Exploring Predictive Models for Diabetes: A Comprehensive Analysis

Abstract

The escalating demand for healthcare professionals is a pressing concern on a global scale, with the United States grappling with its own substantial challenges. According to the American Hospital Association, the U.S. is projected to confront a shortage of up to 124,000 physicians by 2033, coupled with the imperative need to recruit at least 200,000 nurses annually to address heightened demand and fill vacancies due to retiring healthcare professionals.

In response to this growing need, innovative approaches to healthcare delivery, including both in-person and remote diagnostic methods, are becoming increasingly vital. This necessitates the application of advanced data mining techniques, specifically leveraging linear regression models (LGR), for a comprehensive analysis of patient data.

Linear regression, as a data mining tool, proves invaluable in deciphering the intricate relationships within healthcare datasets. By employing LGR, it becomes possible to explore and understand the intricate dynamics influencing the demand for healthcare professionals. This entails scrutinizing various factors such as population demographics, regional healthcare infrastructure, and lifestyle variables, offering critical insights into the evolving landscape of healthcare demands.

In this paper, we delve into the application of linear regression models in addressing the burgeoning need for healthcare professionals. By scrutinizing patient data, we aim to uncover meaningful patterns and correlations that contribute to a more nuanced understanding of the factors influencing healthcare demand. This research not only tackles the imminent shortage of healthcare professionals but also sheds light on the broader implications for healthcare delivery in an era of evolving medical requirements.

We begin with a proper literature review to find similarities in picking methods to properly dissect and find a solution for our problem:

Literature Review:

End-to-End Data Science Example: Predicting Diabetes with Logistic Regression.

i. Using Python and its respective libraries to predict whether females of age

21 and above from the Pima Indians Diabetes Database have diabetes.

Diabetes Prediction Using Machine Learning (Amaan Preeti Gulati)

i. This article explains how to use machine learning to predict diabetes in

individuals using the Pima Indians Diabetes Database.

Using Machine Learning to predict if someone has Diabetes (Edward Leoni)

i. The article explains how to predict if someone has diabetes or not based

on 8 variables using Keras, a high-level neural networks API written in

Python and TensorFlow, CNTK, or Theano.

1. Introduction

Diabetes, a severe chronic disease, significantly diminishes both the quality of life and life expectancy of individuals by impairing the effective regulation of glucose levels in the blood. When food is digested, it breaks down into sugar, entering the bloodstream and prompting the pancreas to release insulin. This insulin is crucial for facilitating the use of blood sugar by the body's cells for energy. Diabetes manifests when the body fails to produce adequate insulin or when it cannot utilize the produced insulin effectively.

The Behavioral Risk Factor Surveillance System (BRFSS), an annual telephone health survey conducted by the CDC, gathers responses from over 400,000 Americans. It addresses health risk behaviors, chronic health conditions, and preventive service utilization. Since its inception in 1984, the BRFSS has been a vital resource for understanding health-related trends.

This project utilized CSV files from a Kaggle dataset available in 2015, originating from the BRFSS survey. The dataset, comprising responses from 441,455 participants and featuring 330 variables, encompasses both direct participant responses and variables calculated based on individual answers.

Data Set breakdown:

Data Collection : Diabetes Health Indicators (from Kaggle):

Three files containing data collected from the Behavioral Risk Factor

Surveillance System (BRFSS) through an automated telephone system. The data

is diverse as it includes fields such as: Income, Education, Sex, Veggies, Alcohol

Consumption, and Exercise, all will be useful in providing accurate predictions.

1. Data Set 1:

- Classes: 0 for no diabetes or diabetes during pregnancy, 1 for prediabetes, 2 for diabetes

- Imbalanced classes

- 21 feature variables

2. Data Set 2:

- Classes: 0 for no diabetes, 1 for prediabetes or diabetes (50-50 split response)

- Balanced classes

- 21 feature variables

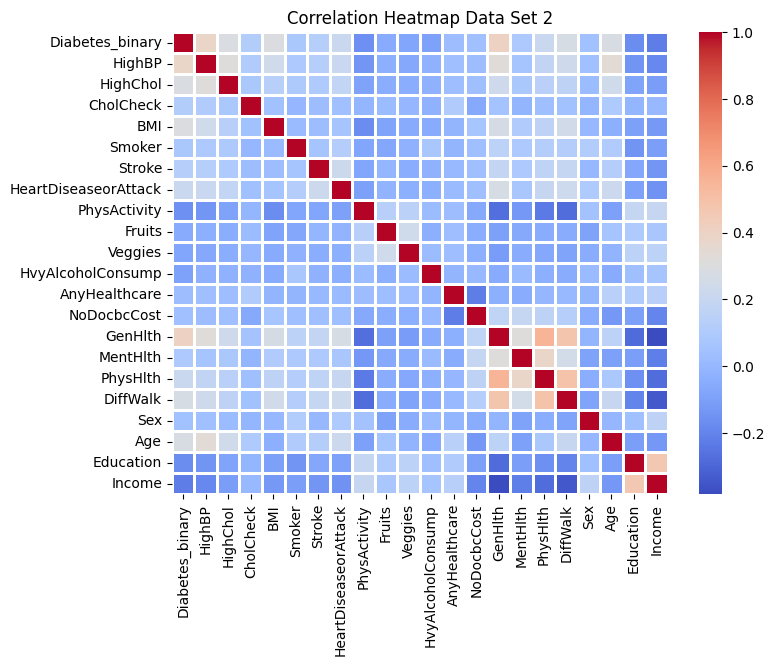
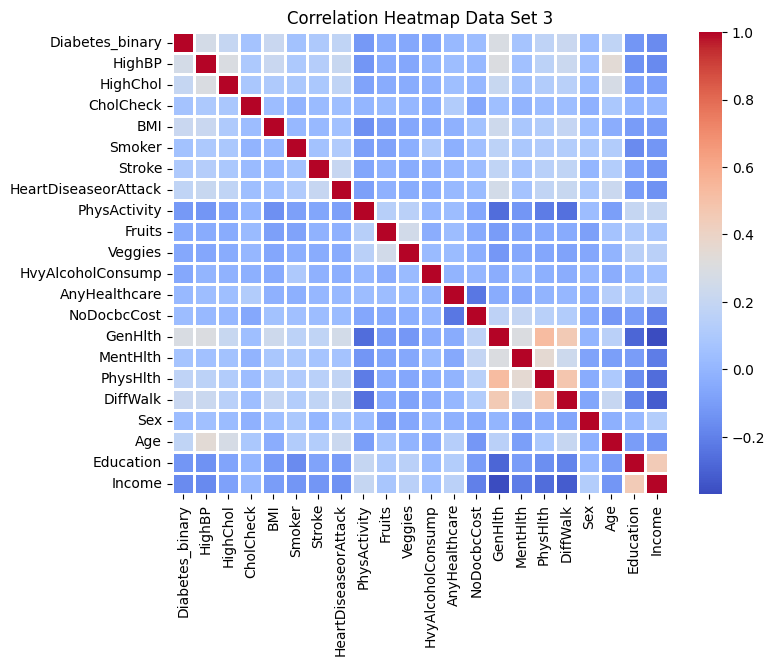
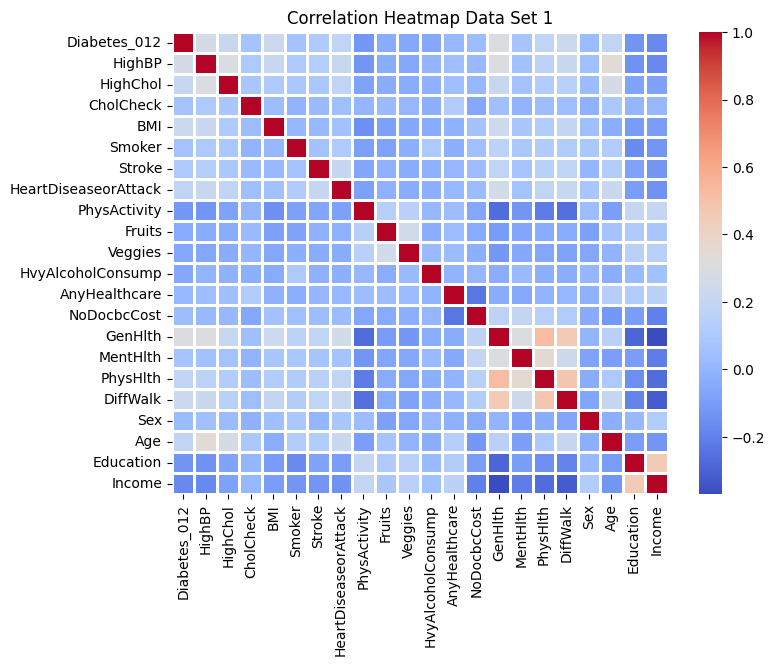
3. Data Set 3:

- Classes: 0 for no diabetes, 1 for prediabetes or diabetes

- Imbalanced classes

- 21 Feature Variables

Correlation Matrices for each Data Set



We have opted to focus our analysis on Data Set 2, which employs a binary model with two classifications: 0 for no diabetes and 1 for diabetes. This choice simplifies our analysis, making it more accessible and allowing for a clearer breakdown of results.We arrived at this choice after running our each data through a correlation analysis to determine which feature corresponds to a positive identification of Diabetes within a person. Data Set 2 has an even split of answers that allows us to run our code better. Having said that, it was the easiest choice out of the three to pick because it was balanced, allowing for better spread of our predicitons.

2. Methodology

In addressing the critical challenge of predicting diabetes within our chosen dataset, our methodology was carefully crafted to encompass key steps such as data preprocessing, feature selection, and the application of machine learning techniques. Our chosen methods to break down and get accurate predictions were using Logistic Regression algorithms, a KNN Model, and a K means Clustering algorithm

Logistic Regression Method:

Logistic Regression is perfect for our problem because it is specifically designed for binary classification problems, in our case Data Set 2 is a balanced and binary data set. In order to enhance the quality of our data set we did some data preprocessing tasks. First we addressed class imbalances to prevent biases within the machine learning model by implementing techniques such as oversampling or undersampling. This approach ensured a fair representation of different classes, mitigating the risk of skewed predictions. Next we checked for missing values.Luckily our data set had been full when we were exploring for good data to experiment with. Next we had to decide what features to include in our model, based off the correlation matrix we picked: 'HighBP', 'HighChol', 'BMI', 'HeartDiseaseorAttack', 'PhysActivity', 'GenHlth', 'PhysHlth', 'DiffWalk', 'Age'. These features exhibited a relationship with our target variable (diabetes) and would help us get the best results through our model. Finally we chose to split our data, approximately 80% of the data comprising the training set, allowing the machine learning model to learn intricate patterns and relationships. The remaining 20% constituted the test set, serving as an independent validation set to assess the model's predictive performance.

KNN Model

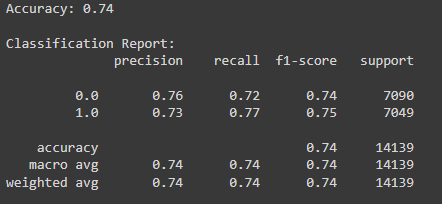
The KNN model was chosen because it is used for both classification and regression challenges. It is a fairly straightforward algorithm that uses the training set, instead of training a new data set, in real time to predict new instances given the feature selections. The idea is the model predicts models based on the ‘k’ nearest neighbor in the feature space. The model is simple to understand and easy to decipher. We used the same features selected once again: 'HighBP', 'HighChol', 'BMI', 'HeartDiseaseorAttack', 'PhysActivity', 'GenHlth', 'PhysHlth', 'DiffWalk', 'Age'. There was no need to split the data set as the KNN model runs its algorithm on the training set itself.

K Means Clustering

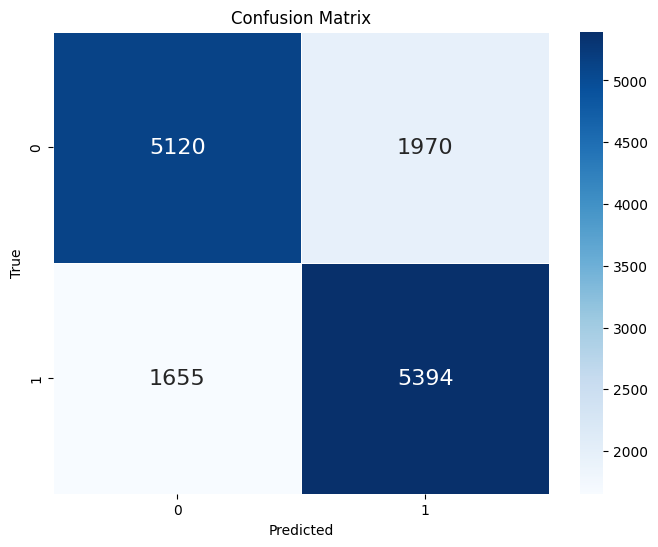
K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into a set of distinct, non-overlapping subgroups or clusters. The goal of the K-means algorithm is to group data points into clusters based on similarity, with the number of clusters (K) specified by the user. We picked the ample amount of clusters (k) using the elbow method. The elbow method is a technique used to determine the optimal number of clusters (k) in a K-Means clustering algorithm. It helps find a balance between having too few clusters (which might not capture the underlying structure in the data) and having too many clusters (which might overfit the data).

3. Results - should prob do verification curves - no results for non pregnancy cases

Logistic Regression



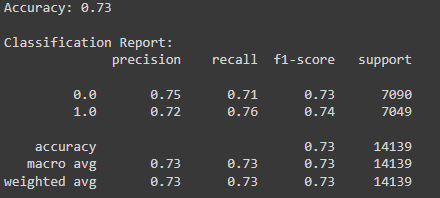
The logistic regression model demonstrated a commendable accuracy of 74%. Precision and recall metrics shed light on the model's ability to correctly predict instances of diabetes (Class 1) and no diabetes (Class 0). A precision of 0.73 for diabetes and 0.76 for no diabetes indicated a balanced approach, minimizing false positives and false negatives.



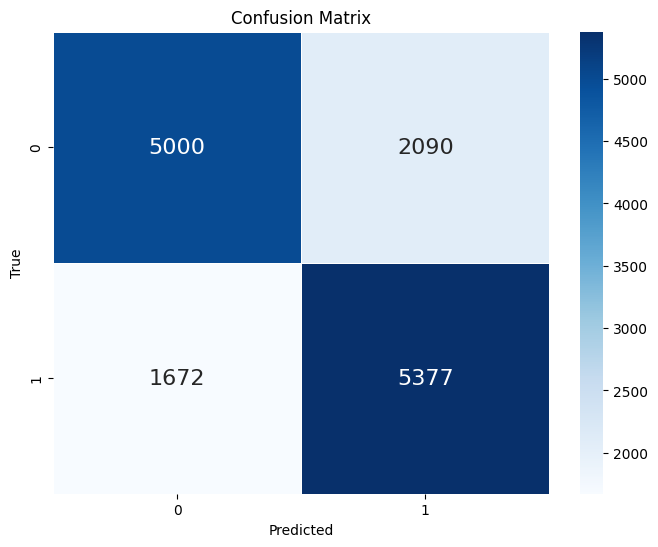
The confusion matrix further revealed the model's efficacy revealing:

* True Positives (TP): 5394 instances were correctly predicted as diabetes.
* True Negatives (TN): 5120 instances were correctly predicted as no diabetes.
* False Positives (FP): 1970 instances were incorrectly predicted as diabetes.
* False Negatives (FN): 1655 instances were incorrectly predicted as no diabetes.

K-Nearest Neighbors (KNN)



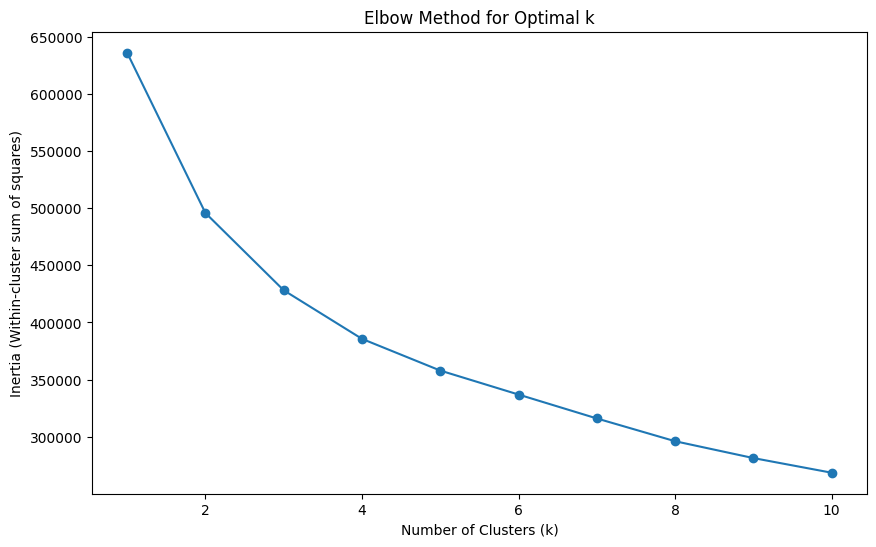
The KNN model, a versatile algorithm suitable for both classification and regression tasks, showcased an overall accuracy of 73%. Precision, recall, and F1-score metrics for both diabetes and no diabetes classes underscored the model's effectiveness. Notably, a precision of 0.75 for no diabetes and 0.72 for diabetes, along with corresponding recall values, indicated a well-balanced approach in capturing both true positive and true negative instances.



The confusion matrix reinforced the model's reliability by highlighting instances of correct and incorrect predictions:

* True Negative (TN): 5000
* False Positive (FP): 2090
* False Negative (FN): 1672
* True Positive (TP): 5377

K-Means Clustering



In the realm of unsupervised learning, K-means clustering was employed to identify inherent patterns within the dataset. The optimal number of clusters was determined using the elbow graph method, setting the stage for subsequent explorations into data groupings. Our elbow algorithm indicates that four clusters was the best choice to implement in the clustering algorithm



Our clustering report and graph was broken down into 4 clusters each having centroids around the BMI of approximately 29-32 and an AGE of 7-10. Those two features were chosen as the axis for the clustering graph as they showed the biggest correlation and most positive outcome when compared to the other feature selections. In the case of BMI, having a BMI Index of 29-32 means you are either in the ‘overweight’ range, or you are in the ‘obese’ range. This significant insight into our model because here we have a real time signifier of whether someone has diabetes or not, knowing that individuals who are overweight or obese most likely develop diabetes as they age.

The comprehensive analysis of logistic regression and KNN models provides valuable insights into their predictive capabilities for diabetes diagnosis. The chosen methodologies, precision-recall metrics, and confusion matrices serve as key indicators of model performance. As we delve deeper into the dataset, these results will guide further enhancements and refinements in our ongoing research. The incorporation of K-means clustering introduces an additional dimension to our exploration, laying the groundwork for uncovering hidden patterns within the data.

4. Limitations

Class Imbalance

One notable limitation lies in the class imbalance present in the target variable, particularly evident in the dataset labeled "diabetes\_012\_health\_indicators\_BRFSS2015.csv." Class imbalance introduces challenges as machine learning models may exhibit biases towards the majority class, potentially compromising the effectiveness of predictions. It is worth noting that our decision to utilize a preprocessed and balanced dataset was a strategic choice to mitigate this issue. However, addressing class imbalance remains an ongoing concern and warrants continuous attention in future iterations.

Survey Bias

The reliance on a telephone survey, specifically the Behavioral Risk Factor Surveillance System (BRFSS), introduces the potential for survey biases. Non-response bias and recall bias are inherent challenges associated with survey-based data collection methods. These biases may impact the generalizability of our findings to the broader population. Acknowledging and addressing these biases is crucial to ensure the reliability and validity of our predictive models.

Timing

A temporal limitation arises from the dataset's origin in 2015. Health-related trends are dynamic, subject to change over time. The patterns and correlations observed in 2015 may not necessarily hold true in the current landscape of health and wellness. Incorporating more recent data into our analysis would offer a more accurate and up-to-date representation of diabetes prevalence and associated factors.

Limited Demographic Information

While the dataset includes various health indicators, it may lack crucial demographic information, such as ethnicity and socio-economic status. These demographic factors are known to influence diabetes risk and could provide a more comprehensive understanding of the multifaceted nature of the disease. Future iterations of the project should consider expanding the dataset to encompass a broader range of demographic variables.

Model Evaluation

Our approach to model evaluation, while providing valuable insights, could have been more robust. Incorporating cross-validation methods, such as k-fold cross-validation, would offer a more comprehensive assessment of the models' generalization performance. Rigorous model evaluation ensures that the predictive capabilities of our models are thoroughly tested on diverse subsets of the dataset, enhancing the reliability of our findings.

Ethical Considerations

The use of health data necessitates a heightened awareness of ethical considerations. Future iterations of this project should incorporate a robust ethical framework, addressing concerns related to data privacy, informed consent, and the responsible use of sensitive health information. Ethical considerations are paramount in ensuring the integrity and societal impact of our research.

While our project has provided valuable insights into predicting diabetes, these limitations underscore the need for continuous refinement. Future iterations should focus on mitigating biases, incorporating recent data, enhancing demographic information, and implementing rigorous model evaluation techniques.

5. Conclusion

In traversing the complex terrain of diabetes prediction, our journey has been marked by insightful analyses, critical reflections, and a commitment to unraveling the intricacies of this chronic condition. The amalgamation of diverse methodologies, including logistic regression, K-nearest neighbors (KNN), and clustering, has provided a multifaceted perspective on the predictive landscape.

Model Performance:

The logistic regression model, boasting a commendable accuracy of 74%, and the KNN model, not far behind at 73%, have demonstrated their efficacy in deciphering patterns within the dataset. These models serve as invaluable tools in identifying potential instances of diabetes, contributing to the broader spectrum of predictive healthcare analytics.

Clustering Insights:

The clustering analysis has shed light on significant indicators of diabetes, with BMI and age emerging as key factors influencing the onset of this condition. This finding underscores the importance of not only individual health metrics but also the interplay between different factors in predicting the likelihood of diabetes.

Overall Implications:

Our overarching observation is that the model, collectively, has proven to be a valuable asset in gauging the impending onset of diabetes. The interplay of methodologies, particularly the use of logistic regression, highlights the relevance of employing advanced analytical techniques in disease prediction. Logistic regression, with its ability to model complex relationships between variables, stands out as a crucial tool in identifying patients at risk of specific diseases, such as diabetes.

Moving Forward:

As we conclude this exploration, it is imperative to recognize the limitations encountered, such as class imbalance and survey biases, which open avenues for future enhancements. The call for continuous refinement, integration of more recent data, and a more comprehensive inclusion of demographic variables resonates as we strive for more accurate and nuanced predictions.

In essence, our foray into diabetes prediction has unveiled promising results and significant insights. The intersection of machine learning, statistical analyses, and clustering techniques has proven to be a potent force in unraveling the complexities of diabetes onset. As we move forward, our commitment to refining methodologies and embracing evolving healthcare trends remains unwavering, ensuring our efforts contribute meaningfully to the field of predictive healthcare analytics.

In the grand tapestry of healthcare analytics, our venture serves as a thread, weaving together data-driven insights and methodological rigor, as we collectively seek to illuminate the path towards proactive disease management and improved patient outcomes.

6. Contribution of Each Team Member

We worked on the project equally, sharing and collaborating to the best of our abilities.

References

Dhandhania, Keshav. “End-to-End Data Science Example: Predicting Diabetes with

Logistic Regression.” Medium, Towards Data Science, 24 May 2018,

towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logisti

c-regression-db9bc88b4d16.

Gulati, Aman Preet. “Diabetes Prediction Using Machine Learning.” Analytics Vidhya, 4

Jan. 2022,

www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/

Leoni, Edward. “Using Machine Learning to Predict If Someone Has Diabetes.” Medium,

Learn stuff with Ed, 12 Apr. 2020,

medium.com/edward-leoni/using-machine-learning-to-predict-if-someone-has-diabetes-21

dc52118d8f.

Teboul, Alex. “Diabetes Health Indicators Dataset.” Kaggle, 8 Nov. 2021,

www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.