National College of Ireland

Project Submission Sheet

Student Name (s): Bolormaa Mendbayar, Thapelo Emmanuel Khantsi, Temitope Oladimeji

Student ID: x23176725, x23131535, x23187204

Programme: MSc in Data Analytics

Year: 2024

Module: Domain Application & Predictive Analytics

Lecturer: Vikas Sahni

Submission Due Date: 23/02/2024

Project Title: Project Design – Diabetic Health Risk Indicator

Word Count: 1975

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Bolormaa, Thapelo, Temitope

Date: 18/02/2024

Diabetic Health Risk Indicators

Bolormaa Mendbayar - x23176725
Thapelo Khantsi – x23131535
Temitope Oladimeji – x23187204

National College of Ireland
MSc in Data Analytics
(MSCDAD_A)
Domain Application in Predictive
Analytics

Abstract — In this research, socio-demographic factors contributing to the risk of diabetes among individuals are explored, aiming to develop predictive models for assessing the likelihood of diabetes and providing insights for healthcare providers. Ethical considerations surrounding privacy, algorithm bias, and data accuracy are addressed, emphasizing the importance of upholding ethical principles, and promoting fairness in predictive analytics for healthcare. Business value is addressed, as well as the applicability of machine learning techniques, complemented by a few visualizations.

Keywords **—** Predictive modelling, diabetes, machine learning, ethics, naïve bayes, neural networks, logistic regression, SVM, random forest, decision trees.

I. Introduction

The U.S. struggles with diabetes, a chronic disease affecting millions and straining healthcare systems. Individuals with diabetes struggle to regulate blood sugar, leading to potential complications like heart disease, vision loss, and amputations. The progression and complications of diabetes are often caused by elevated blood glucose levels [1]. While incurable, managing weight, diet, and activity, alongside medical treatment, can lessen the impact, early diagnosis is crucial, enabling lifestyle changes and better treatment [2]-[3]. The CDC estimates 34.2 million Americans had diabetes in 2018, with 88 million having prediabetes, often undiagnosed [4].

Type 2 diabetes, the most common form, varies in prevalence across demographics. Lower socioeconomic groups often bear the heaviest burden [4]. The economic cost is staggering: diagnosed diabetes alone costs an estimated $327 billion annually, with total costs approaching $400 billion when including undiagnosed cases and prediabetes [5].

A. Scope

The scope of this project was designed following the CoNVO framework.

- **Context**: Diabetes disease is one of the most chronic diseases of which its prevalence has been rising rapidly. This is due to its nature of being dynamic and complex. As mentioned earlier in the report; diabetes has a great financial burden on healthcare. Predicting the disease progression is also still a challenge as of today.
- **Need**: Based on the issues identified for diabetes. There is need to reduce the progression of diabetes using demographic factors and bio markers to detect and diagnose the disease early. This will benefit healthcare professional in tailoring personalised treatment plans for patients.
- **Vision**: The predictive model proposed and built from this project should contribute to the domain of health by adding better Predictive capabilities to the diabetes diagnosis and identifying patients at risk. Recent reports state that there is a financial burden caused by this disease, therefore the outcomes of these project should contribute to cost reduction and resource optimisation in the health sector.
- **Outcome**: The final deliverable of the project will be a predictive model that uses both machine learning and deep learning concepts to allow diabetes health risk indicators to be easily diagnosed at an early stage and prevent complications aided by the predicting the progression of the disease.

II. Goal of The Project

The goal of the project is to utilize data analytics techniques to explore and identify Socio-Demographic factors contributing to the risk of diabetes among individuals. The primary objectives include developing predicting models to assess the likelihood of diabetes based on health related, social and demographic variables, as well as giving healthcare providers practical insights on how to effectively control and reduce the risk of diabetes on their patient groups.

The study will examine a variety of lifestyle, medical, and demographic variables associated with diabetes risk, such as body mass index (BMI), smoking status, physical activity, and pre-existing medical disorders, cholesterol, blood pressure, mental health, education, income, age and gender etc. By applying predictive analytics and machine learning algorithms, we aim to uncover patterns and relationships within the data that can help healthcare providers.

With the help of the study, we hope to provide prediction models that may correctly identify people who are more likely to develop diabetes, as well as interpretive summaries and visualizations to help medical professionals comprehend and make decisions. By aligning the research with the business value of improved patient outcomes and healthcare management efficiency, we aim to contribute to the advancement of data-driven healthcare solutions.

III. Ethical Concerns

Understanding and adhering to ethical considerations is important in making the best possible decisions. In healthcare, researchers have outlined the significance of professionals to understand and comply to medical regulations [6]. Though different professions have their own ethical principles, D'Souza [7] states the importance of maintaining a professional conduct with workers and patients alike.

In a narrow sense, these ethical concerns also extend to the predictive testing of diabetes. With an estimated 450 million diagnosed with the disease [8], severe complications can occur. Research on diabetes still has to adhere to privacy regulations, as medical records and patient data must be collected and anonymized where possible. More importantly, healthcare personnels must avoid medical discrimination based on demographic factors such as age, gender or race [9]. Data sharing with or without patient consent can be a common practice within facilities, and care must be taken to avoid mishandling of sensitive health data, and a potential risk of data breach. In a recent study [10], the issue of inequity was also addressed, emphasizing the moral caution when conducting research of such level.

There are ethical challenges to be addressed in predictive analytics used for this research, associated mainly with algorithm bias which may lead to unfair, or even false, outcomes. Privacy concern is a key AI issue [11], and the use of machine learning poses significant risks to the protection of data. Delving deep into AI's ethical concerns, pattern recognitions might violate privacy regulations, seen in Jernigan and Mistree's study [12].

While biased outcomes are a possibility, machine learning models might also be susceptible to inaccurate data from the start. These models learn from historical data [13] which already might contain biases or discrimination. Feeding this type of data into the AI model could lead to disparities in results. However, this can be mitigated by cleaning the data before performing any technique on the data, though it is essential to ensure every form of bias is addressed from the beginning. Promoting fairness and upholding ethical principles remains a vital part of equality, ensuring individuals are safe in every regard.

IV. Business Value Of The Project

Diabetes predictions using machine learning concepts have significant business value. These predictions can help medical professionals make accurate early predictions and judgments, leading to better management of diabetes and improved patient outcomes [14]. When prevention and intervention is done early, the likelihood of disease development is brought to halt from the onset: Eliminating progression all together. More importantly, the timely intervention will lead to predicting severity of the existing conditions, potentially saving lives. The predictive analytics will bring valuable insights to the healthcare sector, enhancing patient care and hospital performance. The doctors will manage to prioritise high risk patients saving costs and time in diagnosis. They will also have a clinical foresight of identifying those who are likely to have chronic conditions resulting into preventive measures. Future events can also be predicted based on the data collected, and this helps improve decision making activities through data driven predictive analytics [14]-[15]. Events such as extended hospital stays and readmissions will be reduced benefiting both patients and healthcare systems.

Resource management will be optimized due to efficient supply chain. When demand is correctly predicted, there is less waste which improves supply chain. The resources will be allocated correctly as healthcare providers would manage to anticipate who are the actual patients that need to be treated and who are those that do not classify as risk patients.

The predictive analytics of diabetes health risk indicators will result in healthcare moving from reactive care to proactive prediction and prevention. Reduced healthcare costs by avoiding avoidable hospitalizations and complications in patients.

V. Preliminary Visualisations

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. For this project, a CSV of the dataset available on Kaggle for the year 2015 was used. The dataset contains 253,680 records and 21 feature variables. The target variable Diabetes_012 has 3 levels 0-2, and 0 is for no diabetes, 1 is for prediabetes, and 2 is for diabetes.

Firstly, missing values and duplicated records were dropped from the dataset using R programming, resulting in 229,782 records.

All the visualizations were used PowerBI tool, and for understanding dataset, we visualized initially distribution of diabetes level to see whether dataset is balanced or not. As below shown in Fig.1, roughly 82% is non-diabetes, around 15% is diabetes, and the rest of 2% is prediabetes, so we can say dataset is unbalanced.

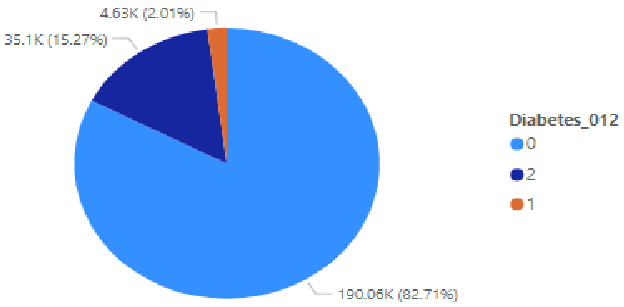

Fig. 1. Distribution of Diabetes level

As illustrated in Fig. 2, a correlation matrix is visualized between independent and dependent variables. Variables such as as High Blood Pressure(HighBP), High Cholesterol(HighChol), Body Mass Index(BMI), and General Health(GenHlth), and Difficult of walking(DiffWalk), with correlation coefficients above ±0.2, are considered correlated to the target variable, Diabetes_012.

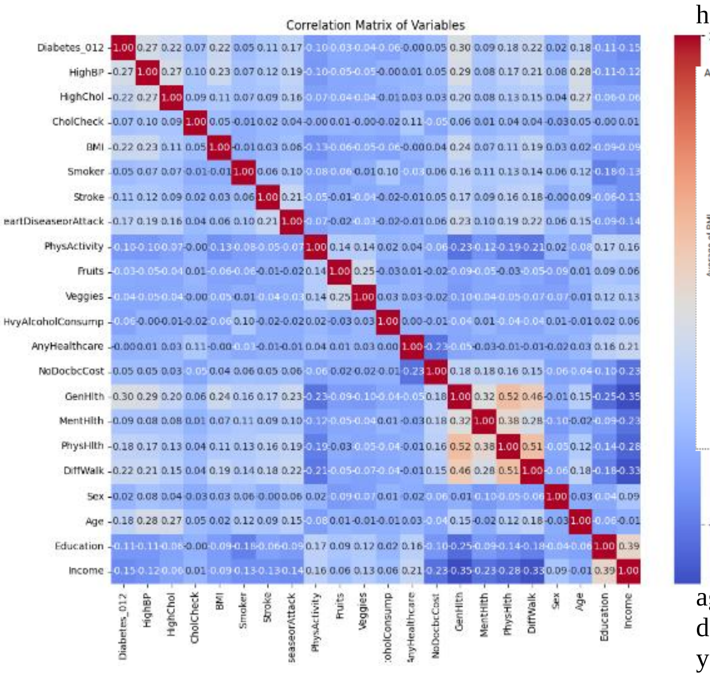

Fig. 2. Correlation matrix

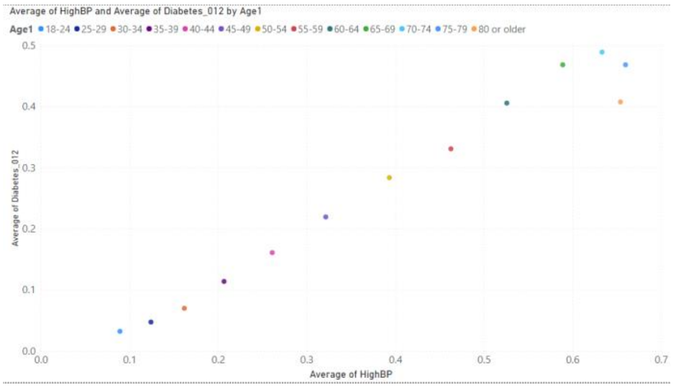

Fig. 3. HighBP and Diabetes by Age

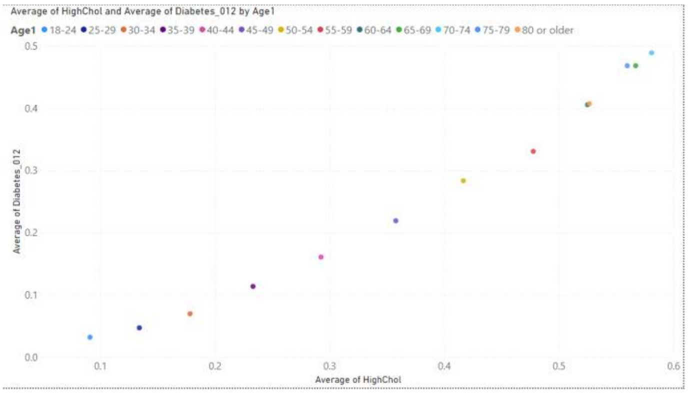

Fig. 4. HighChol and Diabetes by Age

From Fig. 3 and Fig. 4, we observe that as age increases, measurements of High Blood Pressure (HighBP) and High Cholesterol (HighChol) also increase. This indicates that compared to younger individuals, older people have a higher risk of developing Diabetes.

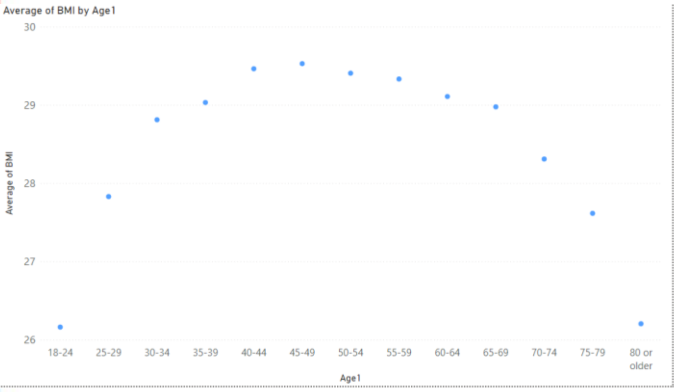

Fig. 5. BMI and Diabetes by Age

In Figure 5, the BMI value consistently rises until the age of 45-49 which is middle-aged, after which it begins to decline. Additionally, it is evident that both the oldest and youngest age groups exhibit similar BMI levels.

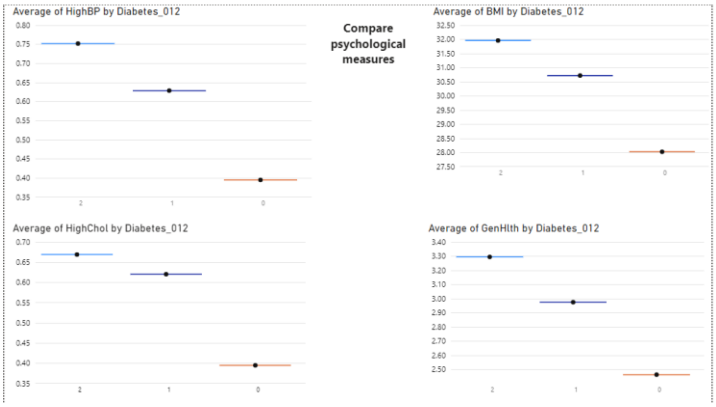

Fig. 6. Psychological measures on HighBP, HighChol, BMI and GenHlth

Figure 6 displays the averages of HighBP, HighChol, BMI, and GenHlth factors. We observe that values exceeding approximately 0.63 for HighBP, 0.61 for HighChol, 30.75 for BMI, and 2.95 for GenHlth are indicative of prediabetes.

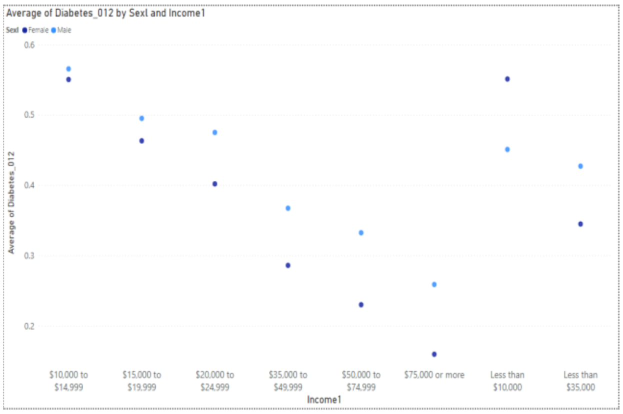

Fig. 7. Income and Diabetes by Sex

We hypothesize that individuals with higher incomes have a lower risk of diabetes, while those with lower incomes face a higher risk. As illustrated in Figure 7, we have demonstrated that individuals with an income exceeding $35,000 experience a lower risk compared to those earning less. Specifically, the income range between $10,000 and $14,999 exhibits the highest risk of diabetes, whereas an income of $75,000 or more correlates with a lower risk.

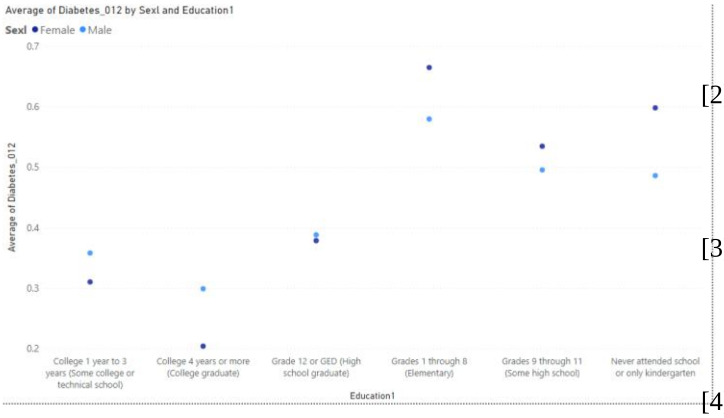

Fig. 8. Education and Diabetes by Sex

As we hypothesized regarding income, similarly, in the case of education, we have also observed that individuals who have attended college have a lower risk of diabetes compared to those who have never attended or only attended kindergarten.

VI. Lists Of Applicable Techniques

Research on diabetes prediction has covered the use of multiple machine learning techniques. KDD appears to be the common framework for such analysis [17], where classification algorithms take priority due to their broad use in the medical field [18]. Its strength in handling large volumes of data and in the diagnosing of diseases validates researcher's use of machine learning [18].

Sisodia [18] proposed three classification algorithms, namely SVM, Naïve Bayes, and Decision Tree classifiers in their work on the prediction of diabetes. Similarly, Khanam and Foo's work [19] analysed seven different techniques to predict diabetes, with Linear Regression and SVM having the best performance. In recent years, neural networks have also become widely used in the medical field, with the prediction of birth weight, Parkinson's, and diabetes, being researched [20]-[21]-[22]. The increasing use of neural networks reflects the trend in using advanced methods for medical diagnosis, highlighting their ability to handle huge datasets and diagnose diseases accurately.

The preference for classification techniques in diabetes prediction emphasizes their proven suitability based on multiple conducted research. Ensemble methods have been used by researchers [17[- [19] in disease prediction due to their classification accuracy. These techniques have all performed well to an extent, with comparable datasets, and thoroughly analysed to justify their use. In line with the methodologies observed in prior research, the study similarly adopts a classification-focused approach within the KDD framework to predict diabetes.

VII. References

[1] A. Alzaid, P. Ladrón de Guevara, M. Beillat, V. Lehner Martin, and P. Atanasov, "Burden of disease and costs associated with type 2 diabetes in emerging and established markets: systematic review analyses," *Expert Review of Pharmacoeconomics & Outcomes Research*, vol. 21, no. 4, pp. 785-798, 2021.

[2] N. A. ElSayed *et al.*, "1. Improving Care and Promoting Health in Populations: Standards of Care in Diabetes—2023," *Diabetes Care*, vol. 46, no. Supplement_1, pp. S10-S18, 2023.

[3] N. A. ElSayed *et al.*, "3. Prevention or delay of diabetes and associated comorbidities: Standards of care in diabetes—2023," *Diabetes Care*, vol. 46, no. Supplement_1, pp. S41-S48, 2023.

[4] A. W. Cdc, "Centers for disease control and prevention," ed, 2020.

[5] "Diabetes - NIDDK," *National Institute of Diabetes and Digestive and Kidney Diseases*. https://www.niddk.nih.gov/health-information/diabetes#topics

[6] G. D. Pozgar, *Legal and ethical issues for health professionals*. Jones & Bartlett Learning, 2023.

[7] R. D'Souza, "Ethical Issues associated with diagnosing and managing diabetes," *Glob. Bioeth. Enquiry*, vol. 2017, pp. 52-56, 2017.

[8] G. Roglic, "WHO Global report on diabetes: A summary," *International Journal of Noncommunicable Diseases*, vol. 1, no. 1, pp. 3-8, 2016.

[9] S. B. Haga, "Ethical issues of predictive genetic testing for diabetes," *Journal of diabetes science*

and technology, vol. 3, no. 4, pp. 781-788, 2009.

[10] E. Shayo *et al.*, "Ethical issues in intervention studies on the prevention and management of diabetes and hypertension in sub-Saharan Africa," *BMJ Global Health,* vol. 5, no. 7, p. e002193, 2020.

[11] B. C. Stahl, *Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies*. Springer Nature, 2021.

[12] C. Jernigan and B. F. T. Mistree, "Gaydar: Facebook friendships expose sexual orientation," *First Monday,* 2009.

[13] T. H. Davenport, *The AI advantage: How to put the artificial intelligence revolution to work*. mit Press, 2018.

[14] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science,* vol. 165, pp. 292-299, 2019.

[15] K. D. Wowak, J. P. Lalor, S. Somanchi, and C. M. Angst, "Business Analytics in Healthcare: Past, Present, and Future Trends," *Manufacturing & Service Operations Management,* vol. 25, no. 3, pp. 975-995, 2023.

[16] P. Maheswari, A. Jaya, and J. M. R. S. Tavares, "Business Intelligence and Analytics from Big Data to Healthcare," *Handbook of Intelligent Healthcare Analytics: Knowledge Engineering with Big Data Analytics,* pp. 115-145, 2022.

[17] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal,* vol. 15, pp. 104-116, 2017.

[18] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science,* vol. 132, pp. 1578-1585, 2018.

[19] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *Ict Express,* vol. 7, no. 4, pp. 432-439, 2021.

[20] M. Al-Shawwa and S. S. Abu-Naser, "Predicting birth weight using artificial neural network," 2019.

[21] R. M. Sadek *et al.*, "Parkinson’s disease prediction using artificial neural network," 2019.

[22] N. S. El_Jerjawi and S. S. Abu-Naser, "Diabetes prediction using artificial neural network," 2018.