Final Report

Predicting Type 2 Diabetes Risk Using Multidomain NHANES Data:

Integrating Demographic, Clinical, Behavioral, Social and Environmental Factors

Scott Enriquez (Leader), Somaya Albhaisi, Darren Parry

Problem Definition and Background

Type 2 diabetes (T2D) represents one of the most pressing global health challenges, contributing significantly to rising rates of morbidity, mortality, and healthcare expenditure. Its onset is frequently gradual and asymptomatic, leading many individuals to remain undiagnosed until serious complications develop. **Early detection and accurate risk assessment** are therefore essential for timely intervention, prevention of disease progression, and reduction of long-term public health and economic burden. This project addresses the need for improved screening tools by developing machine learning models to identify individual adults likely to have diabetes or prediabetes, using the comprehensive, nationally representative data from the National Health and Nutrition Examination Survey (NHANES).

We frame this as a **binary classification** task, predicting whether a participant has been or would be told by a health professional that they have diabetes or prediabetes, as captured by the target variable **DIQ010**. While this variable reflects existing diabetes rather than future diagnosis, it serves as a clinically relevant proxy for identifying high-risk individuals who may benefit from confirmatory testing and preventive care. The problem is particularly **meaningful** given that an estimated 20–30% of diabetes cases in the United States remain undiagnosed. Models trained on multidomain NHANES data spanning demographic, clinical, behavioral, social, and environmental factors can support targeted screening efforts and enable more efficient allocation of public health resources. This work **builds upon prior research** that has utilized NHANES data for diabetes risk prediction, such as studies focusing on youth diabetes risk (Vangeepuram et al., 2021) or easy-to-measure risk factors (Turi et al., 2017). However, many existing approaches have relied on limited feature sets or traditional statistical methods. Our project extends these efforts by integrating a broader array of features from across NHANES datasets and applying advanced, automated machine learning techniques to optimize model performance with a specific emphasis on recall, a critical metric for minimizing false negatives in a screening context. We also address practical challenges inherent to the NHANES database, including high dimensionality, class imbalance (approximately 10:1 no diabetes diagnosis to diabetes diagnosis ratio), missing data, and inconsistencies across survey cycles.

The key **contributions** of this project are threefold. First, we develop and implement a reproducible data preprocessing and feature engineering process that merges, normalizes, and selects relevant variables from over 1,000 NHANES features. Second, through iterative experimentation using AutoGluon, we produce a performant gradient boosting model that achieves high recall with a parsimonious set of 20 clinically interpretable features, balancing predictive power with practical usability. Finally, our recall-driven modeling approach and focus on feature applicability provide a scalable framework that can inform real-world clinical decision support and public health screening strategies, bridging the gap between epidemiological data and actionable risk identification.

Description of Dataset

The NHANES dataset, renowned for its rigorous collection of health, nutritional, and environmental data across diverse demographics, provides a foundation for this detailed investigation into T2D prediction. The data is anonymized but uniquely identifiable, using a sequence number (SEQN). There are over 100 tables available with thousands of features total. Tables in NHANES can be joined using the SEQN column to gain further insight into the anonymous individual. The majority of the features that we researched are categorical, with clear explanations of how the data was collected, typically involving a specific question asked of the participants and the meaning of each value. However, when the question wording changes even slightly, the previous variable is deprecated in favor of a new column. As a result, many columns needed to be merged together to address the sparsity of similar columns, since these differences are often minor and do not alter their semantic meaning. The diabetes table was our primary focus, containing a **quantity** of ~99,000 participants and a target called DIQ010. Our target variable indicates whether a participant has been told by a health professional that they have diabetes or prediabetes (i.e., diagnosis). The diabetes dataset is imbalanced and has a 10:1 negative to positive class ratio. During the exploratory data analysis (EDA), the team analyzed over 1,000 features using a combination of statistical measures and domain expertise. NHANES provides statistical analysis and distribution of values (e.g., continuous variable distributions and valid ranges, categorical variable meanings, etc.), so much of the EDA was centered around how to prepare the data for machine learning. We targeted tables that have a proven association with diabetes and researched features within those tables to build the classification models.

The datasets often had data **quality** issues that we addressed during the EDA and preprocessing. Many variables contain a mixture of both continuous and categorical values within the same column. For example, a column called ALQ130 captures the average number of drinks per day the participant has had over the past 12 months, but the value of 77 indicates that the participant refused to provide an answer. In such cases, interpreting this value as purely continuous would skew the results. Normalization was another aspect to consider for the dataset. Some columns indicated the units of measurement for other variables.

Given that most of the features that the team targeted in NHANES are categorical, we binned the continuous variables using a combination of domain knowledge and statistical techniques (e.g., quartiles) in order to support standardized metrics across the features. The team calculated coverage (i.e., the percentage of rows that contain data for a given feature), chi-square for relation, and Cramér's V for strength of association. The team found that many features had a strong association with limited coverage.

By the end of the project, we selected 20 features for our final model of the exhaustive list of 1,003 features (Figure 1). These selected features are categorized into four distinct domains. The first one includes physical activity levels (PAQ), alcohol consumption (ALQ), smoking status, and blood pressure and cholesterol questionnaire data (BPQ). Next, we explicitly highlighted social determinants of health, capturing socioeconomic data through features such as family poverty level (INDFMMP), financial stability, income sources (INQ30), health insurance coverage (HIQ), home ownership (HOQ), and education level (DMDEDUC2). The rest of the features include standard demographic questionnaire data, like age (RIDAGEYR), race/ethnicity (RIDRETH1), and marital status (DMDMARTL), and examination data comprised of body mass index (BMXBMI), waist circumference (BMXWAIST), and systolic blood pressure (BPXSY1).

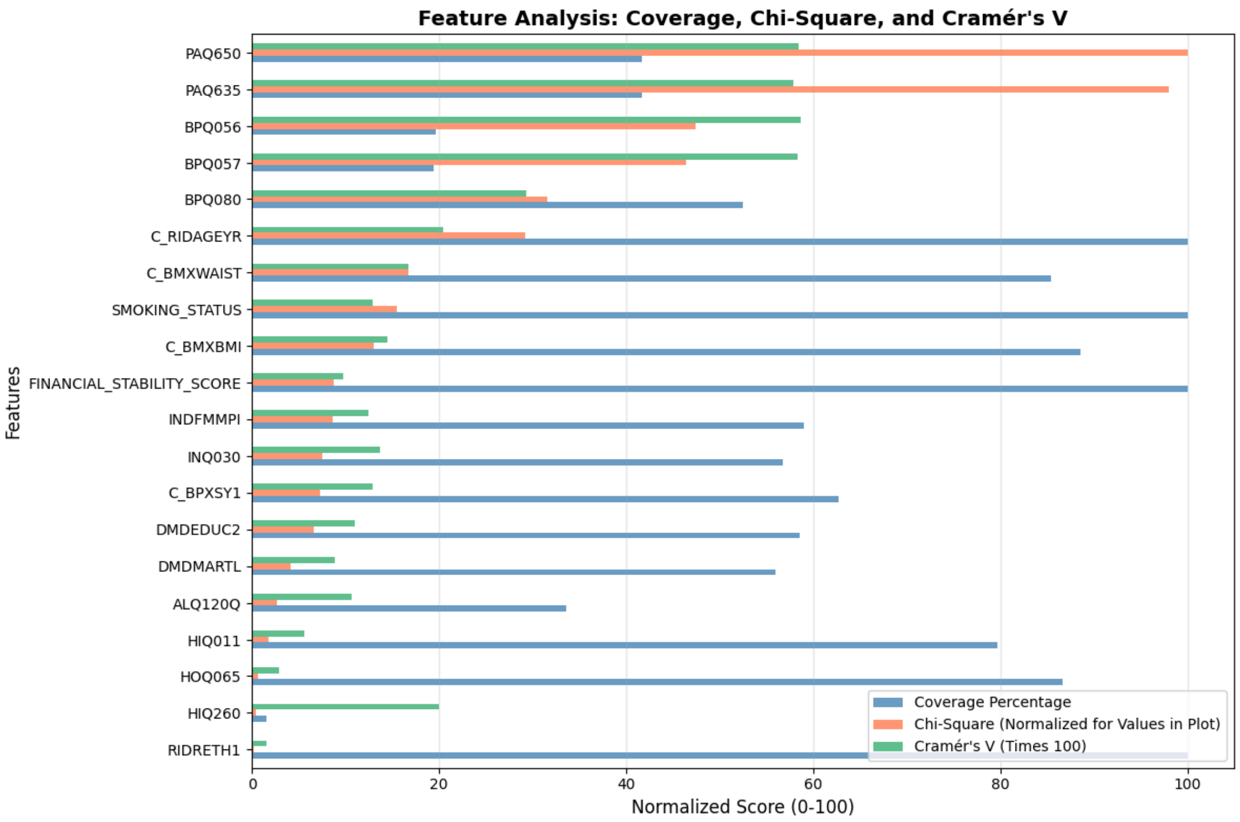

Figure 1: Final 20 expert-selected features out of 1,003 features (full list available in deliverable)

Methods Used, Experiment Setup, and Analysis Results

In order to **rapidly create and evaluate as many models as possible**, the team used AutoGluon as our primary model development tool. AutoGluon is an open-source library originally developed by Amazon Web Services (AWS) that trains and combines multiple state-of-the-art models (i.e., ensembles), performs hyperparameter tuning, and generates leaderboards for model evaluation. The ability to generate thousands of models in a completely automated fashion enabled the team to iterate quickly on new ideas. All code for developing the models is located in the *models* folder of our project ZIP file. The team used the train/test split functionality of scikit-learn to ensure that models were evaluated on unseen data to **minimize overfitting**. Given the unbalanced dataset, we used an 80/20 split for training and evaluation with stratification for the target variable to ensure that both sets have roughly the same distribution of positive and negative classes. Given the **high negative impact of false negatives and the low negative impact of false positives**, we trained most models targeting F1 as the primary metric and ranked the results based on recall. While it is possible to train the models targeting recall as the primary metric, the team found that AutoGluon overoptimized and produced models with a perfect recall of 1 and nearly zero accuracy in such cases. The balance of F1 allowed us to produce models with a reasonable baseline accuracy of at least 0.70 and high recall at the cost of precision. For imputation, we found that using a random forest model regression model yielded better results than simple statistical methods like mean, median, or mode. Imputation was used primarily to fill in physical activity features (i.e., PAQ*) with high chi-square and Cramér's V but low coverage. Over nine systematic iterations, our project converged on a high-performance, practical model for diabetes risk screening (Table 1).

Iteration	**Algorithm**	**Features**	**Recall**	**Accuracy**	**Precision**	**F1**	**Description**
1	NeuralNet Torch	1,003	0.61133	0.89668	0.44221	0.51320	Initial baseline
2	CatBoost	1,003	0.86912	0.80861	0.30108	0.44723	Sample weights
3	LightGBM	150	0.84986	0.78332	0.27135	0.41135	Top features by Cramér's V
4	CatBoost	150	0.87365	0.80684	0.29965	0.44625	Top features by chi-square
5	LightGBM	150	0.27365	0.91995	0.61372	0.37853	Targeting accuracy
6	CatBoost	150	0.87195	0.80735	0.3	0.44641	Adding imputation to Iteration 4
7	CatBoost	20	0.87649	0.77045	0.26323	0.40487	Domain expert features
8 (Best)	CatBoost	20	0.89122	0.74300	0.24301	0.38189	Adding imputation to Iteration 7
9	NeuralNet FastAI	20	0.43003	0.89224	0.40201	0.41555	Oversampling without weights

Table 1: Best model and performance summary for each iteration during the project

The most important and meaningful results are as follows:

1. Class imbalance correction was the single biggest lever for performance. The baseline model (Iteration 1) using all features achieved a low recall of 0.611, failing to identify nearly 40% of diabetes/prediabetes cases. By simply applying **sample weights** in AutoGluon to counter the 10:1 negative-to-positive class ratio (Iteration 2), recall skyrocketed to **0.869**, an improvement of over 0.25. This was our most impactful finding, underscoring that without explicitly prioritizing the minority class, even advanced algorithms will default to predicting the majority, rendering a screening tool useless.

2. A parsimonious model can match or exceed the performance of a complex one. A key objective was balancing predictive power with practical usability. Iterations using 150 top-ranked features (by Cramér's V or chi-square) maintained recall around 0.87. However, the most significant practical result came from Iterations 7 and 8. By refining the feature set to just **20 clinically interpretable variables** selected through domain expertise and statistical association, we achieved our **highest recall of 0.891** (Iteration 8). This demonstrates that a small, meaningful subset of inputs (demographic, clinical, socio-behavioral factors, and relevant lab values) can provide excellent sensitivity without the burden of collecting hundreds of data points.

3. Metric selection directly defines clinical utility. Iteration 5 explicitly targeted accuracy, resulting in the highest accuracy (0.920) but the worst recall (0.274). This experiment perfectly illustrates the critical trade-off: a model optimized for overall correctness will overwhelmingly predict "no diabetes," missing most actual cases. For a screening tool where the cost of a false negative (a missed diagnosis) is high, **recall is the paramount metric.** Our final model (Iteration 8) accepts lower accuracy (0.743) and precision (0.243) to achieve the high recall necessary for effective early detection.

4. Feature engineering strategies had nuanced effects. We found that: (i) **Imputation** (Iterations 6 & 8) on key features with high importance but low coverage provided a small but consistent boost to recall, confirming that intelligently handling missing data preserves signal. (ii) **Oversampling (RandomOverSampler)** on the reduced feature set (Iteration 9) was detrimental, causing recall to collapse to 0.430. For our specific data and models, reweighting the loss function was a superior strategy for handling imbalance.

In summary, the most meaningful outcome is our **final CatBoost model using 20 features**. It embodies the project's goal: a highly sensitive (Recall = 0.891), practical screening tool that identifies nearly 9 out of 10 individuals with diabetes or prediabetes by asking a manageable set of questions. To make the model accessible for consumption and further evaluation, the team deployed the final CatBoost model to AWS Lambda, Amazon's serverless offering, to perform ad-hoc inference. Using the 20 final features as inputs, the API can be invoked via an HTTP POST call. Instructions and an example invocation are included in the code deliverable.

Observation and Conclusion

While we proved that it is possible to distill the essence of diabetes risk into a simple checklist without losing predictive power, the NHANES data has several limitations. The target variable (DIQ010) reflects self-reported diagnosis, which may misclassify undiagnosed diabetes. NHANES data also contain substantial missingness and inconsistent variable coding, requiring imputation and variable merging that may introduce bias. Because NHANES is cross-sectional, the model identifies associations rather than causality and may not generalize fully to clinical settings or populations unlike NHANES. Class imbalance and dataset-specific feature distributions further limit generalizability, and the feature selection process, while rigorous, relied partly on domain expertise and coverage constraints, which may have excluded relevant variables. We mitigated these issues through a reproducible preprocessing automation, careful harmonization of related variables, and targeted model-based imputation to reduce bias. Emphasizing recall as the primary metric helped compensate for self-report limitations by maximizing sensitivity to high-risk individuals. Class imbalance was addressed using sample weighting, which produced more stable models than oversampling. Combining statistical association with clinical domain expertise allowed us to refine the feature set to 20 interpretable variables, resulting in a high-recall, practical model. These steps collectively strengthen the robustness, interpretability, and real-world relevance of our final screening tool.

A critical observation is the inherent limitation of our target data. Our model predicts the variable DIQ010: "Has a health professional *told* you that you have diabetes or prediabetes?" Therefore, a prediction of "no diabetes" does not mean the person is disease-free; it means they are unlikely to have a *known* diagnosis. This accurately mirrors the real-world public health crisis of undiagnosed diabetes. Our model's high recall means it is excellent at identifying the profile of someone who *should* be diagnosed, making it a potent tool for finding those who have slipped through the cracks of the healthcare system. For practical implementation, this tool could be deployed as a rapid digital risk assessment. In a clinic waiting room or a community health fair, individuals could answer a brief 20-item questionnaire. Those flagged as high-risk could then be formally referred for a definitive HbA1c or fasting glucose test. This creates an efficient, two-stage screening process that maximizes resource allocation, a core goal of public health. For our next steps, we plan on publishing the findings from our research and refining our decision tool so that it can be tested for pragmatic use by clinicians to develop more reliable and efficient ways to identify diabetes risk.

In conclusion, this project was a deep exploration into turning a large, complex public health dataset into a focused, actionable tool. Our understanding is that the technical challenge was not just about building an accurate model, but about building a useful one. The real success lies in the journey from 1,003 features to 20 features, proving that with careful analysis, we can distill the essence of diabetes risk into a simple checklist without losing predictive power. We learned that addressing class imbalance is non-negotiable, that simpler models are often more deployable, and that every metric choice carries a real-world consequence. The final model is not a diagnostic crystal ball, but a smart, sensitive sieve designed to improve the efficiency of early detection. By focusing on recall and simplicity, we have created a prototype for a tool that can genuinely contribute to reducing the burden of undiagnosed type 2 diabetes in adults.

References:

Vangeepuram, N., Liu, B., Chiu, P., 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

Turi, K. N., Buchner, D. M., & Grigsby-Toussaint, D. S. (2017). Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors. *Preventing Chronic Disease*, *14*, E23. https://doi.org/10.5888/pcd14.160244