Machine Learning-Based Diabetes Prediction: Advancing Precision Healthcare with PIMA Dataset

Iram Wajahat[1], **Fazel Keshtkar**[2], **Syed Ahmad Chan Bukhari**[*2]
[1]Division of Computer Science, Mathematics and Science, Collins College of Professional Studies
[2]Institute of Biotechnology, Collins College of Professional Studies
St. John's University, Utopia,Parkway, Queens, 8000, NY, USA
wajahati@stjohns.edu, keshtkaf@stjohns.edu, bukharis@stjohns.edu*

Abstract

The prevalence of diabetes poses a significant global health challenge, demanding practical prognostic tools for timely intervention. This study explores the development of machine learning models for diabetes prognosis, utilizing the PIMA Indian dataset. Emphasizing early detection, the research underscores diabetes as modifiable through lifestyle adjustments. By analyzing diverse healthcare data, including electronic health records and imagery, machine learning algorithms can unveil crucial patterns for timely diagnosis. This project aims to identify significant features contributing to diabetes, develop a predictive model, and compare model performance. Statistical and machine-learning algorithms applied to the PIMA dataset reveal glucose levels as the foremost predictor of diabetes across all models, highlighting logistic regression's efficacy in feature extraction and prediction accuracy over random forest, K nearest neighbors (KNN), and deep neural networks. Omics data integration shows promise in enhancing deep learning model performance, offering robust diagnostic solutions for automated prognostic tools.

Keywords: Machine Learning, Diabetes Prediction, Precision Healthcare, PIMA Dataset

Introduction

Diabetes is a pervasive and escalating global health crisis, its prevalence and impact underscored by staggering statistics. In 2017 alone, approximately 450 million individuals received a diagnosis of diabetes, contributing to 1.37 million deaths worldwide (Cho, 2018) The United States, confronting over 100 million adults grappling with diabetes, witnessed the condition ascend to the rank of the seventh leading cause of mortality by 2020. Alarmingly, projections indicate a trajectory where by 2050, as many as one in three US adults could be affected by this metabolic disorder (CDC, 2020). Beyond mortality, diabetes exacts a toll through severe complications such as kidney failure, vision impairment, cardiovascular diseases, and limb amputations, accompanied by staggering economic burdens reaching billions of dollars annually (Krasteva, 2011)). Despite extensive research into the risk factors associated with diabetes, critical gaps persist, necessitating further investigation and refinement of predictive models. While various methodologies ranging from logistic regression to machine learning have been utilized to develop these models, challenges such as inadequate covariate selec-tion and model specification errors remain prevalent (Alghamdi , 2017). Moreover, concerns linger regarding the lack of objective and unbiased evaluation, impeding the reliability and widespread adoption of existing predictive models (Nguyen, 2019). In this study, our aim is to address these gaps by adopting a comprehensive approach that combines traditional logistic regression with advanced machine learning techniques to predict the risk factors associated with type 2 diabetes mellitus. Through meticulous evaluation and validation procedures, our goal is to identify significant predictors while ensuring the robustness and generalizability of our models, with a particular emphasis on elucidating the unique diagnostic factors prevalent within the PIMA Indian community. By leveraging diverse methodologies and datasets, our study endeavors to enhance the accuracy and reliability of predictive models, furnishing healthcare practitioners and policymakers with actionable insights to effectively prevent, diagnose, and manage diabetes (Habibi , 2015).

Methods

To address our research question, we utilized the PIMA Indian Diabetes dataset, consisting of data on females aged 21 and above. This dataset comprises nine features, including the outcome variable, which indicates the presence or absence of diabetes. The features include the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin levels, body mass index, diabetes pedigree function, and age (Figure. 1).Through exploratory data analysis and preprocessing, we aimed to understand the relationships between these features and the outcome of interest (Zehra, 2014).

Statistical Analysis: We conducted statistical tests to compare the distribution of some features (glucose, blood pressure) between diabetic and non-diabetic groups. We did an unpaired t-test. **Hypothesis 1:** The mean glucose levels for diabetic and non-diabetic individuals are the same. The mean glucose levels for diabetic and non-diabetic individuals are the same. **Group 1:** Diabetic individuals and **Group**

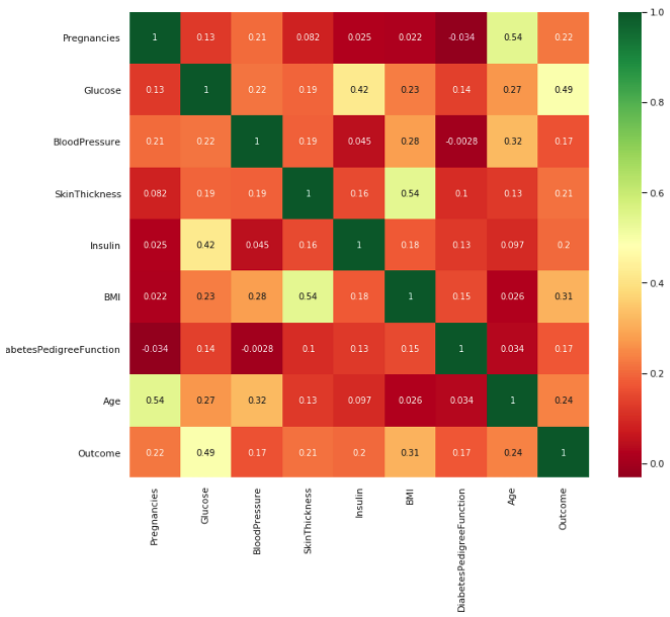

Figure 1: *Heatmap presents an overview of the quality of the PIMA Indian diabetes dataset across different variables*

2: Non-diabetic individuals. In this case, we performed two-sample t-tests to compare the means of two independent groups. Null Hypothesis (H0): The mean glucose levels for diabetic and non-diabetic individuals are equal.·Alternative Hypothesis (H1): The mean glucose levels for diabetic and non-diabetic individuals are unequal. **Hypothesis 2:** There is no significant difference in blood pressure between diabetic and non-diabetic patients. In this case, the t-value is negative, suggesting that the mean blood pressure level in group 0 (non-diabetic individuals) is lower than in group 1 (diabetic individuals), but the difference is not very large. The p-value is 0.08735. This value is not less than the conventional significance level of 0.05. Therefore, we do not have sufficient evidence to reject the null hypothesis. This suggests that the two groups may not significantly differ in mean blood pressure levels (Figure 2). (Lakens, 2013)

Finding Optimal Machine Learning: Subsequently, we employed four predictive models: logistic regression, random forest, k-nearest neighbors, and deep neural networks, to ascertain the most important predictors of diabetes and evaluate the performance of each model (Miao, 2021)(Chang, 2023) . Our analysis sought to elucidate the impact of family history on diabetes risk and determine the optimal model for diabetes prediction based on the PIMA dataset (Figure 3).

Results and Conclusion

Our research highlights glucose levels as the top predictor of diabetes, with diabetic individuals showing significantly higher levels compared to non-diabetics. There's no strong evidence for a difference in mean blood pressure between the two groups. Glucose, BMI, and age are significant predictors of diabetes, while blood pressure shows a weaker association. Logistic regression proves most effective in predicting diabetes, achieving an F1 score of 0.67. Deep neural networks show potential despite small datasets, indicating adaptability. Our findings align with previous studies emphasizing the importance of glucose, BMI, and age in diabetes prediction. While logistic regression is emphasized, other models like random forest and k-nearest neighbors also perform well, highlighting the need for diverse modeling approaches. Our study contributes to diabetes prediction by showcasing the potential of deep learning techniques, especially in healthcare analytics with limited data.

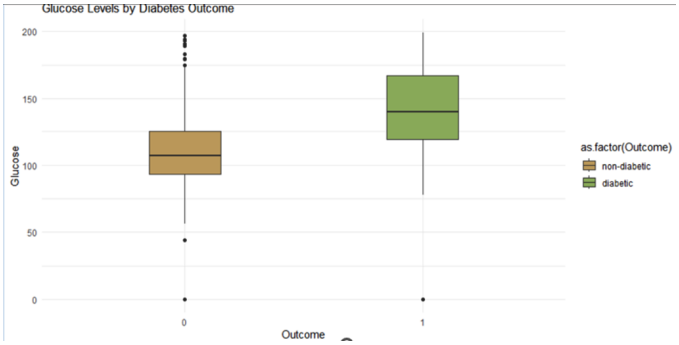

Figure 2: *The box plot summarizes the distribution of glucose levels for diabetic and non-diabetic individuals. It shows the median (middle line), spread (box), and outliers (individual points) for each group, allowing for easy comparison between the two categories.*

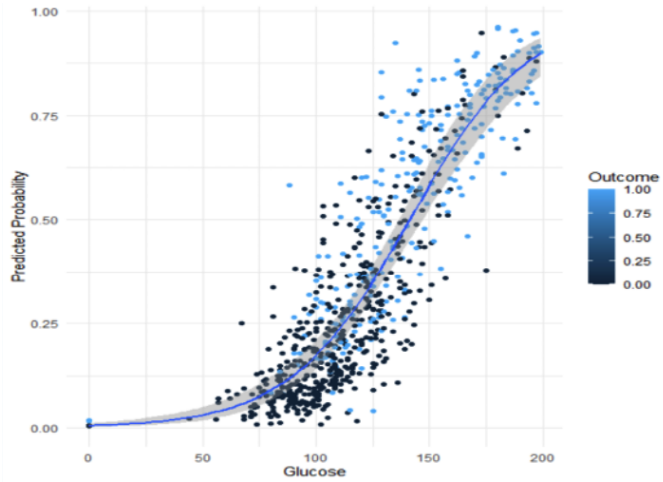

Figure 3: *Our experiments suggest that Logistic Regression outperforms the K nearest neighbors (KNN), and deep neural (DL) networks*

References

Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pract. 2018, 138,

271–281.

CDC. Centers for Disease Control and Prevention and Others; National Diabetes Statistics Report; Centers for Disease Control and Prevention, US Department of Health and Human Services: Atlanta, GA, USA, 2020; pp. 12–15.

Krasteva, A.; Panov, V.; Krasteva, A.; Kisselova, A.; Krastev, Z. Oral cavity and systemic diseases—Diabetes mellitus. Biotechnol. Biotechnol. Equip. 2011, 25, 2183–2186.

Alghamdi, M.; Al-Mallah, M.; Keteyian, S.; Brawner, C.; Ehrman, J.; Sakr, S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. PLoS ONE 2017, 12, e0179805.

Nguyen, B.P.; Pham, H.N.; Tran, H.; Nghiem, N.; Nguyen, Q.H.; Do, T.T.; Tran, C.T.; Simpson, C.R. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput. Methods Programs Biomed. 2019, 182, 105055.

Habibi, S.; Ahmadi, M.; Alizadeh, S. Type 2 diabetes mellitus screening and risk factors using decision tree: Results of data mining. Glob. J. Health Sci. 2015, 7, 304.

Zehra, A., Asmawaty, T., & Aznan, M. A. M. (2014). A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset. In 3rd International Conference on Software Engineering & Computer Systems (pp. 1-10).

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in psychology, 4, 62627.

Miao, Y. (2021, March). Using machine learning algorithms to predict diabetes mellitus based on Pima Indians Diabetes dataset. In Proceedings of the 2021 5th International Conference on Virtual and Augmented Reality Simulations (pp. 47-53).

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157-16173.