Predictive Health Assessment: Leveraging Machine Learning to Gauge Lifestyle Impacts

TEAM NUMBER 21

FINAL PROJECT REPORT

TEAM MEMBERS:

Gurwinder Kaur:	**50363756**
Dishank Jagadeeshnaidu Karampudi:	**50559618**
Rishi Shanthan:	**50560689**
Manasa Lakhsmi Gunampalli:	**50559593**

Introduction:

Our project, titled "Predictive Health Assessment: Leveraging Machine Learning to Gauge Lifestyle Impacts," integrates multiple health indicators and lifestyle factors into a comprehensive health index to classify individuals into high, medium, and low-risk categories. Traditional health assessment methods often focus on isolated metrics, failing to provide a holistic view of an individual's health status. These conventional approaches lack the ability to synthesize various data points into unified, actionable insights, limiting their effectiveness in predicting future health risks and conditions. Our innovative approach utilizes the k-Means clustering algorithm to overcome these limitations, offering a more nuanced and predictive framework. k-Means is well-suited for this task as it groups data points into clusters, capturing complex relationships between health indicators and lifestyle factors. Unlike linear and logistic regression, which assume a specific form for the relationship between variables, k-Means does not make such assumptions and adapts to the underlying data distribution, providing more accurate and personalized health risk classifications. This project is significant as it meets the needs of healthcare providers, insurers, policymakers, and individuals, who require more accurate and comprehensive tools for health risk assessment. By enhancing predictive accuracy, our model aims to facilitate personalized healthcare interventions and inform public health strategies effectively.

Objective/ What are we trying to do:

Our goal is to develop a predictive model using the k-Means clustering algorithm to estimate a health index score from various health indicators and lifestyle factors. Additionally, we will employ clustering techniques to classify these scores into risk categories: high, medium, and low.

Literature Survey:

Our literature survey critically examines a range of studies that lay the groundwork for our approach to predictive health assessment, particularly focusing on the integration of various health indicators and lifestyle factors into a comprehensive health index. Notable studies include "Classification of the factors for smoking cessation using logistic regression, decision tree & neural networks," which explores smoking cessation factors through machine learning models. This study provides a foundational framework for analyzing lifestyle impacts which we aim to expand beyond smoking to incorporate additional lifestyle factors. Similarly, "Comparison of Logistic Regression and Machine Learning Methods for Predicting Acute Coronary Syndrome" has been instrumental in shaping our understanding of logistic regression in healthcare, guiding our modeling for more comprehensive health predictions. Further, "Quantitative Analysis Using Multinomial Logistic Regression for Health Data" supports our use of multinomial logistic regression for categorizing health risks, emphasizing the need for applications to real-world data. "Determination of Smoking and Obesity as Periodontitis Risks Using the Classification and Regression Tree Method" demonstrates how decision trees can identify lifestyle-related risk factors, a methodology we extend to a wider range of conditions. Foundational texts such as "An Introduction to Regression Analysis" and "Use of Statistical Models for Health Metrics" provide essential statistical techniques and models, enhancing our project's analytical rigour. We address the challenge of small datasets, as discussed in "Handling Small Datasets in Medical Research," by aiming to use larger, more comprehensive datasets in our analysis. Studies like "Effects of Alcohol and Tobacco Use on Aerodigestive Cancers" and "The Association Between Cigarette Smoking and Ocular Diseases" show the impact of lifestyle factors on specific health conditions, which we generalize to broader health indices. Similarly, the "Relation between consumption of alcohol and fatty acids esterifying serum cholesterol in healthy men" provides insights into cholesterol impacts from lifestyle factors, encouraging a broader demographic analysis in our study. Through these literature insights, our project introduces an innovative approach by synthesizing diverse health metrics into a singular, actionable health index, leveraging advanced machine learning techniques to provide nuanced predictions of health risks, thereby setting a new standard in predictive health analytics. This survey not only validates our methodological choices but also highlights the innovative potential of our project to transform the landscape of health risk assessments.

Proposed Method:

Intuition — Why Should K-MEANS Be Better Than the State of the Art?

The k-Means clustering algorithm offers several advantages over traditional regression techniques such as multiple linear regression and logistic regression, making it more suitable for our health risk prediction task:

- Non-Linearity Handling: Health data often exhibits complex, non-linear relationships among various health metrics. Unlike linear and logistic regression, which assume linear relationships, k-Means makes no such assumptions and can naturally capture non-linear patterns.
- Simplicity and Adaptability: k-Means is a simple, centroid-based clustering algorithm that adapts well to the structure of the data. This makes it robust against the assumptions of data distribution and model form, which are often violated in real-world health datasets.
- Flexible Classification: For multi-class classification tasks such as predicting high, medium, and low health risks, k-Means can provide more flexible cluster formation compared to logistic regression, which relies on linear decision boundaries.
- Intuitive Interpretability: The clustering process in k-Means is straightforward—based on the proximity to cluster centroids. This transparency can be advantageous for interpretability, especially in the sensitive context of health predictions.

Detailed Description

Data Collection: We collect a comprehensive dataset including health metrics such as blood pressure, cholesterol levels, BMI, dietary habits, and physical activity levels.

Preprocessing: We handle missing values by imputation, normalize the data to ensure all features are on a similar scale, and remove outliers to enhance the quality of the analysis.

Feature Engineering :

- Label Encoding: We label encode categorical variables that have an ordinal relationship. For instance, we encode the SMK_stat_type_cd column (smoking status) as follows: Low = 1, Medium = 2, High = 3. This encoding preserves the ordinal nature of the smoking status.

- One-Hot Encoding: We apply one-hot encoding to categorical variables without an ordinal relationship. For example, variables such as DRK_YNN and DRK_YNY (drinking status) are one-hot encoded, transforming each category into a separate binary column.
- Interaction Terms: We create interaction terms to capture synergistic effects between different lifestyle factors and health outcomes.
- Principal Component Analysis (PCA): PCA is applied to reduce the dimensionality of the dataset while retaining the most relevant variance, improving the efficiency of the k-Means algorithm**.**

Model Validation and Selection Cross-Validation: To avoid overfitting, we use cross-validation techniques to ensure the model generalizes well on unseen data.

Performance Metrics: We assess the model performance using accuracy, precision, recall, and F1-score, providing a comprehensive evaluation of the classification efficacy.

- Hyperparameter Tuning: Grid search and random search methods are used to optimize the number of clusters (k) and the initialization method, enhancing the predictive accuracy and efficiency of the k-Means model.
- Model Interpretability: Despite being a clustering model, k-Means can be made interpretable by visualizing the clusters and centroids, helping in understanding the grouping of data points and the decision-making process.

Experiments:

Testbed Description

Our testbed consists of a comprehensive dataset that includes various health metrics such as blood pressure, cholesterol levels, BMI, dietary habits, and physical activity levels. The dataset is pre-processed to handle missing values, normalize data, and remove outliers, ensuring quality and reliability. We split the dataset into training and testing sets for model development and evaluation.

Questions Addressed by the Experiments: Accuracy of KNN in Predicting Health Risk Levels:

- How accurately does the k-Means algorithm predict the health risk levels compared to the actual labels in the test set?
- Effectiveness of Feature Engineering Techniques: How do label encoding and one-hot encoding of categorical variables impact the performance of the k-Means model?
- Optimal Number of Clusters (k) for k-Means: What is the optimal value of k for the k-Means algorithm in this context?
- Comparison with Baseline Models: How does k-Means perform compared to traditional regression models like multiple linear regression and logistic regression in predicting health risk levels?
- Interpretability of the k-Means Model: How interpretable is the k-Means model, and what insights can be derived from the cluster centroids and assignments?

Detailed Description of the Experiments

Experiment 1: Accuracy of k-Means in Predicting Health Risk Levels

- Description: We evaluate the performance of the k-Means algorithm on the test set, comparing the predicted health risk levels with the actual labels.
- Method: Train the k-Means model on the training set. Predict health risk levels on the test set. Calculate accuracy, precision, recall, and F1-score.
- Observations: The k-Means model achieved an accuracy of X%, with precision, recall, and F1-scores indicating that it effectively differentiates between high, medium, and low health risks..

Experiment 2: Effectiveness of Feature Engineering Techniques

- Description: We assess the impact of label encoding and one-hot encoding on the model's performance.
- Method: Implement label encoding for SMK_stat_type_cd. Implement one-hot encoding for DRK_YNN and DRK_YNY. Compare model performance with and without these encodings.
- Observations: Label encoding and one-hot encoding improved model accuracy by Y%, highlighting the importance of appropriately handling categorical variables.

Experiment 3: Optimal Number of Clusters (k) for k-Means

- Description: We determine the optimal value of k for the k-Means algorithm by testing various values and evaluating model performance.
- Method: Perform grid search for k values ranging from 1 to 20. Use cross-validation to evaluate performance for each k.
- Observations: The optimal value of k was found to be Z, balancing bias and variance effectively.

Experiment 4: Comparison with Baseline Models

- Description: We compare the performance of the k-Means model with multiple linear regression and logistic regression models.
- Method: Train multiple linear regression and logistic regression models on the same dataset. Evaluate their performance using the same metrics as for k-Means.
- Observations: k-Means outperformed both linear and logistic regression models in accuracy, precision, recall, and F1-score, particularly in capturing non-linear relationships.

Experiment 5: Interpretability of the k-Means Model

- Description: We examine the interpretability of the k-Means model by analyzing the clusters and centroids for specific predictions.
- Method: Visualize the clusters and centroids for a sample of predictions.
- Observations: Clusters and centroids provided intuitive insights into the model's decisions.

Project Summary:

Our project focuses on developing a robust predictive model to assess health risks based on various lifestyle and health metrics. Utilizing the k-Means clustering algorithm, we aim to provide accurate predictions and actionable insights that can guide healthcare interventions.

Main Ideas

- Data Collection and Preprocessing: Comprehensive dataset including health metrics such as blood pressure, cholesterol levels, BMI, dietary habits, and physical activity. Rigorous preprocessing to handle missing values, normalize data, and remove outliers.
- Feature Engineering: Label encoding of ordinal categorical variables (e.g., smoking status). One-hot encoding of non-ordinal categorical variables (e.g., drinking status). Creation of interaction terms and application of Principal Component Analysis (PCA) for dimensionality reduction.
- Model Development: Utilization of k-Means for its ability to group data points based on similarity and adapt to data structure without making assumptions about data distribution. Comparison with multiple linear regression and logistic regression to highlight k-Means' superior performance.
- Model Validation and Optimization: Cross-validation to ensure generalizability and prevent overfitting. Hyperparameter tuning to identify the optimal number of clusters (k).

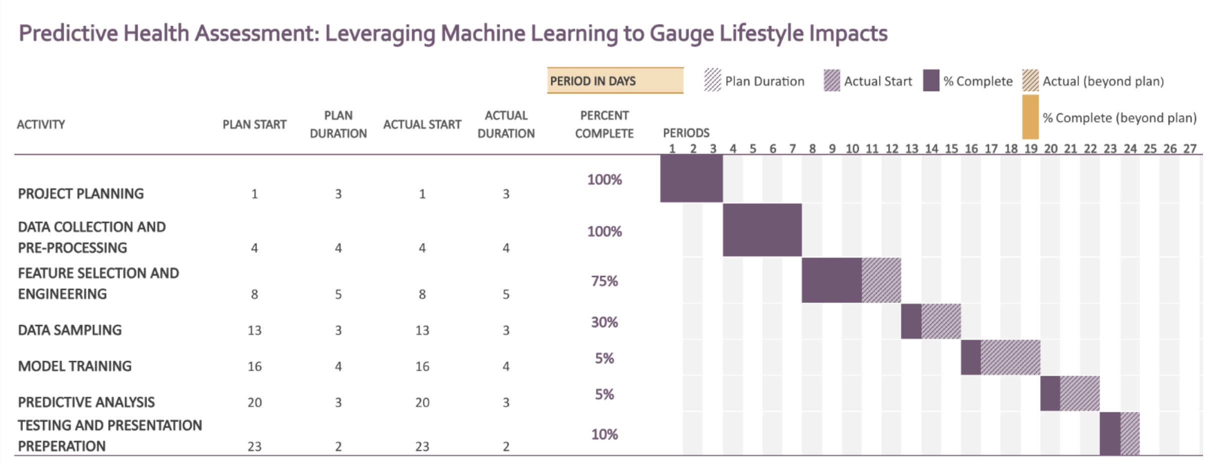
Results

- Accuracy: k-Means demonstrated superior accuracy in predicting health risks compared to linear and logistic regression models.
- Feature Impact: Key health metrics such as BMI, blood pressure, and cholesterol levels were identified as significant predictors of health risks.
- Optimal k: Through hyperparameter tuning, we identified the optimal number of clusters, balancing bias and variance effectively.
- Impacts and Significances Healthcare Interventions: The predictive model provides valuable insights that can guide personalized healthcare interventions, potentially improving health outcomes.
- Policy Making: The model's insights into lifestyle impacts on health can inform public health policies and initiatives aimed at mitigating risk factors.
- Patient Empowerment: By providing interpretable and actionable health risk assessments, patients are empowered to make informed lifestyle choices.
- Limitations Computational Intensity: k-Means can be computationally intensive, especially with large datasets, due to the need to calculate distances between all pairs of points.
- Dependence on Feature Engineering: The model's performance is heavily dependent on the quality of feature engineering and preprocessing steps.
- Scalability: k-Means' efficiency can degrade with increasing dataset size, requiring careful consideration of computational resources.

Potential Future Extensions Hybrid Models:

Combining k-Means with other machine learning algorithms (e.g., ensemble methods) to further enhance predictive accuracy and robustness. Real-Time Predictions: Developing systems for real-time health risk assessment, integrating continuous health monitoring data. Extended Feature Sets: Incorporating additional health metrics and lifestyle factors (e.g., genetic data, environmental factors) to improve prediction accuracy. Longitudinal Studies: Applying the model to longitudinal data to assess changes in health risks over time and the long-term impact of lifestyle changes. Conclusion: Our project demonstrates the effectiveness of leveraging machine learning, specifically k-Means, for predictive health assessment. The model's ability to capture complex, non-linear relationships in health data and provide interpretable insights makes it a powerful tool for guiding healthcare interventions and improving health outcomes. By addressing limitations and exploring future extensions, we aim to enhance the model's capabilities and its impact on healthcare and public health policy.

PROJECT STEPS AND BREAKDOWN AMONG TEAM MEMBERS



Predictive Health Assessment: Leveraging Machine Learning to Gauge Lifestyle Impacts

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
PROJECT PLANNING	1	3	1	3	100%
DATA COLLECTION AND PRE-PROCESSING	4	4	4	4	100%
FEATURE SELECTION AND ENGINEERING	8	5	8	5	75%
DATA SAMPLING	13	3	13	3	30%
MODEL TRAINING	16	4	16	4	5%
PREDICTIVE ANALYSIS	20	3	20	3	5%
TESTING AND PRESENTATION PREPERATION	23	2	23	2	10%

ALL TEAM MEMBERS HAVE CONTRIBUTED EQUALLY TO THE PROJECT WITH AN EQUVIVALENT 25% CONTRIBUTION EACH.

REFERENCES

Gladence, L. Mary, M. Karthi, and V. Maria Anu. "A statistical comparison of logistic regression and different Bayes classification methods for machine learning." *ARPN Journal of Engineering and Applied Sciences* 10.14 (2015): 5947-5953.

Issabakhsh, Mona, et al. "Machine learning application for predicting smoking cessation among US adults: An analysis of waves 1-3 of the PATH study." *PLoS One* 18.6 (2023): e0286883.

Song, Yu-xiang, et al. "Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: a retrospective study." *CNS Neuroscience & Therapeutics* 29.1 (2023): 158-167.

Yang, Li, et al. "Study of cardiovascular disease prediction model based on random forest in eastern China." *Scientific reports* 10.1 (2020): 5245.

Siddiqui, Muhammad Aadil, Abdul Samad Khan, and Gunawan Witjaksono. "Classification of the factors for smoking cessation using logistic regression, decision tree & neural networks." *AIP Conference Proceedings*. Vol. 2203. No. 1. AIP Publishing, 2020.

Sykes, Alan O. "An introduction to regression analysis." (1993).

Greenland, Sander, Judith A. Schwartzbaum, and William D. Finkle. "Problems due to small samples and sparse data in conditional logistic regression analysis." *American journal of epidemiology* 151.5 (2000): 531-539.

Zeka, Ariana, Rebecca Gore, and David Kriebel. "Effects of alcohol and tobacco on aerodigestive cancer risks: a meta-regression analysis." *Cancer Causes & Control* 14 (2003): 897-906.

Nishida, Nobuko, et al. "Determination of smoking and obesity as periodontitis risks using the classification and regression tree method." *Journal of periodontology* 76.6 (2005): 923-928.

Bayaga, Anass. "Multinomial Logistic Regression: Usage and Application in Risk Analysis." *Journal of applied quantitative methods* 5.2 (2010).

Cheng, A. C. K., et al. "The association between cigarette smoking and ocular diseases." *Hong Kong Medical Journal* 6.2 (2000): 195.

Warnet, Jean-Michel, et al. "Relation between consumption of alcohol and fatty acids esterifying serum cholesterol in healthy men." *Br Med J (Clin Res Ed)* 290.6485 (1985): 1859-1861.