Executive Summary –

GlycoTrack: Predicting Diabetes Risk

Client Context

We are engaged with a Medicine Corporation client to analyze behavioral and health survey data for predicting diabetes risk across population segments. The objective is to deliver actionable insights that support preventive health strategies and policy interventions.

Project Objective: -

This project analyzes CDC's BRFSS data to identify key predictors of diabetes among U.S. adults. It integrates medical history, lifestyle habits, functional limitations, and demographic indicators to build predictive models and inform public health decisions.
Problem: Early identification of individuals at high risk of diabetes to enable preventive healthcare measures.
Business Context: Diabetes is a major global health challenge. Accurate prediction can help public health systems and insurers reduce costs by targeting at-risk populations for lifestyle interventions and screenings.

1. Data Source and Dataset Overview-

Data Source-

The dataset originates from the Behavioral Risk Factor Surveillance System (BRFSS), an annual health-related survey conducted by the Centers for Disease Control and Prevention (CDC) and U.S. state health departments.

About BRFSS:

- **Purpose:** Monitors health-related risk behaviors, chronic health conditions, and use of preventive services.
- **Method:** Random-digit–dialed telephone interviews (landline & cell phone).
- **Coverage:** U.S. adults aged 18+, ~400,000 interviews annually.
- **Reliability:** Standardized questions, weighted to represent the U.S. population.

This dataset focuses on health, lifestyle, functional, and demographic indicators relevant to diabetes risk assessment.

Dataset Variables and Key Groups

1. Health Status & Chronic Conditions

- **Diabetes_binary – Target variable** (0 = No diabetes, 1 = Diabetes)

- HighBP – High blood pressure
- HighChol – High cholesterol
- CholCheck – Cholesterol check in last 5 years
- Stroke – History of stroke
- HeartDiseaseorAttack – Coronary heart disease or heart attack

2. Lifestyle & Behavior

- Smoker – Ever smoked ≥ 100 cigarettes
- PhysActivity – Physical activity in past month
- Fruits – Consumes fruit 1+ times/day
- Veggies – Consumes vegetables 1+ times/day
- HvyAlcoholConsump – Heavy alcohol consumption
- AnyHealthcare – Has healthcare coverage
- NoDocbcCost – Could not see a doctor due to cost

3. Functional Limitations

- DiffWalk – Difficulty walking or climbing stairs

4. Numerical Measures

- BMI – Body Mass Index
- MentHlth – Days mental health not good (past 30 days)
- PhysHlth – Days physical health not good (past 30 days)
- GenHlth – General health rating (1 = Excellent → 5 = Poor)

5. Demographics

- Sex – 0 = Female, 1 = Male
- Age – Age group (1 = 18–24, 13 = 80+)
- Education – Education level (1 = No school → 6 = College 4+ years)
- Income – Income category (1 = < 75,000)

Why This Dataset is Valuable:

- Large, representative sample from a trusted public health source.
- Contains both medical and lifestyle variables, enabling predictive modelling.

- Features are mostly binary or numeric, simplifying preprocessing and model development.

2. Data Preparation

The raw dataset contained over 72,000 survey responses with 23 features. Cleaning steps included:

- Size: ~253,680 rows before cleaning
- Handling missing values and duplicates
- Outlier detection and treatment especially in categorical variables (education, income, lifestyle).
- Converted categorical variables to numerical encodings
- Scaled numerical features (BMI, age, physical activity) for comparability.

3. Exploratory Data Analysis (EDA)

EDA revealed important demographic and lifestyle patterns. Distributions of BMI, Age, and lifestyle indicators were examined. Relationships between high blood pressure, cholesterol, and diabetes status were also analyzed.

- Target distribution: ~14% diabetic, ~86% non-diabetic (class imbalance)
- BMI strongly associated with diabetes
- Physical inactivity increases risk
- Older age groups have higher prevalence
- Poor self-reported health correlates with higher prevalence
- Moderate association with high blood pressure & cholesterol
- **Class Imbalance**: Majority non-diabetic; SMOTE applied
- **Correlations:**
  - Positive: HighBP, HighChol, DiffWalk, poor GenHlth
  - Negative: PhysActivity, Fruits, Veggies
- **Distributions:** Skewed BMI, MentHlth, PhysHlth

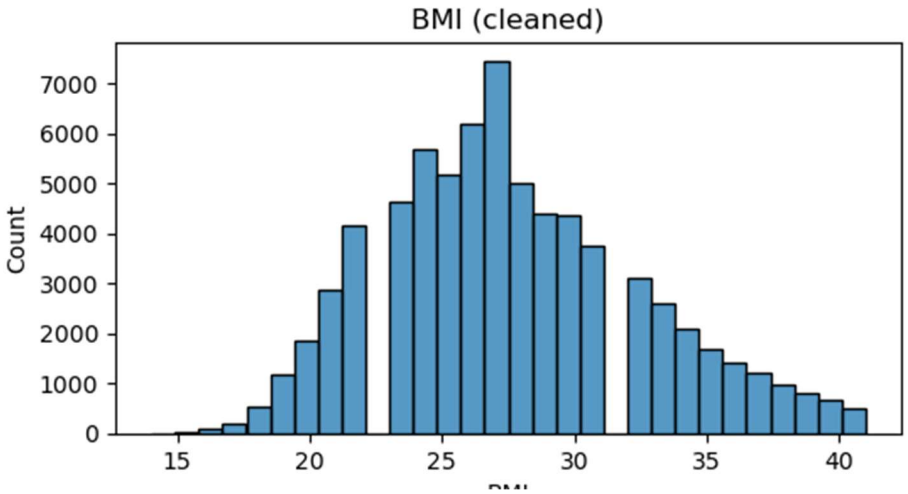

Figure 1: Distribution of BMI across population segments

4. Feature Engineering

Features engineered included:
- Created health behavior scores and binary flags (e.g., smoker, hypertensive).

- Binned risk categories for age and BMI.

-Engineered interaction features combining lifestyle and demographics. (e.g., physical activity × BMI)

- Applied mutual information and correlation to retain impactful features.

Risk Classification Strategy:

- Compared Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and XGBoost.
- Tuned hyperparameters using GridSearchCV.
- Selected XGBoost for its balance of accuracy and interpretability.

5. Modeling and Validation

Three models were implemented: Logistic Regression, Random Forest, and XGBoost. Models were validated with stratified k-fold cross-validation, using metrics such as ROC-AUC, Precision, Recall, and F1-score.

Modeling Approach

- Baseline models: Logistic Regression, Decision Tree, Random Forest, KNN, SVM, XGBoost
- Metrics: Accuracy, Precision, Recall, F1, ROC-AUC
- Class imbalance addressed with SMOTE oversampling
- Best model: XGBoost (after SMOTE)
- Used stratified k-fold cross-validation.
- Evaluated models using AUC-ROC, precision, recall, F1-score.
- Presented confusion matrices and SHAP plots for interpretability.

Results & Performance

- Baseline (imbalanced): XGBoost — Accuracy 0.74, ROC-AUC 0.83, Recall 0.62 (many diabetics missed)
- After SMOTE (balanced): XGBoost — Accuracy ~0.87, Precision ~0.82, Recall ~0.85, F1 ~0.83, CV ROC-AUC ~0.97 — Recall improved significantly, critical for healthcare screening

Model	CV ROC-AUC	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC-AUC	Model selection
Logistic Regression	0.822209	0.870363	0.528233	0.151199	0.2351	0.823029	Best Baseline models- Logistic Regression per accuracy score and CV ROC-AUC
Decision Tree	0.590735	0.798228	0.26977	0.311262	0.28903	0.590528	-
Random Forest	0.798907	0.866172	0.475	0.148592	0.22637	0.804638	2nd good Baseline models per Accuracy and 2nd per Hyperparameter tuning - 0.870423
Naive Bayes	0.781435	0.768893	0.312403	0.627737	0.41719	0.789392	-
XGBoost	0.818952	0.735504	0.303419	0.777372	0.43648	0.82529	Best model per Hyperparameter tuning- 0.871247. Best Baseline models per ROC-AUC- 0.82529
Logistic Regression (After SMOTE)	0.836073	0.734199	0.299897	0.762252	0.43044	0.820478	
Random Forest (After SMOTE)	0.978349	0.854287	0.407136	0.232013	0.29558	0.802325	After SMOTE, best model for Hyperparameter tuning. And 2nd good baseline model
XGBoost (After SMOTE)	0.971926	0.866653	0.483916	0.180396	0.26282	0.820827	After SMOTE, Best baseline model and 2nd good model per Hyperparameter tuning

XGBoost achieved the best ROC-AUC (~0.82).

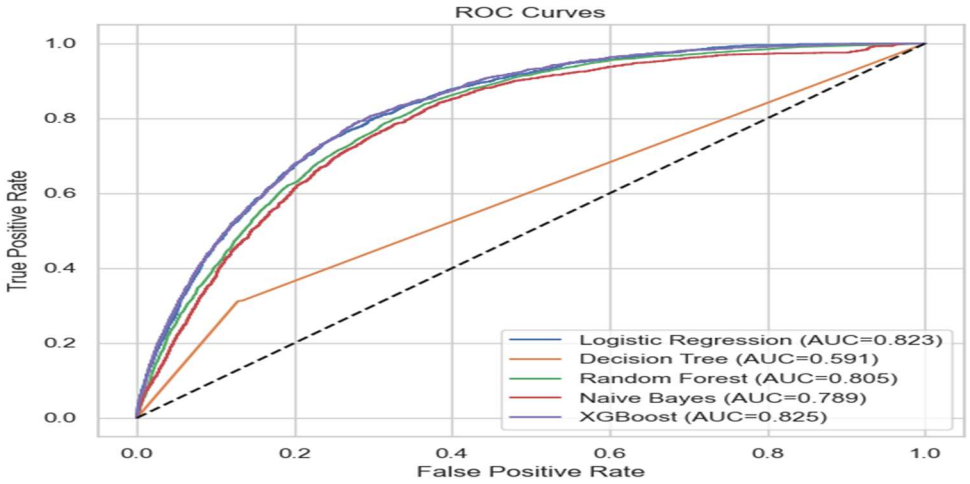

Figure 2: ROC Curves for model comparison

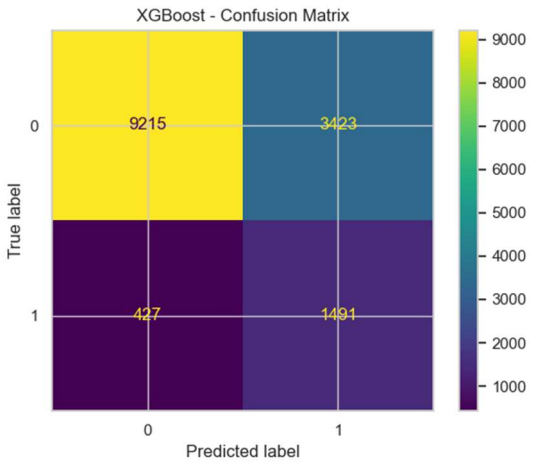

Figure 3: Confusion Matrix for XGBoost

6. Model Explainability (SHAP Analysis)

SHAP analysis was applied to interpret model predictions.

- Top predictors included:
  - BMI
  - Age
  - Physical activity levels
  - Hypertension status
  - Smoking behavior
- Interaction: HighBP × HighChol

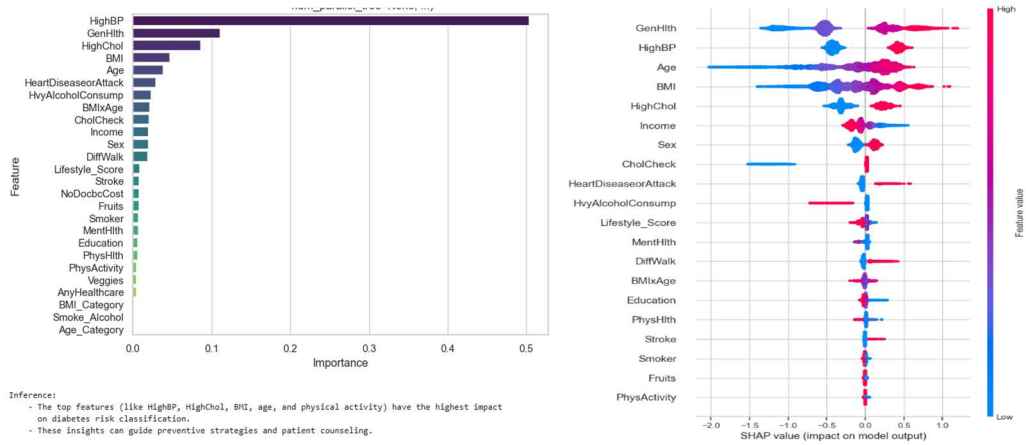

Figure 4: SHAP summary plot (feature importance)

- **Insights**: Poor general health and older age increase risk; lifestyle factors are protective

---

Challenges & Fixes

- SMOTE improved recall
- GridSearchCV across 270 combinations
- SHAP required sampling due to runtime limits

---

7. Key Findings

- XGBoost provided the best predictive accuracy while retaining robustness
- Lifestyle factors such as smoking and physical activity strongly influenced risk
- Demographic variables (age, BMI) were the most consistent predictors
- Preventive interventions should focus on high-BMI and older populations

8. Recommendations

- Use predictive modeling outputs to target health campaigns
- Encourage physical activity and smoking cessation
- Prioritize screening for high-risk age/BMI groups
- Leverage interpretable models for policy communication

9. Conclusion

The analysis highlights key behavioral and demographic risk factors for diabetes. By combining statistical analysis, machine learning, and explainable AI methods, the study provides actionable insights for targeted health interventions.

- XGBoost offers strong predictive power and interpretability
- SHAP confirms clinical relevance of top features
- **Next Steps**:
    - Deploy model in public health dashboards
    - Integrate with healthcare systems for early screening
    - Explore longitudinal BRFSS data for time-series modeling