Diabetes Prediction

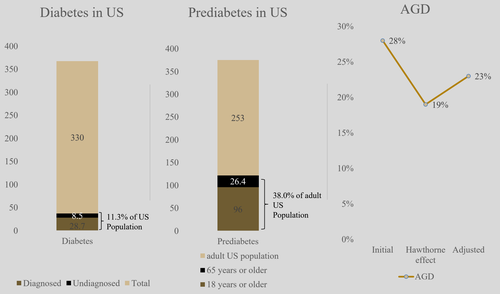
Zhaoming Hu
Purdue University, Daniels School of Business
hu787@purdue.edu

ABSTRACT

My motivation behind this project stemmed from the urgent need to identify individuals at high risk of developing diabetes. I aimed to develop a machine learning model for predicting the risk of diabetes based on various demographic and health factors. The dataset was collected from a diverse population, enabling me to conduct robust analysis and establish generalizability. My findings revealed significant associations between age, BMI, hypertension, and elevated blood glucose levels, which emerged as strong predictors, highlighting the importance of these factors in diabetes risk assessment. The machine learning model achieved promising accuracy, demonstrating its potential for aiding healthcare professionals in making informed decisions and improving patient outcomes.

BUSINESS PROBLEM

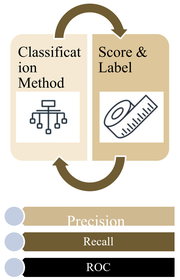
Diabetes has become a global health concern, affecting million of individuals and placing a significant burden on healthcare systems worldwide. Yet most people begin treatment only after diabetes is confirmed. The ability to predict diabetes risk accurately is crucial for early intervention, personalized treatment plans, and preventive measures.
Researchers and public health organizations can use these models to gain insights into risk factors and develop preventive strategies. Ultimately, patients will benefit from early detection, leading to improved health outcomes and a higher quality of life.



The effect of a comprehensive early intervention for diabetic patients was studied according to a randomized trial using electronic identification and bedside expert integrated diabetes team (IDT) management. The results showed that early identification and management significantly reduced AGD per patient (with a more liberal glycemic target) about 9%, indicating improved glycemic control in the hospital. In addition, the number of patient days with average blood glucose levels above the target threshold was significantly reduced. Therefore, I think it is necessary and effective to predict diabetes by the level of the patient's characteristics and to intervene early.

ANALYTICS PROBLEM FRAMING

The analytics problem centers around developing a predictive model for diabetes risk based on the available demographical and health factors. The assumptions in this project are based on the previous knowledge about the relationship between diabetes and various risk factors. Based on this, I use the dataset in classification-based machine learning models and clustering algorithms to correctly classify individuals as either having diabetes or being at risk. The success metrics are the accuracy of the model by using statistical measurement to do the evaluation and be utilized the assess the mode's performance comprehensively.



RESEARCH QUESTIONS

- What are the significant demographic and health factors that contribute to the risk of developing diabetes?
- Can a predictive model using the available demographic and health factors accurately classify individuals as having diabetes or being at risk of developing diabetes?

DATA

Source: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
Provenance: Electronic Health Records (EHRs) are the primary source of data for the Diabetes Prediction dataset. This dataset contains 100,000 patients' information along with their diabetes status (positive or negative).

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80.0	0	1	never	25.19	6.6	140	0
Female	54.0	0	0	No Info	27.32	6.6	80	0
Male	28.0	0	0	never	27.32	5.7	158	0

Raw Data — Input Diabetes Prediction.csv file to JMP Pro

Using Automated Data Imputaion fill in the missing value. → Data Cleaning

Convert gender to binary variable where males are represented by the number 1 and females by the number 0. Separate the smoking history column to three column, one for 'never', one for 'current', one for 'ever','former'and'not current'. → Data Type Conversion

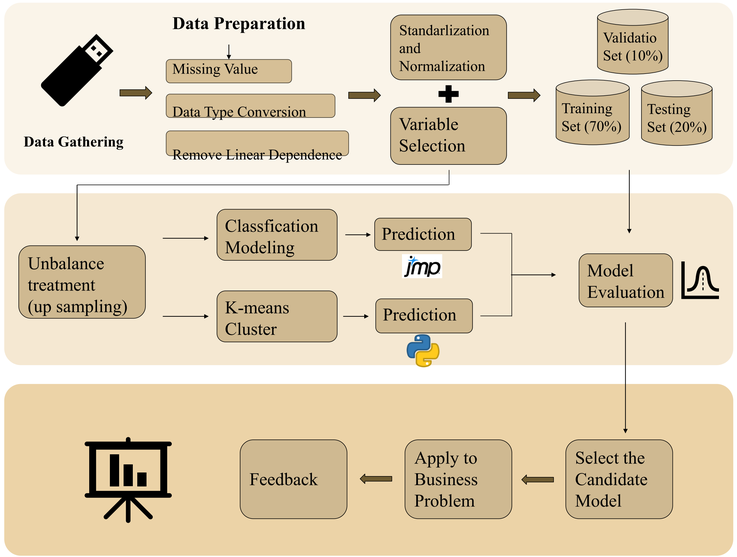
Apply standardization or normalization techniques to ensure variables are similar scale. → Standardizing

Analyze the relevance and importance of each variable in relation to the diabete. By using forward selection, gender is insignifiant. → Feature Selection

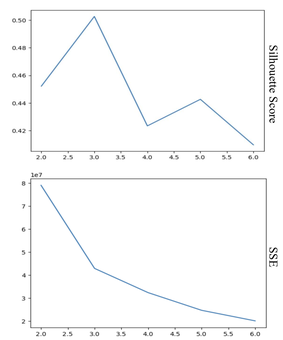
Using Cross-validate the cleaned dataset by performing data quality check, running basic descriptive statistics and validating against external sources to ensure accuracy. And I used 70:20:10 to randomly split the dataset. → Data Validation

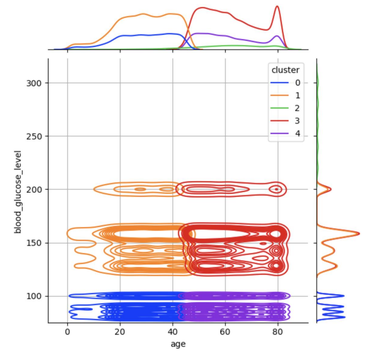
PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

METHODOLOGY



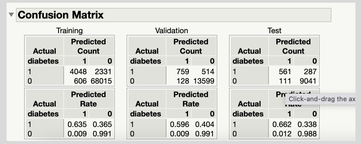
MODEL BUILDING AND EVALUATION – STATISTCIAL PERFORMANCE





After selecting 5 clusters as the optimal number based on the Silhouette Score and SSE charts, I applied the K-means algorithm to perform the diabetes prediction on the dataset. To evaluate the performance of the K-means model, I use several statistical metric. First, I calculated the Within-Cluster Sum of Squares, which measure compactness of the cluster. Next, I computed the Silhouette Coefficient, which assesses the quality of clustering y measuring the separation between clusters and the cohesion within clusters. Additionally, I evaluated the cluster centroids to understand the characteristics of each cluster. Finally, I used the joint plot to visualize the distribution between two variables base on the identified clusters. Specifically, I focused on the relationship between blood glucose level and age, which exhibited clear variation across the clusters. The joint plot revealed the certain clusters had higher blood glucose levels and older age groups, indicting a higher risk of diabetes.

MODEL EVALUATION – BUSINESS IMPLICATIONS



In the evaluation of my diabetes prediction model, I employed the sensitivity measure, which quantifies the ability of the model to correctly identify individuals with diabetes. Given the importance of early intervention and treatment in diabetes cases, accurately capturing true positive instances is crucial. In addition, the decision tree algorithm exhibited favorable performance in this context, as indicated by the model's R-square values. The low discrepancy between the training and test sets suggest that the decision tree model generalizes will.

According to the American Diabetes Association, the annual cost for a diabetes patient is approximately $16752, with $9601 attributed to diabetes management and $2926 related o complications arising from diabetes. On the other hand, a diabetes screening test typically costs around $70. It is worth nothing that the average medical expenditure for a US citizen is $4393. Based on my decision tree model, I achieved a 66% probability of accurately predicting individuals with diabetes and a 99.8% probability of correctly identifying individuals without diabetes.

CONCLUSIONS

- With an accuracy of 85% and sensitivity of 66%, indicating its potential to identify individuals at risk of developing diabetes.
- High specificity of 99.8% ensures a low false positive rate minimizing unnecessary interventions for individuals.

In terms of limitations, further analyses and future research could focus on expanding the dataset to include a larger and more diverse population, incorporating additional variables or risk factors, and conducting longitudinal studies to assess the long-term predictive accuracy of the model. Additionally, exploring ensemble methods or other advanced machine learning techniques may enhance the predictive performance and robustness of the model.