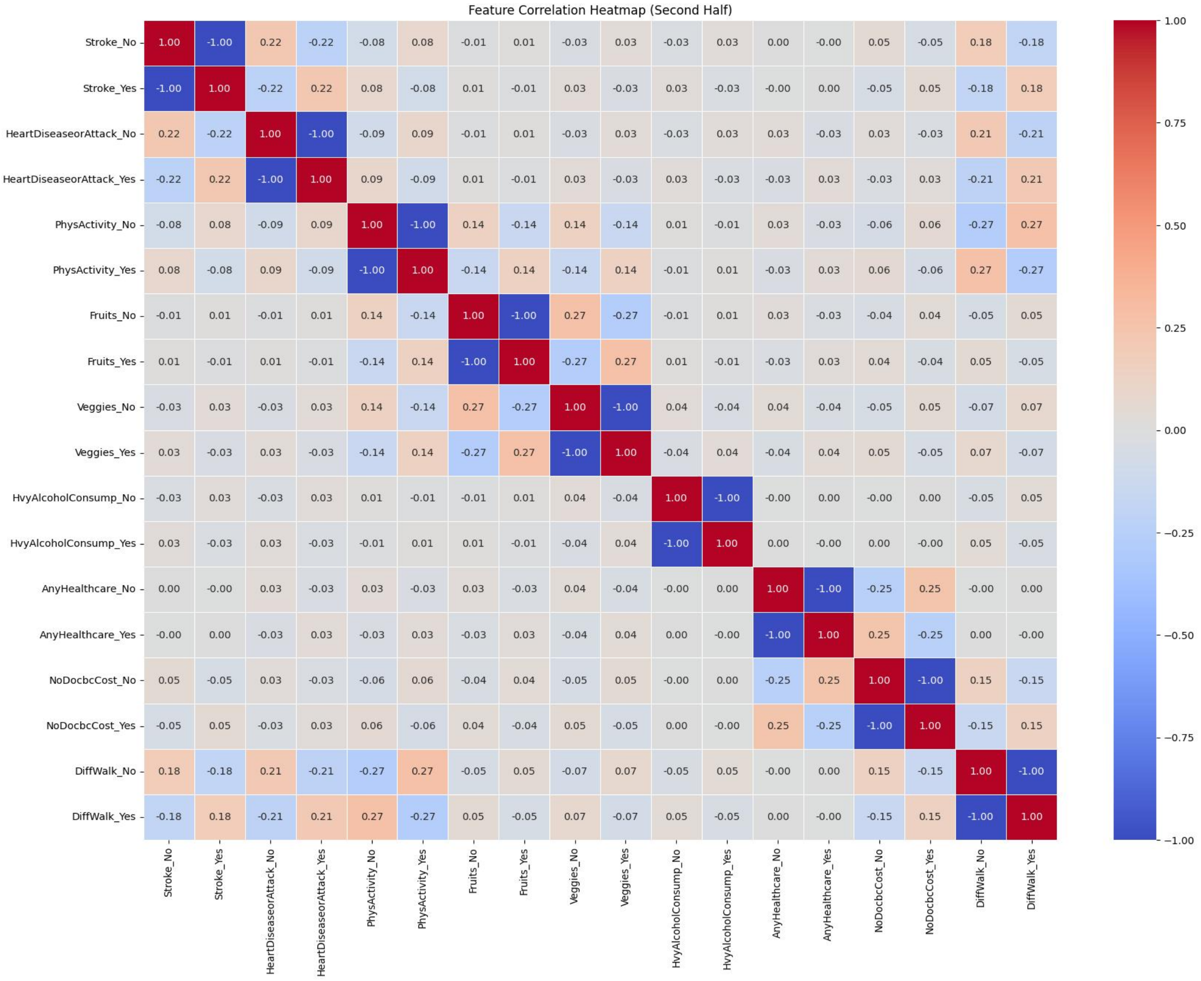


Figure 5.2: Feature correlation heatmap illustrating relationships between health, demographic, and lifestyle variables

The feature correlation heatmap visually represents the correlation coefficients between various features in a dataset, using a color scale from red (indicating a strong positive correlation) to blue (indicating a strong negative correlation), with white representing no correlation. Key observations include strong positive correlations such as those between binary features like Smoker_Yes and Smoker_No, which are perfectly negatively correlated as expected. Strong negative correlations are seen between factors like Genhlth and PhysActivity_No, suggesting that worse general health is associated with lack of physical activity. Age shows moderate positive correlations with high blood pressure, high cholesterol, and heart disease, indicating that these conditions are more common with increasing age. Higher BMI is positively correlated with higher blood pressure and cholesterol. Income and education are positively correlated, indicating that higher education levels are associated with higher income. General trends show that many health conditions are interrelated, with worse health outcomes linked to a lack of physical activity. Binary variables naturally exhibit strong negative correlations between their opposites. For example, there is a moderate correlation between general health and mental health, suggesting that worse general health is associated with worse mental health. Higher education levels are also associated with increased physical activity. This heatmap highlights potential areas for further analysis or intervention, such as promoting physical activity to improve various health outcomes and addressing the interplay between socioeconomic factors and health.

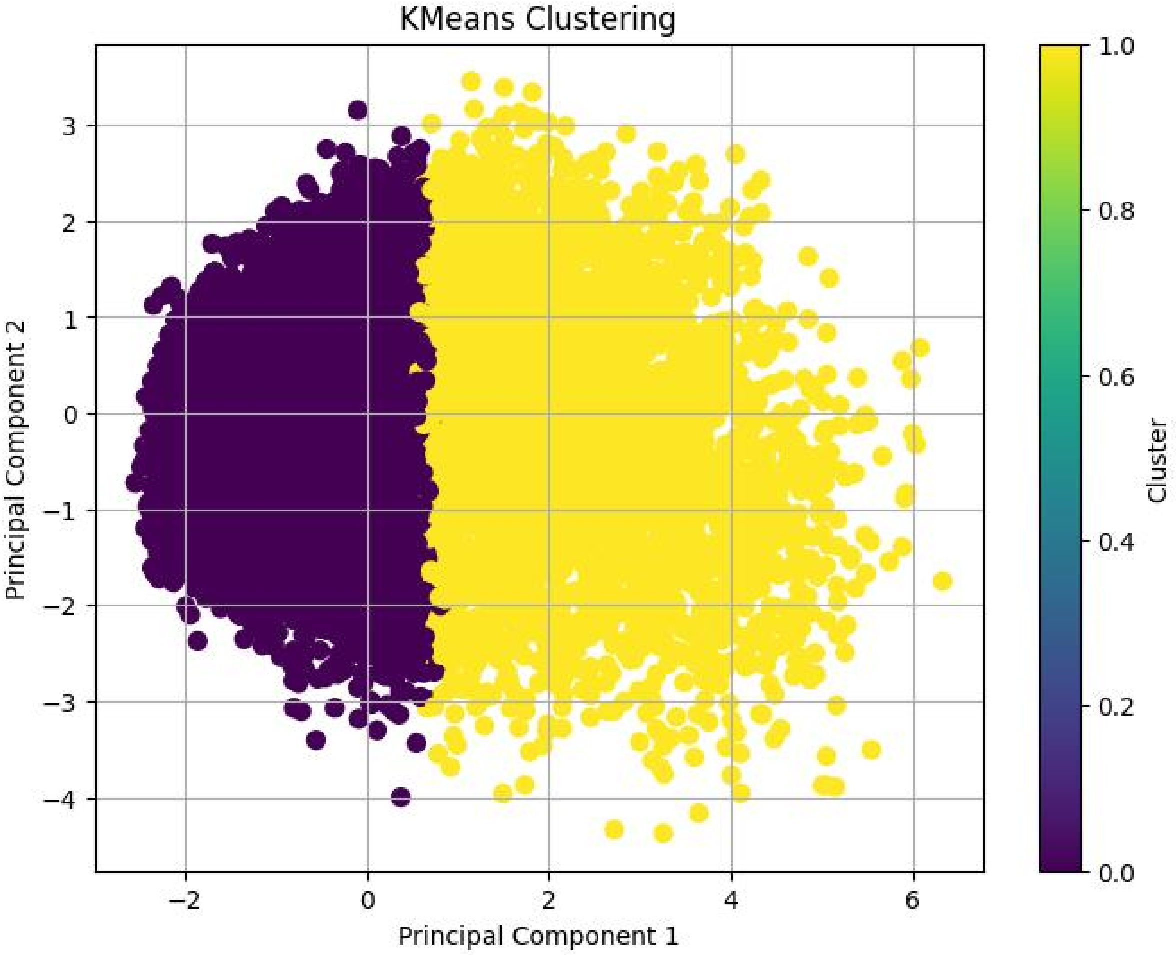


Figure 5.3: KMeans clustering results visualized using principal component analysis, showing two distinct clusters.

The provided image is a scatter plot illustrating the results of KMeans clustering on a dataset, visualized using two principal components. The x-axis represents the first principal component, while the y-axis represents the second principal component. Each data point in the scatter plot corresponds to an observation in the dataset, projected onto the two principal components for easier visualization. The points are colored according to their cluster assignment, with two distinct clusters shown in yellow and purple.

The KMeans algorithm has successfully divided the data into two clusters, with the yellow cluster predominantly occupying the right side of the plot and the purple cluster predominantly occupying the left side. This clear separation suggests that the two clusters are well-defined within the feature space represented by the principal components. The color bar on the right side of the plot indicates the cluster labels, ranging from 0 (purple) to 1 (yellow).

The principal component analysis (PCA) has effectively reduced the dimensionality of the dataset, allowing for a more straightforward interpretation of the clustering results. The distinct boundary between the two clusters implies that the features used in the PCA and KMeans clustering have significant discriminative power, enabling the algorithm to distinguish between the two groups effectively. This visualization helps in understanding the structure and distribution of the data, and it can be useful for further analysis or validation of the clustering process.

Chapter 6

CONCLUSION

The personalized diabetes management and recommendation system developed in this project marks a notable advancement in leveraging machine learning to improve healthcare outcomes. Utilizing a Random Forest classifier, the system achieves an overall accuracy of 82.7% in predicting diabetes status, underscoring its reliability in identifying individuals without diabetes. This makes it a valuable tool for large-scale screening initiatives.This shortcoming points to the need for further refinement, particularly in addressing class imbalance and enhancing feature selection to boost the model's performance across all categories.

Beyond prediction, the system employs a K-Means clustering algorithm to categorize individuals based on their health profiles. This clustering enables the generation of personalized health recommendations tailored to the distinct characteristics of each group. By identifying common risk factors within each cluster, the system offers actionable and customized guidance aimed at mitigating diabetes risk and promoting better health management. These personalized recommendations are crucial for empowering individuals to take proactive steps in managing their health.

In conclusion, this personalized diabetes management and recommendation system showcases the potential of machine learning in healthcare. It not only delivers accurate diabetes predictions but also provides individualized health recommendations, thereby contributing to improved health outcomes. As the system continues to evolve, it promises to become an indispensable tool for diabetes prevention and management, ultimately enhancing the quality of life for individuals at risk.

REFERENCES

[1] KM Jyoti Rani, "Diabetes Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 294-305, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206463 Journal URL : http://ijsrcseit.com/CSEIT206463 .

[2] Raj Keshav, Jitendra I. Rathod, Ashutosh D. Tathe, Tejaswini S. Pawar,Piyush S. Mahajan,Diabetes prediction system using random forest classifier,ISSN: 2582-5208,Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com.

[3] Jian, Y.; Pasquier, M.;Sagahyroon, A.; Aloul, F. A MachineLearning Approach to PredictingDiabetes Complications. Healthcare2021, 9, 1712. https://doi.org/10.3390/healthcare9121712Academic Editors: Keun Ho Ryu andNipon Theera-UmponReceived: 27 October 2021Accepted: 4 December 2021Published: 9 December 2021.

[4] https://doi.org/10.1016/j.imu.2019.100179,Received 20 January 2019; Received in revised form 27 March 2019; Accepted 4 April 2019

[5] E-mail addresses: Zhucs_2008@163.com (C. Zhu), xtianidemudia@yahoo.co.uk (C.U. Idemudia), 1036784024@qq.com (W. Feng).Informatics in Medicine Unlocked 17 (2019) 100179,Available online 05 April 2019,2352-9148/ © 2019 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license,(http://creativecommons.org/licenses/by-nc-nd/4.0/).

[6] Berk OZTURKa,1, Tom LAWTONa,b, Stephen SMITHa, Ibrahim HABLIa,Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning,E-mail addresses: berk.ozturk@york.ac.uk (B. Ozturk), tom.lawton@bthft.nhs.uk (T. Lawton),stephen.smith@york.ac.uk (S. Smith), ibrahim.habli@york.ac.uk (I. Habli).

[7] Kayla Esser1∗ , Monica Duong1∗ , Khalil Kain1∗ , Son Tran1∗ , Aryan Sadeghi2, Aziz Guergachi2,3,Karim Keshavjee2, Mohammad Noaeen1, and Zahra Shakeri2,Predicting Diabetes in Canadian Adults Using Machine Learning,https://doi.org/10.1101/2024.02.03.24302302.

[8] Mbuya, Emmanuel; Mokheleli, Tsholofelo; Bokaba, Tebogo; and Ndayizigamiye, Patrick, "A MulticlassApproach to Predicting Diabetes Using Machine Learning" (2023). ACIS 2023 Proceedings. 140,https://aisel.aisnet.org/acis2023/140.