Final Report: Predicting Diabetes in the Pima Indian Population

1. Literature Review & Public Health Question

The Public Health Context

Type 2 diabetes is a chronic disease of serious type where early diagnosis is essential in averting the chronic complications that may include kidney failure and blindness. One of the many problems with normal screening is the fact that it usually does not detect insulin resistance at an early stage. Here, the worst case is a False Negative (FN) which is informing a sick patient that he is healthy since they are deprived of the treatment that they need.

The Pima Indian Anomaly

The study aims at the Pima Indians of Arizona, one of the demographic groups in the world with the highest prevalence of Type 2 diabetes documented. Fifty percent of the adults more than 35 years old have the disease in this population.

Statistical Question

Is it possible to use machine learning to correctly forecast diabetes among high-risk groups using standard biometric data, and under specified conditions to reduce the number of missed cases (False Negatives)?

Limitations of Prior Work

Previous studies utilizing the NIDDK dataset have faced two primary hurdles:

1. **Performance Ceilings:** Most models plateau at approximately 75% accuracy due to the limited sample size (768 records).
2. **Bias:** Models consistently favor the majority class (Healthy), resulting in dangerously low Recall for diabetic patients.

2. Data Sources and Variables

Data Origin

The dataset was sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

Response and Variables of Interest

- **Response Variable:** `Outcome` (Binary), defined as 0 (Healthy) or 1 (Positive diagnosis for Diabetes).
- **Predictors:** Standard clinical markers for metabolic syndrome, including Glucose, `BMI (Body Mass Index),` Insulin, Blood Pressure, and Skin Thickness.

Confounding Variables

We identified Age, Pregnancies, and DiabetesPedigreeFunction as potential confounders.

- *Rationale:* Age and pregnancy history significantly alter physiological baselines. These variables were included as features to allow the algorithm to adjust for these variances rather than letting them skew the results.

Data Issues: The "Zero" Problem

A critical data quality issue was the presence of biologically impossible zero values. Living patients cannot have a Blood Pressure, BMI, or Glucose level of 0. These zeros act as noise that confuses the model.

- **Severity of Missingness:** Significant missing data was observed in Insulin`(approx 50%) and SkinThickness.`
- **Handling Strategy:** We employed **Median Imputation**, replacing 0s with the median value of non-zero entries for each column. This approach preserved the data distribution better than mean imputation or dropping rows.

3. Data Analysis and Results

3.1 Exploratory Analysis

Initial exploration of the dataset revealed a class imbalance: 65% of the dataset was labeled "Healthy" (0) and 35% "Diabetic" (1). This imbalance typically encourages models to be "lazy" by predicting "Healthy" for everyone to maximize accuracy at the expense of safety.

3.2 Analytic Methods

A combination of Accuracy and Safety was achieved by testing three different algorithms:

- Logistic Regression: Interpretability is done on the basis of this. It has not been effective in capturing the complex non-linear relationships and therefore the missed cases were too many (27 Missed Cases).

- Random Forest: Chosen a Safety Net. It also had great Recall but less overall accuracy because of higher numbers of false alarms (14 Missed Cases).

- XGBoost + Feature Engineering (Final Model): The chosen model is the XGBoost + Feature Engineering. The most stable and safe was a combination of gradient boosting and engineered features (11 Missed Cases).

Feature Balancing and Engineering.

- In order to overcome the 75% accuracy ceiling and overcome bias, we applied:

- SMOTE (Synthetic Minority Over-sampling Technique): We created fake images of diabetic patients within the training group in order to compel the model to acquire the traits of diabetes.

- Engineered Features:
    - *Insulin Resistance Index* (Glucose / Insulin): Captures the body's struggle to utilize insulin.
    - *High Risk Combo:* A binary flag for patients with BMI > 30 AND Age > 40.

3.3 Analytical Results

The final model (XGBoost + Engineered Features) yielded the following performance metrics on the test set:

Metric	**Baseline (Logistic Regression)**	**Final Model (XGBoost + FE)**	**Improvement**
Accuracy	69.5%	**74.7%**	+5.2%
Recall	50.0%	**79.6%**	+29.6%
Missed Cases (FN)	27	**11**	-59%

Discussion of Results

The results demonstrate a successful shift in priority from pure accuracy to clinical safety.

- **Safety Improvement:** We reduced the critical error rate (False Negatives) by nearly 60%.
- **Consistency with Literature:** The overall accuracy of 74.68% is consistent with the "Performance Ceiling" noted in literature, where most models using this dataset cap at ~75%. This suggests the dataset lacks specific genetic markers or diet logs needed for higher precision.

3.4 Improvement Strategy

While Recall is strong (~80%), the F1-score (0.688) indicates a trade-off with Precision (more False Positives). To improve the model:

- **External Validation:** Test on non-Pima datasets to ensure the model generalizes well.
- **Deep Learning:** If the sample size could be increased significantly beyond 768 records, deep learning approaches might capture more subtle patterns.

4. Conclusion

This project successfully addressed the public health question by demonstrating that machine learning can be optimized for safety in high-risk populations. By moving beyond simple accuracy and incorporating **Feature Engineering** and **Class Balancing**, we developed a clinically viable model.

The final model identifies significantly more at-risk patients than standard baselines, directly translating to earlier intervention and better long-term health outcomes for the Pima community. Specifically, for every 100 diabetic patients screened, our model identifies 30 more patients than the baseline model.