Predicting Diabetes Diagnoses

Maryellen Marino & Caleb Carr

Rensselaer Polytechnic Institute

COGS 4210: Cognitive Modeling

Dr. Stefan Radev

April 24, 2024

Introduction

Diabetes is a pervasive and escalating global health crisis. Approximately 38 million U.S. adults are living with diabetes, with a significant portion unaware of their condition (CDC, 2023). This chronic illness, which is the primary cause of adult blindness, kidney failure, and lower-limb amputations, poses a severe public health challenge (CDC, 2023). In light of these statistics, predictive modeling for early detection and management of diabetes is an essential avenue for research.

The National Health and Nutrition Examination Survey (NHANES) dataset offers a comprehensive collection of medical and lifestyle data that has been a cornerstone for various health-related research. Previous research has demonstrated the value of machine learning in utilizing NHANES for health predictions. For instance, Vangeepuram et al. (2021) explored the NHANES dataset to predict diabetes risk among the youth, illustrating the potential of these data in forecasting health outcomes across different demographics. Similarly, Qin et al. (2022) employed advanced machine learning models to predict diabetes based on lifestyle data, affirming the capability of analytical models to handle complex datasets like NHANES effectively.

This project leverages NHANES data to predict the presence of diabetes among individuals using a series of biomedical markers and lifestyle factors. This study's original dataset has been pared down to include age, gender, body mass index (BMI), glucose levels, and dietary habits. These parameters provide a multidimensional approach to understanding and predicting diabetes. The implementation employs Python for data manipulation and analysis, using libraries such as Pandas and Scikit-learn data structuring and machine learning model implementation. This project aims to create a robust model that can accurately predict diabetes

presence, thereby aiding in early diagnosis and potentially mitigating the severe complications associated with the disease.

The research question guiding this project is: How effectively can machine learning models, developed using the comprehensive NHANES dataset, predict the presence of diabetes in individuals based on their physiological data and lifestyle choices?

Methods

Six neural network models were developed using the Keras library to thoroughly analyze a cleaned version of the NHANES dataset. This dataset, initially containing around 40,000 observations, was reduced to 2,720 observations with 28 essential features. These models varied in complexity, incorporating differing numbers of layers and nodes to strategically address various aspects of the dataset for classifying individuals as diabetic or non-diabetic.

Software and Libraries Utilized

The project utilized a range of Python libraries, including TensorFlow for building neural network models, Pandas for data handling, Scikit-learn for data preprocessing and model evaluation, Matplotlib and Seaborn for visualization, and imbalanced-learn for applying SMOTE. These tools were selected for their robust functionality and compatibility with large-scale data analysis.

Data Preparation and Transformation

Data was initially pre-processed in R and saved as a CSV for further manipulation in Python. Pandas was used for data manipulation, enabling the refinement of the dataset to the essential variables for the study. For data preprocessing, numerical features were scaled to a mean of zero and a standard deviation of one using StandardScaler, and categorical variables were transformed into machine-readable formats through one-hot encoding using

OneHotEncoder. These transformations were integrated into a ColumnTransformer to streamline the application across the dataset.

Model Development and Training

A basic neural network model served as the initial template, structured with an input layer designed to accept the number of features determined post-transformation. The architecture involved multiple hidden layers with varying nodes and dropout rates to prevent overfitting, such as 128-node layers with 0.1 dropout and 64-node layers with 0.4 dropout. Each model utilized the 'adam' optimizer for learning optimization and 'binary_crossentropy' as the loss function, reflecting standard practices for binary classification tasks.

Model Evaluation and Enhancement

The dataset was split into training, testing, and validation sets to evaluate model performance comprehensively. A baseline model, alongside enhanced models incorporating bias regularization and adjusted class weights, were trained and validated to gauge their efficacy in handling class imbalances and enhancing predictive accuracy. Models were trained over multiple epochs, with performance evaluated using accuracy and loss metrics on unseen data.

The Synthetic Minority Over-sampling Technique (SMOTE) was applied to training datasets to correct class imbalances by synthesizing new minority class samples. This step was crucial in ensuring that models did not favor the majority class disproportionately. In addition to incorporating SMOTE, certain models incorporated L2 regularizations in the bias terms of each layer to prevent overfitting. Similarly, class weights were introduced to the loss function during training to improve the accuracy of the minority class.

Predictive Performance Analysis

Models' predictive capabilities were assessed using confusion matrices and classification reports, providing detailed insights into precision, recall, and F1-scores across diabetic and non-diabetic classifications. Loss trajectories were plotted for each model to identify potential overfitting or underfitting during the training phases.

Results

Baseline Model Evaluation

In our first model design, the accuracy was notably high, measuring at about 0.928 using solely the data in the original dataset. However, the validation and training cross-entropies did not converge. This indicates potential challenges in the baseline model generalizing to new data beyond the training set, thus suggesting the initial model is overfit (**Figure 1**).

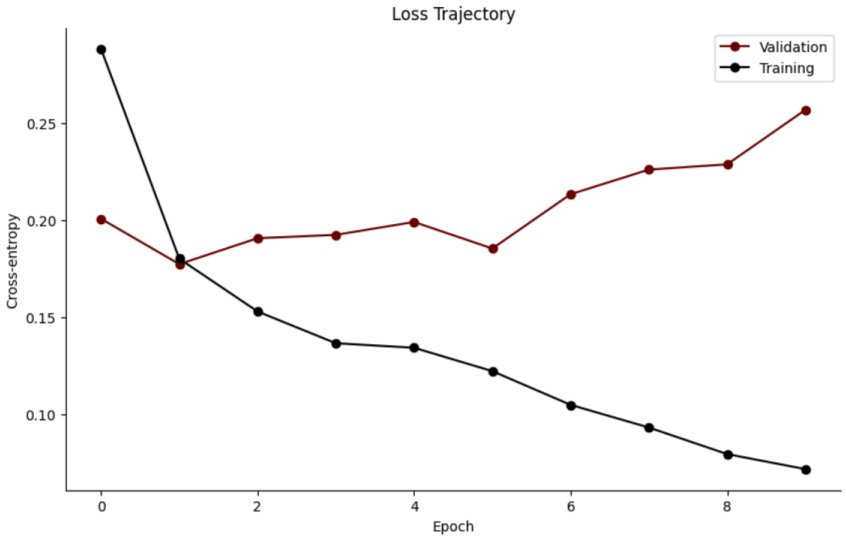

Figure 1. Loss trajectory of the training and validation processes on the baseline model. A consistent decrease in the training loss is observed from the start while the validation loss trends upwards.

This confusion matrix for the first model reveals that the accuracy for those with no diabetes is much higher than that for those with diabetes. A true negative (that is to say, the percent of non-diabetics predicted to not have diabetes) is predicted 97 percent of the time

compared to 67 percent for true positives (that is to say, the percent of diabetics predicted to have diabetes) (**Figure 2**).

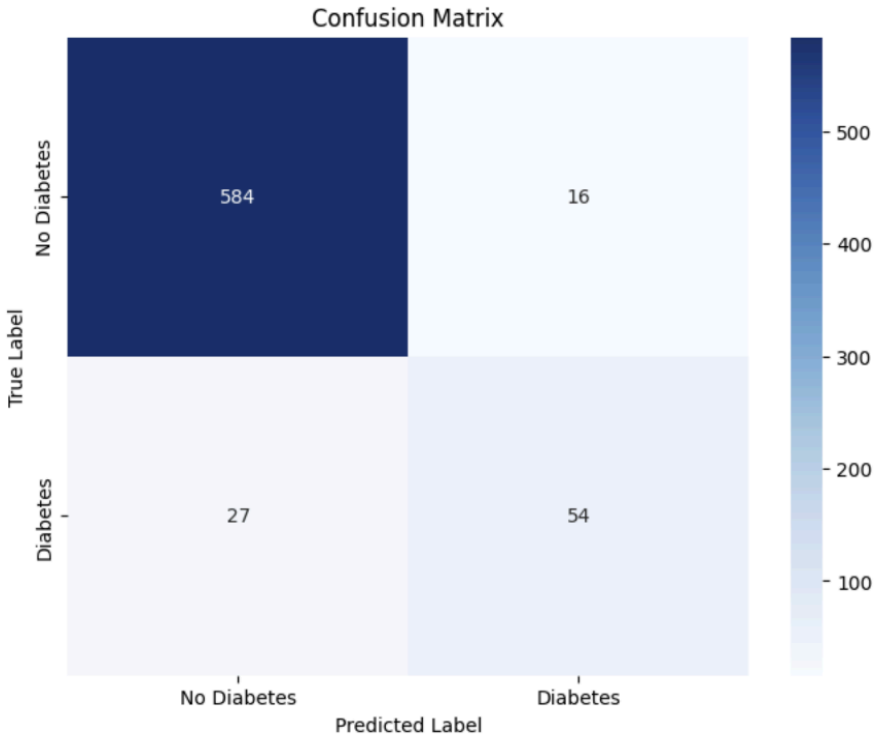

Figure 2. Confusion matrix of the baseline model. High predictive accuracy for classifying "No Diabetes" is observed, while classification in other areas appears less significant.

Model with Bias Regularization and Class Weights

After bias regularization and incorporating class weights into the model, the accuracy was 0.919. While the accuracy did not show a distinct change, the loss trajectory did converge. This significant improvement shows that the model is less prone to overfitting and, ideally, a better representative of the minority class (**Figure 3**).

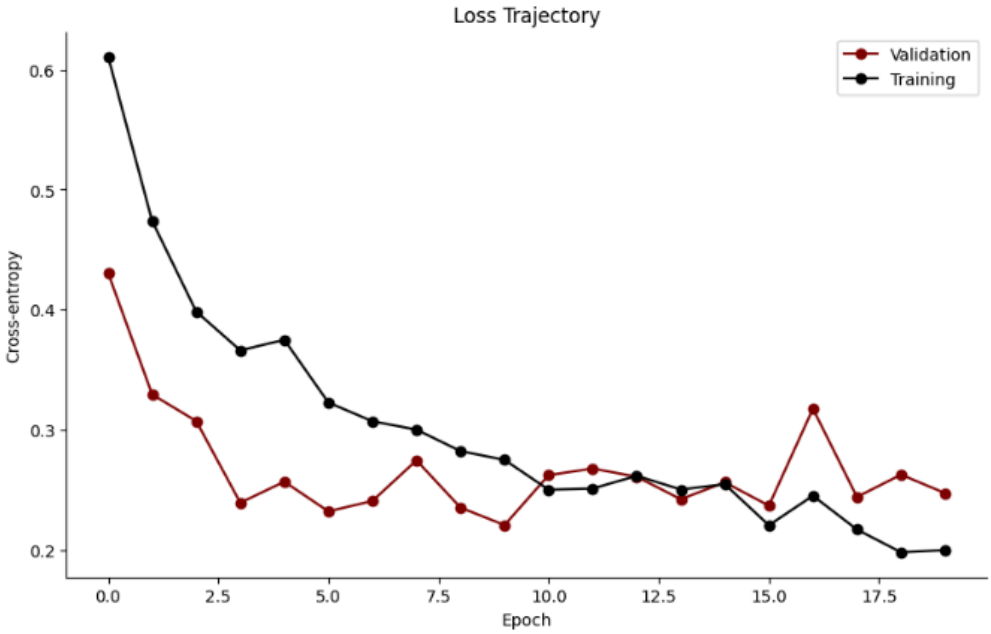

Figure 3. Loss trajectory of the training and validation processes on the model with bias regularization and class weights. A consistent decrease in both training and validation loss is observed with convergence.

This confusion matrix for the model incorporating bias regularization and class weights reveals that the accuracy for those with no diabetes is still much higher than that for those with diabetes. A true negative is found 93 percent of the time compared to 81 percent for true positives (**Figure 4**). This result has many advantages, considering the rate of false negatives has also decreased. False negatives should be strictly avoided when predicting diagnoses.

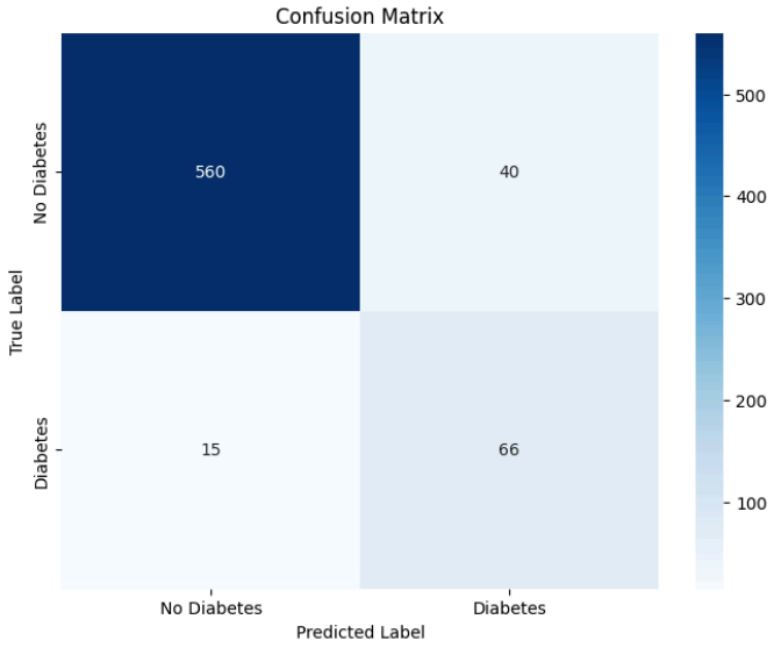

Figure 4. Confusion matrix of the model with bias regularization and class weights. High predictive accuracy for classifying "No Diabetes" is observed, while classification in other areas appears less significant.

Model with SMOTE

After incorporating SMOTE into the model to help balance the dataset, the accuracy was 0.922. The loss trajectory graph did not show significant improvement. With increased epochs, convergence occurred, but so did overfitting, so epochs were limited to 14 (**Figure 5**).

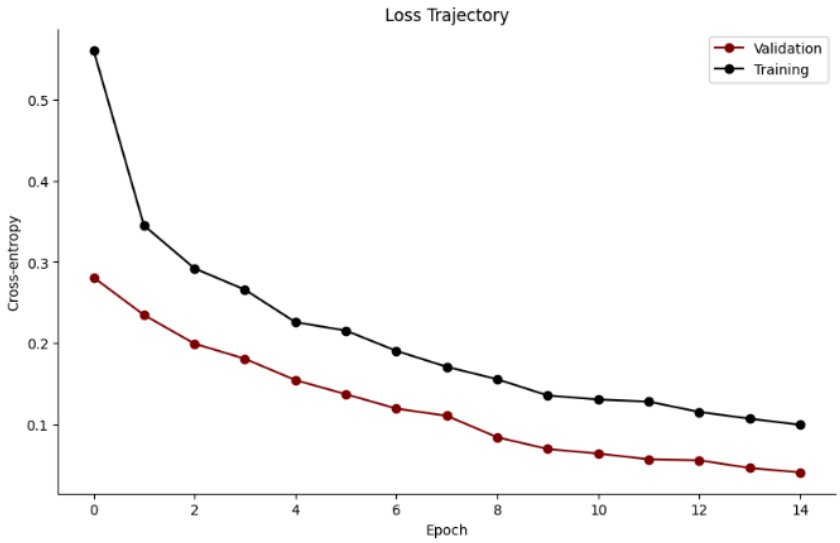

Figure 5. Loss trajectory of the training and validation processes on the model SMOTE. A consistent decrease in both training and validation loss is observed.

This confusion matrix for the model incorporating SMOTE reveals that the accuracy for those with no diabetes is still much higher than that for those with diabetes. Resulting in the same precision as the last model, a true negative is found 94 percent of the time compared to 76 percent for true positives (**Figure 6**). While overall accuracy is increased, it is not preferred to the non-smote model with weights and regularization as the goal is to maximize the rate of true negatives.

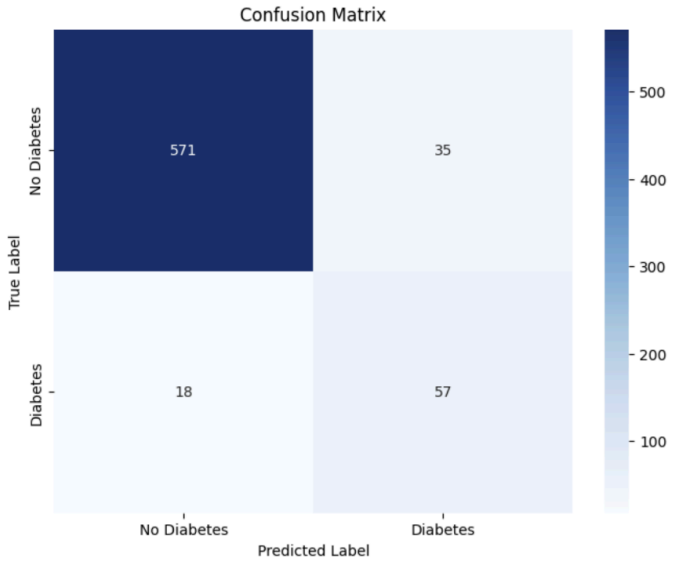

Figure 6. Confusion matrix of the model with SMOTE. High predictive accuracy for classifying "No Diabetes" is observed, while classification in other areas appears less significant.

Regularized SMOTE Model

After adding regularization to the SMOTE model, the accuracy was 0.912. However, the loss trajectory this time did show convergence. (**Figure 7**).

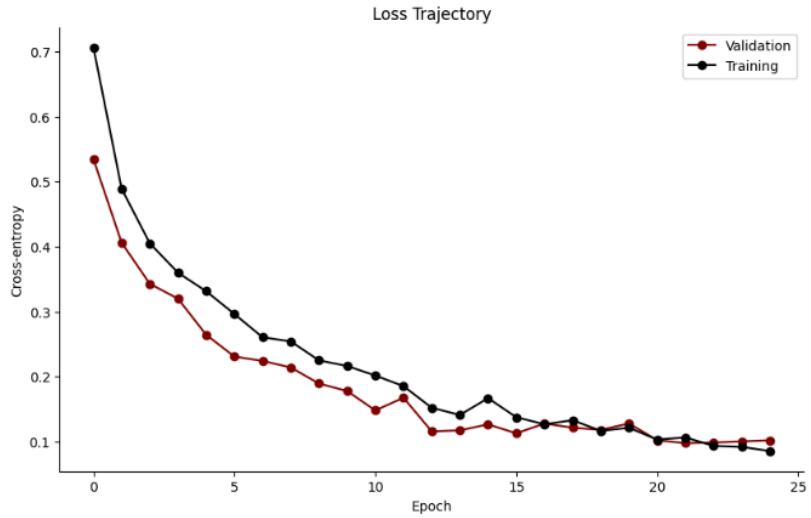

Figure 7. Loss trajectory of the training and validation processes on the model SMOTE with regularization. A consistent decrease in both training and validation loss is observed. Convergence occurs.

This confusion matrix for the model incorporating bias regularization and class weights reveals that the accuracy for those with no diabetes is still much higher than that for those with diabetes. The same precision as the last model results in a true negative being found 93 percent of the time compared to 79 percent for true positives (**Figure 8**).

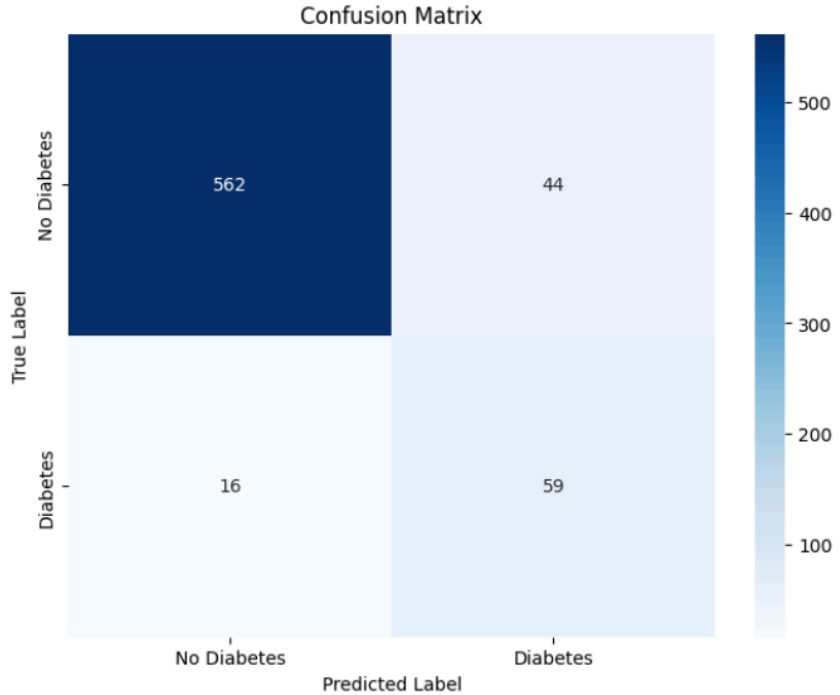

Figure 8. Confusion matrix of the model with SMOTE and regularization. High predictive accuracy for classifying "No Diabetes" is observed, while classification in other areas appears less significant.

Model with Enhanced Data Cleaning & Model Build

To further improve predictive accuracy, continued data processing was performed to trim potentially unnecessary features from the data. After trimming the data further and including regularization and class weights from earlier, an accuracy of 0.88 is observed. (**Figure 9**).

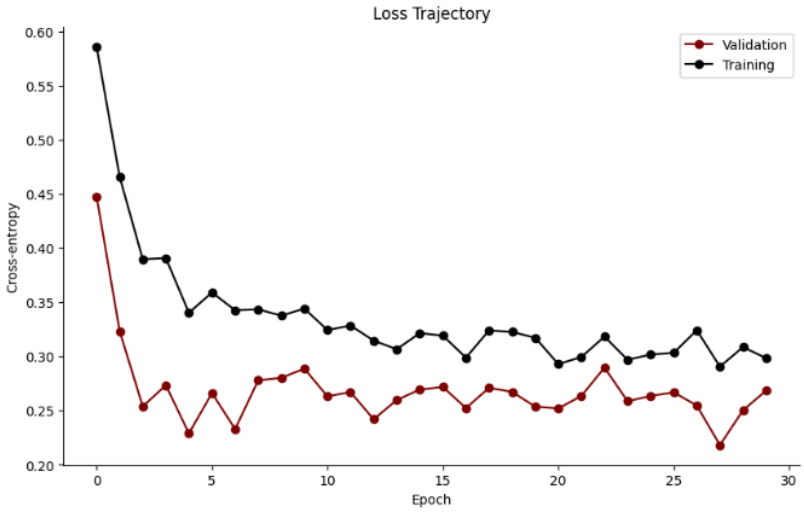

Figure 9. Loss trajectory of the training and validation processes on the model with data trimming and regularization. A consistent decrease in both training and validation loss is observed.

This confusion matrix for the model incorporating further data trimming, bias regularization, and class weights reveals that the accuracy for those with no diabetes is still much higher than that for those with diabetes. This results in slightly slower precision than the last model, where a true negative is found is a much lower 89 percent of the time compared to 81 percent for true positives (**Figure 10**).

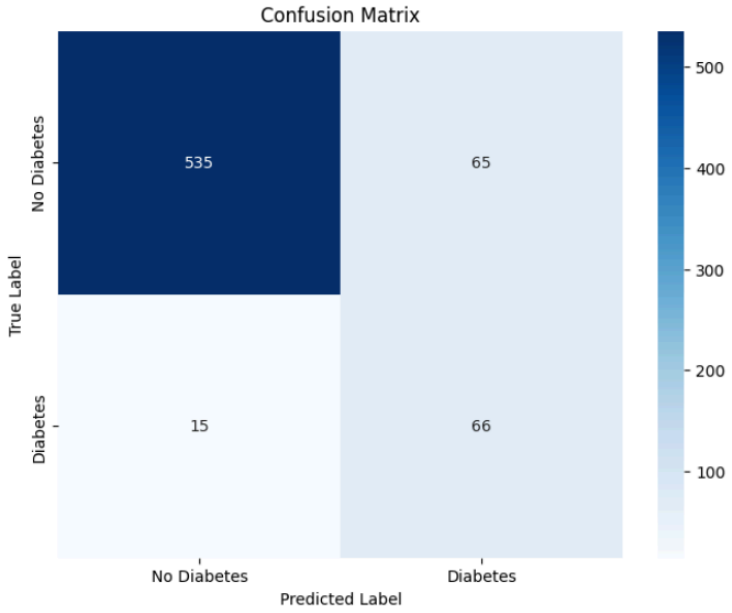

Figure 10. Confusion matrix of the model with data trimming and regularization. High predictive accuracy for classifying "No Diabetes" is observed, while classification in other areas appears less significant.

Model with Enhanced Data Cleaning & SMOTE

The same analysis was performed on the further cleaned data while incorporating SMOTE to determine if trimmed data on a more balanced dataset would be beneficial. The model with these new features resulted in an accuracy of 0.859 over 100 epochs (**Figure 11**).

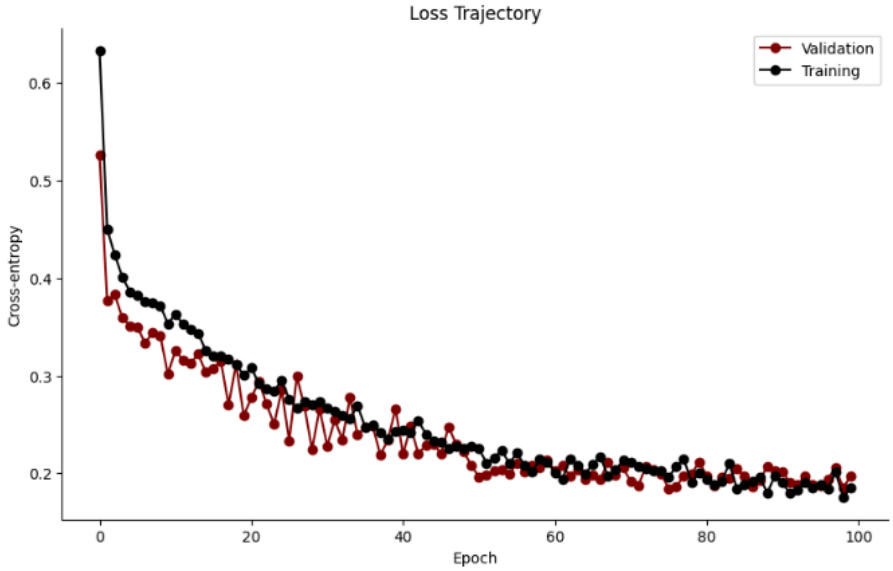

Figure 11. Loss trajectory of the training and validation processes on the model with data trimming and SMOTE. A consistent decrease in both training and validation loss is observed. Convergence occurs.

This confusion matrix for the model that incorporates further data trimming and SMOTE reveals that the accuracy for those with no diabetes is now lower than that for those with diabetes. While this is beneficial in approaching the goal of reducing false negatives, the model still needs much more work to be practically useful. The precision shows true negatives are found 86 percent of the time compared to 88 percent for true positives, the highest of all the models. (**Figure 12**).

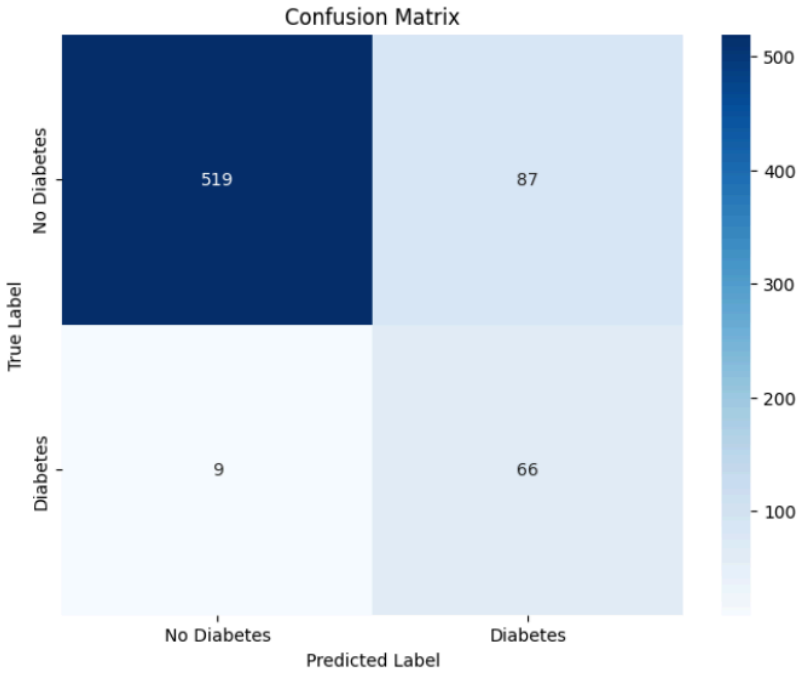

Figure 12. Confusion matrix of the model with data trimming and SMOTE. High predictive accuracy for classifying "No Diabetes" is observed, while classification in other areas appears less significant. Reduced false negatives.

Discussion

The analysis conducted on the NHANES dataset using various machine learning models has revealed several key insights and challenges associated with predicting the presence of diabetes based on physiological data and lifestyle choices. The models demonstrated a high level of predictive accuracy overall, particularly in correctly identifying individuals without diabetes.

This finding underscores the models' efficiency in detecting the absence of the condition, which is a crucial aspect of preventive health screening.

Key Learnings and Challenges

The analysis using the NHANES dataset presented significant insights, particularly the challenges associated with data imbalance in medical predictive modeling. The models demonstrated high overall accuracy in identifying non-diabetic cases. Yet, the prevalence of non-diabetic over diabetic cases in the dataset led to a performance bias toward predicting non-diabetic outcomes more effectively. This skew highlights the inherent difficulties in developing models that accurately predict conditions across unbalanced classes. Introducing the SMOTE and adjusting class weights were key strategies to enhance model sensitivity toward diabetic cases. SMOTE, in particular, provided a more balanced dataset by generating synthetic samples, which equipped the models to recognize better and predict diabetic outcomes marginally for the minority dataset. However, in practicality, this data balancing was less effective than expected.

Despite these strategic interventions, the reduction in the rate of false negatives was moderate, highlighting the limitations of current methods in effectively addressing class imbalance. The persistently high false negative rate—a critical issue in medical diagnostics where missing a diagnosis can have severe implications—emphasizes the need for more robust approaches. The high accuracy levels observed suggest that the models could achieve greater precision and reliability with a more balanced dataset.

Future Outlooks for Model Improvement

As previously mentioned, the largest challenge in improving the predictive models used in this study is the marked imbalance in the dataset, particularly the underrepresentation of

diabetic cases. This shortage of data for positive diabetes cases has resulted in less effective models capturing the complexities and variations associated with the disease, leading to a higher likelihood of false negatives.

Enhancing data acquisition efforts to address this limitation must be a primary focus. Specifically, gathering a larger volume of data from diabetic individuals is essential. More comprehensive data would give the models the necessary diversity and detail to learn the subtle distinctions in physiological and lifestyle factors that characterize diabetes. This dataset enrichment would help balance the classes and refine the models' ability to generalize and predict more accurately across varied populations.

While techniques like the Synthetic Minority Over-sampling Technique (SMOTE) have been implemented to mitigate the effects of data imbalance, they have limitations in replicating the genuine variation between non-diabetic and diabetic cases. Synthesized data, although useful, cannot fully capture the intricate patterns and anomalies present in naturally occurring data. Therefore, while SMOTE and similar methods are beneficial, they cannot substitute the need for actual data from diabetic cases.

Enhanced feature engineering also plays a vital role. By developing more sophisticated features that can capture the nuances of diabetes more effectively, models can be better equipped to detect the disease. Additionally, employing hybrid modeling techniques that combine different analytical approaches could help address data irregularities and improve the detection accuracy.

Another promising approach is to implement models capable of continuous learning. Such models can adapt and update their learning based on new data inputs, maintaining their effectiveness as additional data becomes available.

Project Resources

Team GitHub:

https://github.com/maryellenmarino/Cognitive_ModelingS24

Final Project Folder Code/Data:

https://github.com/maryellenmarino/Cognitive_ModelingS24/tree/main/finalProjecthttps://github.com/maryellenmarino/Cognitive_ModelingS24/tree/main/finalProject

NHANES Notebook:

https://github.com/maryellenmarino/Cognitive_ModelingS24/blob/main/finalProject/nhanes.ipynb

References

CDC. (2023). *What is diabetes?* Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html.

Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., Yu, J., Li, C., Yu, F., & Ren, Z. (2022). Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. International Journal of Environmental Research and Public Health, 19(22), 15027. https://doi.org/10.3390/ijerph192215027.

Vangeepuram, N., Liu, B., Chiu, P. H., Wang, L., & Pandey, G. (2021). Predicting youth diabetes risk using NHANES data and machine learning. Scientific Reports, 11(1), 11212. https://doi.org/10.1038/s41598-021-90406-0.