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**:** 29/04/2024

Project Title: Project Report – Diabetic Health Risk Indicator

Word Count:

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:

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. You must ensure that you retain a HARD COPY of ALL projects, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. Late submissions will incur penalties.
5. All projects must be submitted and passed in order to successfully complete the year. Any project/assignment not submitted will be marked as a fail.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Project Report – 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 & Predictive Analytics

***Abstract*— This research investigates the influence of socio-demographic factors on diabetes risk among individuals, focusing on the implementation and interpretation of Artificial Neural Network (ANN) modeling. The study goal is to achieve accurate classification of diabetes across all three factors. Through comprehensive data analysis and advanced computational techniques, the research aims to provide actionable insights into mitigating diabetes risk and improving healthcare outcomes. Furthermore, the study explores the business value of accurate diabetes risk prediction, highlighting potential cost reductions, early interventions, and enhanced resource allocation in healthcare settings. By integrating socio-demographic insights with predictive analytics, this research contributes to a deeper understanding of diabetes risk factors and their implications for healthcare policy and practice.**

***Keywords*—Diabetes risk assessment, ANN, Classification, Business value**

I. Research into Applicable Techniques

The development of diabetes can cause serious damage to the organs. It has become a global concern, and a rise in numbered cases appears to be inevitable as millions of people are expected to be diagnosed with the disease within the next few years [1]. Patients diagnosed with diabetes are usually classified as having either a type-1, where not enough insulin is produced leading to an almost complete functional loss of the pancreas, or a type-2, where insulin resistance has been developed which leads to glucose deprivation, subsequently causing weakened bones, or a lack of growth especially in children [2]. Insulin injections are usually taken to cure the disease at its early stage, but a proper diet as well as a healthy lifestyle is highly recommended.

Assessing patients who are at risk of contracting the disease is not very straightforward. Factors such as stress levels, diet, and weather have been considered significant enough to complicate any form of diabetes management or treatment. The use of predictive analytics in this context can help identify patients who are most at-risk, manage the disease more effectively, and provide quick intervention. Machine learning is necessary for these improvements, and it is fast becoming a widely accepted tool in medical diagnosis, especially in detecting diseases such as diabetes [3].

Different models have been identified as being suitable for diabetes risk-level predictions. Research has predicted the likelihood of patients being diagnosed with the disease, or its risk factors, using LSTM, RNN, and XGBoost among others. Ensemble methods have typically been preferred for many applications due to their increased accuracy and reliability of their results. Most research focuses on comparative analyses to investigate which methods are most preferable [4]-[5]. This study is informed by the work of others, focusing on limited artificial neural network (ANN) deployment in diabetes prediction, to bridge that existing gap in knowledge. The use of ANN eases this classification process and is widely accepted as an efficient technique for predicting diabetes risk levels [6].

The succeeding section details the methodology used to carry out the analysis, as well as a rigorous interpretation of the application. Section III discusses the quantitative findings and provides conclusions appreciating the study's limitations and implications. The business value and the qualitative interpretation of findings encompass section IV.

II. Implementation of the Technique

Few alternatives for implementation were considered before deciding to use ANN.

- Support Vector Machine: This is effective in classification tasks for finding optimal hyperplanes to separate data points belonging to different classes. This model struggled with high-dimensional data and ANN handled this complexity better.
- Decision Trees: The model uses a series of branching conditions to classify data or predict continuous values. The downside of the model is that it is not easily adaptable to new information, but ANN on the other hand can continuously learn and improve with additional data.
- KNN: The method classifies data based on the similarity to their nearest neighbor in the training dataset. Finding the nearest neighbors can be computationally expensive and ANN is more efficient in processing large amounts of data.
- Random Forest: It is an ensemble learning method that combines different trees to improve the performance and reduce overfitting. This model would have been the most preferable one, but RF usually provides a challenge when trying to decode how predictions are made. Therefore, ANN provides

a much clearer picture of predictions through techniques like visualizing the feature importance within the network.

Considering all the possible techniques, ANN was still the best and optimal technique to use as interpretability, flexibility, and adaptability were preferable for the use case in the health domain.

The implementation of an ANN algorithm, much like any other algorithm, involves the pre-processing of the data to be used. This involves cleaning the data, addressing any missing values present in the dataset. The features were normalized to ensure that they are on a similar scale to prevent specific features from dominating the learning process.

Exploratory data analysis (EDA) is an important step that precedes the implementation of any technique. This involves the visualization of any linear or non-linear relationships between variables. Clustered representations could also be useful here to uncover any additional insights. Histograms, boxplots, and density plots are common visuals used in this preprocessing stage [7].

Prior to the model evaluation phase, the data is separated into two subsets – the first is the training set and is used to build the model, the other is the testing set which is used for the prediction and assesses how well the model performs on unseen data. This helped in evaluating how well the model generalizes to new information. The model's performance is typically evaluated using the following metrics: accuracy, precision, recall, and F1-score. These are common assessment measures for classification tasks. These can also be composed in a table format known as the confusion matrix. Modelling of ANN required defining the network architecture to determine the number of neurons in the input layer, which corresponded to the various diabetes risk factors present in the dataset used.

The study conducted in this paper uses this approach to accurately investigate significant factors that contribute to diabetes, with the aim of confidently providing reliability and practicality in real-world settings. The tools used for analysis would be primarily carried out using Python and its necessary libraries.

III. QUANTITATIVE FINDINGS

The quantitative findings of the study reveal insightful patterns and factors influencing diabetes risk levels. Through the implementation of an Artificial Neural Network (ANN) model using Google Collab, results were obtained that contributed to the existing understanding of diabetes prediction and management.

1. Exploratory Data Analysis (EDA) Insights:

As in CA1, the preprocessing stage of dataset involved removing missing values and outliers. The exploratory data analysis (EDA) revealed notable correlations among different demographic, lifestyle, and health-related variables like HighBP (High Blood Pressure), HighChol (High Cholesterol), GenHlth (General Health), and BMI (Body Mass Index) to diabetes.

2. Model Performance Evaluation:

By incorporating all features into the ANN model allows for the full information captured in the dataset to be leveraged. This approach enables consideration of wide range of demographic, lifestyle, and health related factors that may influence diabetes risk, ensuring a comprehensive analysis of the predictive factors. ANN models are known for their flexibility and adaptability to diverse datasets and problem domains. As well as that, ANN model can effectively handle high-dimensional data and capture complex patterns and interactions, making them well-suited for this analysis.

Below are shown the performance metrics of ANN model, indicating its efficacy in predicting diabetes risk levels. Note: The target variable which diabetes has 3 levels such as 0 is no diabetes, 1 is prediabetes, 2 is diabetes.

Table 1. Metrics values

Metrics values of ANN model		
Accuracy	0.702	
	Each level	Overall
Sensitivity	[0.837 0.382 0]	0.406
Specificity	[0.897 0.034 0]	0.310
F1-score	[0.866 0.062 0]	0.309

Accuracy measures the proportion of correctly classified instances among all instances. An accuracy of 0.702 indicates that the model correctly predicts the class label for approximately 70.2% of the instances.

Sensitivity (also known as recall) measures the model's ability to correctly identify true positive instances among all actual positive instances. This indicates how well the model is correctly identifying individuals with diabetes. An overall sensitivity of 0.406 suggests that the model is moderately effective in identifying positive instances across all classes. However, it varies significantly across different classes, with the highest sensitivity for the first class (0.837).

Specificity measures the model's ability to correctly identify true negative instances among all actual negative instances. This metric gives context into whether the model is correctly identifying individuals who do not have diabetes cases. A specificity of 0.310 indicates that the model is less effective in identifying true negative instances, and it varies across classes. It is worth noting that specificity is not commonly used in multiclass classification and might be less informative in this context.

The F1-score is the integrated means of precision and recall and provides a balanced measure of the model's accuracy. An overall F1-score of 0.309 indicates the model's effectiveness in terms of precision and recall across all classes. However, it is modest, indicating an opportunity for enhancing the model's performance.

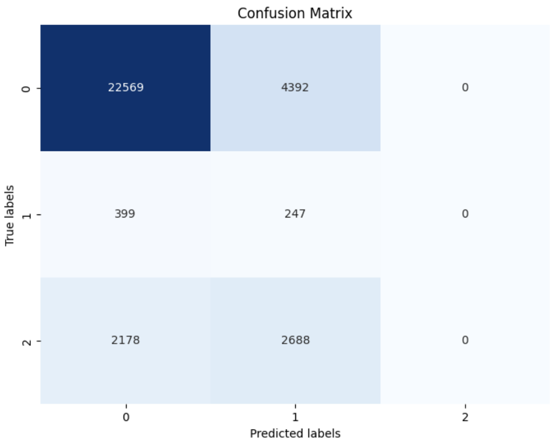

Figure 1. Confusion matrix

From Figure 1, this analysis provides insight into the model's performance in classifying instances across the three levels of diabetes status.

The confusion matrix reveals an apparent class imbalance, especially for the "diabetes" level (2), where the model did not make any correct predictions. This indicates a potential issue with the model's ability to distinguish this class from others.

The model shows a tendency to misclassify instances, particularly between the "no diabetes" (0) and "prediabetes" (1) classes. This is evident from the relatively high number of misclassifications between these two classes.

Overall, the model's performance appears below standard, especially in correctly identifying instances of "diabetes" (2). The low number of correct predictions for this class indicates a significant limitation in the model's ability to accurately classify diabetic cases.

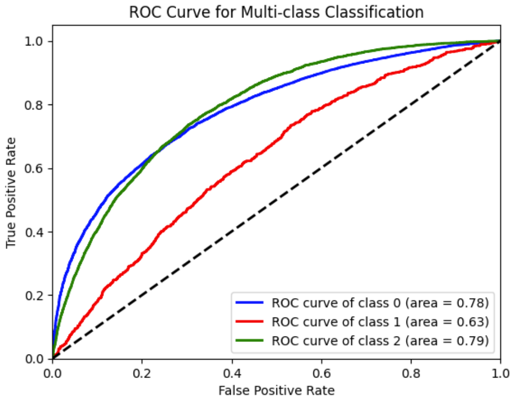

Figure 2. ROC curve for all class

The AUC value quantifies the overall performance of the model. An AUC value closer to 1 indicates a better-performing model, as it suggests higher sensitivity and specificity across various threshold values. Conversely, an AUC value closer to 0.5 indicates poor discrimination ability, resembling a random classifier. As shown above in Figure 2 indicates that levels 0 and 2 demonstrate relatively strong differentiation ability, whereas level 1 shows a weaker performance.

IV. The Business Value Qualitative Interpretation Of The Findings

According to numerous studies, deep learning methods have been shown to achieve superior results in various tasks. These techniques demonstrably reduce classification errors and exhibit greater resilience to noise compared to other approaches [8].

ANN has the advantage of analyzing large datasets quickly, and in the healthcare domain, this will bring value in freeing up the doctors' time to focus on patient care and healthcare resources can be allocated effectively by managing to pinpoint the high-risk patients from the low or moderate risk patients. Since there is no cure yet for diabetes, the capability of ANN in analyzing a vast amount of medical data could add value for identifying potential targets for new diabetes medications in the research domain for the healthcare domain.

Consequently, from the ability to identify high-risk individuals, ANNs can help in early detection and intervention altogether avoiding high costs associated with diagnosing diabetes at later stages. A general observation in healthcare is that delayed diabetes detection leads to patients having complications which leads to prolonged stays at hospitals. The proposed ANN model can assist in faster identification of diabetes health indicators; meaning reduced hospital stays due to faster, efficient, and early diagnosis from the model’s ability to not be biased by human preconceptions or emotions as well as identifying patterns in data accurately.

Currently, in the healthcare domain, health practitioners are using a reactive treatment approach due to diabetes being diagnosed at later stages. The developed ANN model alternatively provides the advantage of a proactive treatment approach to the healthcare domain, allowing for more data-driven decisions made from trends and patterns identified in the medical data.

This model provides a novel approach with immense business value in the healthcare domain. The traditional risk assessments in healthcare rely on averages, and often the missing variations and ANN adapts to this by learning complex patterns from vast medical records, genetic information, lifestyle factors, and environmental data. As a result, patients can be easily profiled and classified for personalized risk assessments and medication. Additionally, Artificial neural networks (ANNs) are particularly adept at handling large datasets as mentioned, leading to improved accuracy, and surpassing the performance of many classical machine learning models in various tasks.

V. Conclusions and Future Work

Diabetes risk factors often have complex, non-linear relationships and ANN exceled at learning these complex patterns from the data selected. With its capability of automated feature extraction, ANN reduced the need for extensive manual feature engineering of the diabetic risk indicators which proves to be more useful in the setting of real-life application in healthcare domain for automation processes.

The findings suggest potential areas for improvement in the model's training process, feature selection, or model choice. Addressing the class imbalance and improving the model's ability to differentiate between classes may enhance its overall performance. Further investigation into the specific factors contributing to misclassifications, such as feature importance and model biases, may provide valuable insights for refining the model and improving its predictive capabilities.

Additionally, exploring techniques like data augmentation or ensemble methods could mitigate class imbalances and enhance the model's generalization capabilities. Interpretability techniques such as LIME (Local Interpretable Model-agnostic Explanations) can unveil hidden patterns or biases within the data. These combined efforts can contribute to building more reliable machine learning models capable of addressing complex real-world challenges.

VI. References

[1] W. Bielka, A. Przezak, P. Molęda, E. Pius-Sadowska, and B. Machaliński, “Double diabetes—when type 1 diabetes meets type 2 diabetes: Definition, pathogenesis and recognition,” Cardiovascular Diabetology, vol. 23, no. 1, Feb. 2024. doi:10.1186/s12933-024-02145-x

[2] H. Klandorf and S. W. Stark, “Diabetes mellitus.,” Magill’s Medical Guide (Online Edition) | EBSCOhost, https://research.ebsco.com/linkprocessor/plink?id=8b3b06df-18a9-32e4-b36c-fc39f618f754 (accessed Mar. 13, 2024).

[3] E. S. Omoora, H. A. Altaweil, T. Nagem, and K. A. Bozed, "Diabetes Mellitus Prediction Based on Machine Learning Techniques," in *2023 IEEE 11th International Conference on Systems and Control (ICSC)*, 18-20 Dec. 2023 2023, pp. 225-231, doi: 10.1109/ICSC58660.2023.10449831.

[4] M. Narasimharao, B. Swain, P. P. Nayak, and S. Bhuyan, "Developing and Evaluating a Machine Learning Based Diagnosis System for Diabetes Mellitus using Interpretable Techniques," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, 9-11 June 2023 2023, pp. 505-511, doi: 10.1109/APSIT58554.2023.10201753.

[5] G. R. Ashisha, S. T. George, X. A. Mary, K. M. Sagayam, and S. Pramanik, Analysis of diabetes disease using Machine Learning Techniques: A Review, Apr. 2022. doi:10.21203/rs.3.rs-1572946/v1

[6] P. Singh, S. Silakari, and S. Agrawal, “An efficient deep learning technique for diabetes classification and prediction based on Indian diabetes dataset,” 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Nov. 2023. doi:10.1109/ictacs59847.2023.10390518

[7] R. Moussaoui, O. Salem, and A. Mehaoua, "Analysis of Different Machine Learning Models for Diabetes Prediction," in *2023 IEEE International Conference on E-health Networking, Application & Services (Healthcom)*, 15-17 Dec. 2023 2023, pp. 43-48, doi: 10.1109/Healthcom56612.2023.10472394.

[8] A. Thammano, A. Meengen, “A New Evolutionary NeuralNetwork Classifier,” Springer-Verlag Berlin, pp. 249-255,(9), 2005