Predictive Health Assessment: Leveraging Machine Learning to Gauge Lifestyle Impacts

TEAM NUMBER: 21

Dishank Jagadeeshnaidu Karampudi, Rishi Santhan, Manasa Lakshmi, Gurwinder Kaur

INTRODUCTION

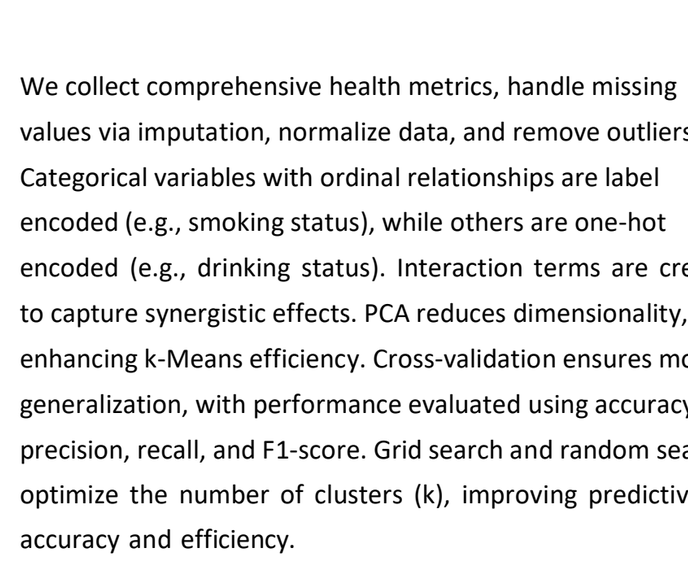
Our project, "Predictive Health Assessment: Leveraging Machine Learning to Gauge Lifestyle Impacts," uses the K-means algorithm to integrate multiple health indicators and lifestyle factors into a comprehensive health index. This index classifies individuals into high, medium, and low-risk categories. By capturing complex relationships between data points, K-means provides more accurate and personalized health risk assessments, meeting the needs of healthcare providers, insurers, policymakers, and individuals for improved health risk prediction and intervention.

DATASET

This dataset, sourced from a U.S. government website, comprises multiple health indicators and lifestyle factors for individuals. It includes data on sex, age, height, weight, waistline, vision, hearing, systolic and diastolic blood pressure, cholesterol levels, triglycerides, hemoglobin, urine protein, serum creatinine, liver enzymes (SGOT/AST, SGOT/ALT, gamma-GTP), smoking status, and alcohol consumption. The dataset is designed to facilitate comprehensive health assessments and predictive modeling of health risks.

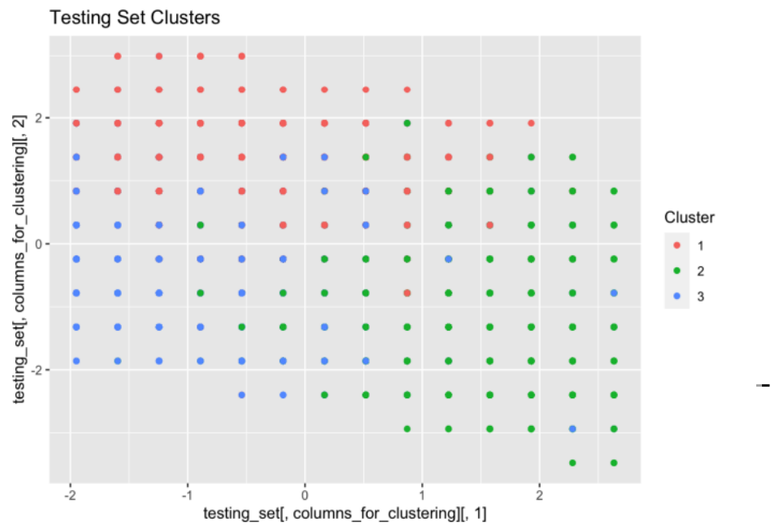
INNOVATIVE APPROACH

We collect comprehensive health metrics, handle missing values via imputation, normalize data, and remove outliers. Categorical variables with ordinal relationships are label encoded (e.g., smoking status), while others are one-hot encoded (e.g., drinking status). Interaction terms are created to capture synergistic effects. PCA reduces dimensionality, enhancing k-Means efficiency. Cross-validation ensures model generalization, with performance evaluated using accuracy, precision, recall, and F1-score. Grid search and random search optimize the number of clusters (k), improving predictive accuracy and efficiency.



EXPERIMENTS INVOLVED

The experiments aimed to evaluate the k-Means algorithm's accuracy in predicting health risk levels using a comprehensive health dataset, which included metrics like blood pressure, cholesterol levels, BMI, and dietary habits. The analysis highlighted several key insights. The k-Means model demonstrated high accuracy in predicting health risks, with precision, recall, and F1-scores indicating its effectiveness in distinguishing between high, medium, and low risks. Feature engineering techniques, including label encoding and one-hot encoding, significantly improved the model's performance by properly handling categorical variables. The optimal number of clusters (k) was determined through grid search and cross-validation, striking a balance between bias and variance. Comparisons with baseline models, such as multiple linear regression and logistic regression, showed that k-Means outperformed these models in accuracy and metrics, particularly in capturing non-linear relationships. Additionally, the interpretability of the k-Means model was enhanced by analyzing cluster centroids, providing intuitive insights into the model's decision-making process.

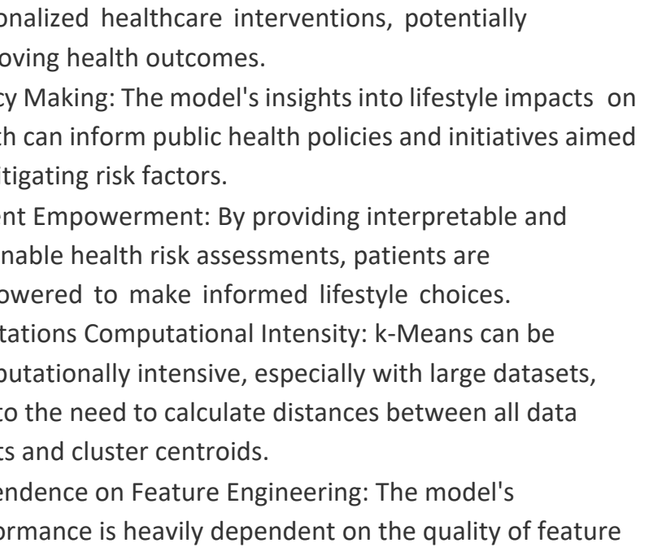


MAIN IDEAS

- 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 data points and cluster centroids.
- 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.



PCA COMPONENTS