Unraveling Diabetes: Written Report

Winnie Taiwo, ISYE 7406

Abstract

This comprehensive report delves into the intricate factors influencing diabetes prevalence using the 2015 Behavioral Risk Factor Surveillance System (BRFSS) data. The study's objective was to unearth the complex interplay between lifestyle choices, behavioral patterns, socioeconomic factors, and other health indicators in relation to diabetes, employing a variety of statistical learning methods. Techniques ranged from logistic regression and decision trees for lifestyle and behavioral analysis, to support vector machines for assessing socioeconomic impacts, and linear models like Ridge and Lasso for health indicators' associations. Key findings include the identification of significant lifestyle and behavioral predictors, the nuanced role of socioeconomic factors, and the clear association between specific health indicators and diabetes risk. The research also introduced dynamic modeling to predict transitions between non-diabetic, pre-diabetic, and diabetic states, aiming at early intervention and personalized healthcare strategies. A notable aspect of the study was the application of a Random Forest model, which effectively utilized health indicators to predict diabetes presence, demonstrating the model's robustness and ability to handle non-linear relationships. Overall, the project offers valuable insights for targeted intervention strategies and contributes to personalized healthcare solutions by highlighting the multifaceted nature of diabetes risk factors and the efficacy of data-driven approaches in public health.

Introduction

In this report, I delve into the escalating issue of diabetes, a global health epidemic that has seen a dramatic increase in prevalence, affecting millions worldwide. The World Health Organization's alarming report of 422 million adults diagnosed with diabetes in 2014 underscores the gravity of this health crisis. Motivated by personal experiences and the broader implications on public health and healthcare systems, this study aims to dissect the complex interplay of factors contributing to diabetes, leveraging the power of data analysis to unearth actionable insights that could lead to more effective prevention and treatment strategies.

The foundation of the inquiry is the Diabetes Health Indicators Dataset from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), which comprises 253,680 survey responses detailing a range of health indicators. Diabetes_012 Dataset with its notable class imbalance and the balanced Diabetes Binary 50/50 Split dataset for binary classification. The dataset presents a multitude of data mining challenges, including but not limited to high dimensionality, the presence of class imbalance, and the complexities inherent in modeling the nonlinear relationships between a diverse set of predictors and diabetes outcomes.

To navigate these challenges, The problem-solving strategy encompasses a multifaceted approach, employing a variety of statistical learning techniques. I utilize logistic regression and decision trees to investigate the influence of lifestyle and behavioral factors, while support vector machines (SVMs) are applied to explore socioeconomic impacts. Furthermore, Random Forest models are deployed to predict the presence of diabetes based on health indicators, showcasing the potential of machine learning in uncovering patterns and relationships within the data.

Throughout the application of these methodologies, I have gained valuable insights and learned key lessons, particularly regarding the significance of lifestyle choices, the nuanced role of socioeconomic factors, and the predictive power of ensemble modeling in identifying diabetes risk. These learnings underscore the importance of a holistic approach to diabetes research, integrating diverse analytical techniques to capture the multifactorial nature of the disease.

This report is structured to provide a comprehensive overview of the study, starting with a detailed problem description and the motivation behind the research. I then outline the data mining challenges encountered and describe the problem-solving strategies adopted to address these challenges. The subsequent sections present the analysis and key findings from the application of various data mining methods, culminating in a conclusion that synthesizes the learnings and insights. Finally, I reflect on the lessons learned throughout this project, offering valuable perspectives on the application of data mining techniques in the field of healthcare.

Problem Statement / Data Sources

This study addresses the escalating global health crisis of diabetes by leveraging the Diabetes Health Indicators Dataset from the 2015 Behavioral Risk Factor Surveillance System (BRFSS). The motivation behind this research is driven by the profound personal and societal impacts of diabetes, a condition that affects millions globally. The World Health Organization reported an estimated 422 million adults living with diabetes in 2014, highlighting the urgency of addressing this condition (World Health Organization, 2014). This project aims to dissect the complex interplay of factors contributing to diabetes through advanced data analysis, with the goal of enhancing personalized medicine and preventive care strategies.

The data sources for this project are derived from three key datasets:

- Diabetes_012_health_indicators_BRFSS2015.csv: This dataset, containing 253,680 responses, categorizes individuals into three groups based on their diabetes status and features 21 variables capturing a range of health indicators (CDC, 2015a).
- Diabetes_binary_5050split_health_indicators_BRFSS2015.csv: A balanced subset of 70,692 responses for binary classification tasks, providing an equal distribution of diabetic and non-diabetic individuals (CDC, 2015b).
- Diabetes_binary_health_indicators_BRFSS2015.csv: Comprising 253,680 responses focused on binary classification of diabetes presence, this dataset reflects the original BRFSS survey distribution, presenting a realistic challenge of class imbalance (CDC, 2015c).

These datasets present several data mining challenges, including high dimensionality, class imbalance, and the complexity of modeling relationships between predictors and diabetes outcomes. To address these challenges, the study employs a comprehensive suite of statistical learning techniques, including logistic regression, decision trees, support vector machines, and Random Forest models. These methodologies are selected for their ability to explore and interpret complex patterns within the datasets, focusing on identifying key lifestyle, behavioral, and environmental predictors of diabetes. The problem-solving strategy of the project involves a methodological approach that encompasses lifestyle analysis, socioeconomic studies, health correlation analysis, and predictive modeling. This multifaceted approach aims to uncover insights into the key factors influencing diabetes risk and provide a deeper understanding of the disease's multifaceted nature.

Through the application of these data mining techniques, I have accomplished significant learnings. I have identified critical lifestyle and behavioral predictors, the impact of socioeconomic factors on diabetes prevalence, and the efficacy of predictive modeling in distinguishing diabetes cases. These findings not only advance the understanding of diabetes but also underscore the potential of data-driven approaches in addressing complex health issues.

This report provides a comprehensive overview of the methodology, key findings, and conclusions drawn from the data mining practice. It outlines the journey through the problem description and motivation, data mining challenges, problem-solving strategies, and the accomplished learning from the applications, concluding with reflections on the lessons learned and potential avenues for future research.

Proposed Methodology

The methodology for this study on diabetes, as outlined in the project proposal and presentation materials, leverages a comprehensive suite of data mining and statistical learning techniques to analyze the Diabetes Health Indicators Dataset from the 2015 Behavioral Risk Factor Surveillance System (BRFSS). The proposed methodology is driven by several scientific research questions aimed at identifying key lifestyle, behavioral, and socioeconomic factors influencing diabetes across diverse demographics.

Scientific Research Questions:

- Lifestyle and Behavioral Predictors: Investigating the association of lifestyle and behavioral health indicators with diabetes in the adult population as of 2015.
- Socioeconomic Influences: Exploring the relationship between socioeconomic factors and the prevalence of diabetes.
- Health Indicator Associations: Assessing which health indicators are most strongly associated with diabetes status among adults.
- Predictive Modeling with Random Forest: Evaluating the effectiveness of a Random Forest model in utilizing health indicators such as BMI, age, physical activity levels, and other health metrics to predict the presence of diabetes.

Model Selection/Methods:

- Logistic Regression and Decision Trees: For identifying lifestyle and behavioral predictors, offering insights into linear and non-linear relationships, respectively. Decision trees allow for capturing complex interactions without the need for linear assumptions.
- Support Vector Machines (SVMs): Employed for analyzing socioeconomic impacts, given their ability to handle high-dimensional data and model complex relationships, which is particularly useful for data that may not be linearly separable.
- Linear Models (Linear Regression, Ridge, Lasso): Applied for examining health indicator associations with diabetes, where linear relationships are assumed, and multicollinearity is addressed through regularization.
- Random Forest: Chosen for its robustness to outliers and ability to handle non-linear relationships, making it suitable for predicting diabetes presence using a wide array of health indicators. The model's capacity to provide insights into feature importance is particularly valuable.
- PCA Combined with Logistic Regression: Utilized for combinatorial analysis of diabetes risk indicators, aiming to reduce dimensionality and enhance model performance. This approach is suitable for datasets with high dimensionality and multicollinearity.

Rationale for Method Selection:

The diverse set of methods chosen for this study allows for a multi-faceted analysis of the complex nature of diabetes risk factors. Each method brings unique strengths to the table, from logistic regression and decision trees' ability to interpret individual predictor impacts to SVMs' capacity for modeling complex relationships and Random Forest's robustness and predictive accuracy. The integration of PCA with logistic regression offers a strategic approach to managing high dimensionality and multicollinearity, further enhancing the study's analytical depth. This methodological diversity is crucial for addressing the research questions comprehensively and providing actionable insights into diabetes prevention and management.

Common Metric for Model Comparison:

The methodology incorporates the use of common metrics such as Accuracy, Precision, Recall, F1 Score, and AUC-ROC across models to ensure comparability and provide a comprehensive view of model performance. These metrics collectively account for the models' ability to predict diabetes accurately and balance between sensitivity and specificity, offering a holistic assessment of their effectiveness.

Model Tuning Process:

The Model Tuning Process for our data mining project involved meticulous adjustments to ensure the precision and reliability of our predictive models. Decision Trees began with default settings, including an unlimited maximum depth and at least one sample per leaf. Through GridSearchCV, I explored varying depths and minimum samples to prevent overfitting, ultimately selecting a maximum depth of 30 and a minimum of four samples per leaf, enabling the trees to elaborate on complex patterns without compromising decision reliability. Random Forest models started with 100 trees, and I refined this by considering 300 trees, a depth of 10, and a minimum split of 5, substantially enhancing performance and accuracy. For Support Vector Machines (SVM), beginning with a linear kernel and a C value of 1, I discovered through GridSearchCV that a C of 10 and a gamma of 0.1 for the RBF kernel were optimal, balancing bias and variance adeptly. The PCA combined with Logistic Regression initiated without a set number of components and a C of 1; further tuning identified 10 PCA components and a logreg__C of 10 as ideal, effectively reducing dimensionality and managing multicollinearity. Finally, Ridge and Lasso Regression models were initially unspecified for alpha values, but detailed tuning led to the selection of the most suitable alphas, ensuring minimal error and optimal interpretability. These models were carefully selected and tuned to address the specific characteristics of our dataset, addressing high dimensionality, potential overfitting, and the imperative for clear model interpretability. This thorough approach has equipped us with robust, reliable, and interpretable models, significantly contributing to our understanding of diabetes prevalence and informing the development of effective predictive tools.

Analysis and Results

In examining the Diabetes Health Indicators Dataset, a range of data mining techniques were employed to investigate key scientific questions. The findings from the study are summarized as follows:

Key Findings:

- Lifestyle and Behavioral Predictors: By utilizing Logistic Regression and Decision Trees, factors such as BMI, age, and high blood pressure were identified as significant indicators, with logistic regression achieving a 75% accuracy rate.
- Socioeconomic Impact: SVMs, with different kernels, were applied to assess the effect of socioeconomic factors on diabetes. A linear kernel SVM demonstrated a 75% accuracy, while RBF and Polynomial kernels showed lower accuracies (73% and 66%, respectively), indicating the intricate nature of these socioeconomic relationships.

- Health Indicator Analysis: The Random Forest approach used various health indicators, including BMI and physical activity levels, to predict diabetes, attaining around 81.7% accuracy. This model was particularly adept at identifying non-diabetic cases.
- Dimensionality Reduction and Prediction: Combining PCA with Logistic Regression proved effective in streamlining the dataset and boosting model performance, with key indicators such as age, blood pressure, and cholesterol levels emerging as significant. The model's overall accuracy stood at 87%.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**MSE**	**R²**	**Best For**
Logistic Regression	75%	0.75	0.75	0.75	-	-	Binary classification and identifying linear relationships.
Decision Trees	73%	0.73	0.73	0.73	-	-	Capturing non-linear patterns and feature interactions without scaling.
Linear SVM	75%	0.75	0.75	0.75	-	-	Handling linearly separable data with efficiency.
RBF Kernel SVM	73%	0.73	0.73	0.73	-	-	Dealing with non-linear and high-dimensional data.
Polynomial Kernel SVM	66%	0.66	0.66	0.66	-	-	Exploring polynomial relationships in data.
PCA + Logistic Regression	87%	0.71	0.57	0.58	-	-	Reducing dimensionality and addressing multicollinearity in logistic regression.
Linear Regression	-	-	-	-	0.1726	0.3096	Estimating the strength and sign of relationships between variables.
Ridge Regression	-	-	-	-	0.1726	0.3096	Shrinking the regression coefficients to reduce model complexity.
Lasso Regression	-	-	-	-	0.1726	0.3095	Performing variable selection and regularization to enhance prediction accuracy.
Random Forest	82%	0.77	0.82	0.78	-	-	Utilizing multiple health indicators to predict diabetes, especially non-diabetic cases.

Explanation for Each Model:

In the Explanation for Each Model section of my report, I've detailed the function and strengths of various models used in our study. Logistic Regression and Decision Trees are pivotal for binary outcomes, excelling in discerning relationships and patterns within the dataset. The flexibility of Decision Trees is particularly noteworthy as they thrive in managing datasets without linear constraints, proving to be an asset for analyzing complex datasets.

The SVM Models, with a Linear SVM, showcased efficiency in processing large datasets that are linearly separable. RBF and Polynomial kernels were selected for their adeptness at mapping data into higher dimensions, a vital feature when encountering non-linear boundaries which are often present in intricate datasets.

The Linear, Ridge, and Lasso Regression models were employed to forecast continuous outcomes. Linear Regression is particularly adept at determining the direction and strength of relationships between variables. Ridge and Lasso Regressions enhance these predictions by mitigating overfitting, with Ridge reducing model complexity and Lasso aiding in pinpointing significant predictors. Random Forest model achieved a high accuracy rate and demonstrated a strong capacity to utilize a wide array of health indicators in predicting diabetes, showcasing its versatility and robustness in handling different types of data.

Finally, the combination of PCA with Logistic Regression allowed us to capitalize on PCA's ability to reduce the dimensionality of our dataset, thus amplifying the predictive prowess of Logistic Regression. This approach was particularly successful in accurately identifying crucial health indicators for diabetes, a testament to the power of integrating dimensionality reduction techniques with predictive modeling.

Model Comparison and Cross-Validation:

Logistic Regression and Random Forest were the front runners in accuracy, outperforming SVMs. This disparity in model performance emphasizes the need for careful model selection tailored to the data's unique features and the research objectives. Through cross-validation, Logistic Regression consistently surpassed Decision Trees, with average accuracy scores of 74.7% against 65.6%, highlighting the former's stability and reliability in diabetes prediction.

The above analysis offers an understanding of the various factors influencing diabetes, paving the way for the creation of focused intervention strategies anchored in solid data-driven findings.

Conclusions

Drawing from my comprehensive analysis of the Diabetes Health Indicators Dataset, I've reached several key conclusions. First, the study confirmed that higher BMI and older age significantly increase the risk of diabetes, highlighting the critical need for lifestyle modifications as part of prevention strategies. My use of Support Vector Machines (SVMs) with various kernels illuminated the complex and possibly non-linear impact of socioeconomic factors on diabetes prevalence, suggesting that intervention strategies may need customization to effectively address different socioeconomic backgrounds. Additionally, employing the Random Forest model enabled me to effectively identify BMI, age, and blood pressure as primary health indicators for diabetes, which can greatly assist in developing screening tools for risk assessment. The insights derived from combining PCA with logistic regression, particularly regarding age and blood pressure, underscore the potential of early intervention models to identify individuals at high risk sooner, allowing for timely and personalized preventive actions.

These findings underline a deep understanding of the multifactorial influences on diabetes and underscore the potential of targeted intervention strategies grounded in solid data-driven insights. Moving forward, I plan to delve deeper into the temporal dynamics of the identified risk factors, potentially developing predictive models that can forecast changes in an individual's diabetes risk profile. Incorporating more comprehensive data types, like genetic information or detailed dietary habits, could further refine the models' predictive accuracy. Collaborating with healthcare providers to implement these findings clinically could also advance intervention strategies, ultimately enhancing patient outcomes and contributing to broader public health benefits.

Lessons Learned and Suggestions for Course Improvement

Throughout the Data Mining & Statistical Learning course (ISYE-7406), I have gained invaluable insights and practical skills that have greatly enhanced my understanding of complex data analysis. Working on the Diabetes Health Indicators Dataset provided me with a real-world application of the theories and methods I learned, reinforcing the importance of a methodical and data-driven approach in tackling public health issues.

1. Importance of Data Preprocessing: I learned that the quality of insights depends significantly on the initial stages of data preprocessing. Handling missing data, outliers, and understanding the distribution of the data are critical steps that can influence the outcome of the analysis.
2. Value of Collaboration: The discussions and interactions with peers and instructors enhanced my learning experience, providing diverse perspectives that enriched my understanding of the subject matter.

Suggestions for Course Improvement:

While the course was highly educational, I have a few suggestions that could enhance the learning experience for future students:

1. More Case Studies: Incorporating more case studies, especially from different domains, could help students see the broader application of data mining techniques across various industries and research fields.

2. Advanced Topics in Data Mining: Introducing more advanced topics, such as deep learning and its applications in data mining, could be valuable. This would provide students with exposure to cutting-edge techniques in the field.