Diabetes Data Analysis Project Report

1. Introduction

This report presents the outcomes of a data analysis project conducted for the DSA210 course. The primary objective was to examine the prevalence and patterns of diabetes within Allegheny County, Pennsylvania, using administrative health insurance claim data for the years 2015 and 2016. The data was sourced from the Western Pennsylvania Regional Data Center (WPRDC), which aggregates and publishes public datasets provided by Allegheny County and the City of Pittsburgh. The diabetes datasets used in this project were originally contributed by multiple health-related organizations and represent de-identified, aggregated claim records.

The project aimed to answer the following questions:

- Does diabetes prevalence differ significantly between Private and Medicaid insurance groups?
- Did Medicaid diabetes prevalence change from 2015 to 2016?
- Can machine learning models classify risk levels based on available prevalence features?

To address these questions, the project covered all stages of data processing, exploratory analysis, statistical testing, and classification model implementation using Python.

2. Dataset Description and Preprocessing

Three separate files were used during the project:

- diabetes2015.csv: De-identified insurance claim records from 2015.
- diabetes_all_2016.csv: Corresponding records from 2016.
- ndc-codes.xlsx: A file that maps drug codes (NDCs) to diabetes-related medications.

Each record in the 2015 and 2016 datasets contains aggregate information per census tract. The data is segmented by insurance type (Private, Medicare, and Medicaid), and further broken down by total enrollees and the number of individuals diagnosed with diabetes (1+ or 2+ diagnoses). For example, columns such as BPAD and BPAN refer to total enrollees and diabetic enrollees under private insurance, respectively.

The following renaming scheme was applied to increase clarity:

- CT → CensusTract
- BPAD → Private_Total, BPAN → Private_Diabetic_1+, BPAN2 → Private_Diabetic_2+
- BWAD, BWAN, BWAN2 → Medicare equivalents
- BMAD, BMAN, BMAN2 → Medicaid equivalents

These changes made it easier to reference and compute new metrics. For each record, I calculated diabetes prevalence rates by dividing diabetic counts by total enrollees for each insurance type. I then created additional features such as:

- Private_Prevalence and Medicaid_Prevalence
- Private_Diff and Medicaid_Diff (difference between 1+ and 2+ diagnoses prevalence)
- A Year column
- Medicaid_RiskLevel, a binned categorical variable: Low (<8%), Medium (8–12%), High (>12%)

Missing prevalence values were treated as zero, assuming no claim activity. Records without a Medicaid_RiskLevel were removed since this was the target variable for classification.

All numeric fields were standardized, and categorical variables were encoded where necessary, resulting in a clean and consistent dataset ready for analysis.

3. Exploratory Data Analysis and Hypothesis Testing

During the analysis phase, I developed and tested the following hypotheses:

- Null Hypothesis (H0-1): There is no significant difference in diabetes prevalence between Private and Medicaid insurance groups.
- Alternative Hypothesis (Ha-1): There is a significant difference in diabetes prevalence between Private and Medicaid insurance groups.
- Null Hypothesis (H0-2): There is no significant difference in Medicaid prevalence between the years 2015 and 2016.
- Alternative Hypothesis (Ha-2): There is a significant difference in Medicaid prevalence between the years 2015 and 2016.

The hypotheses were tested using independent-sample t-tests. The results were:

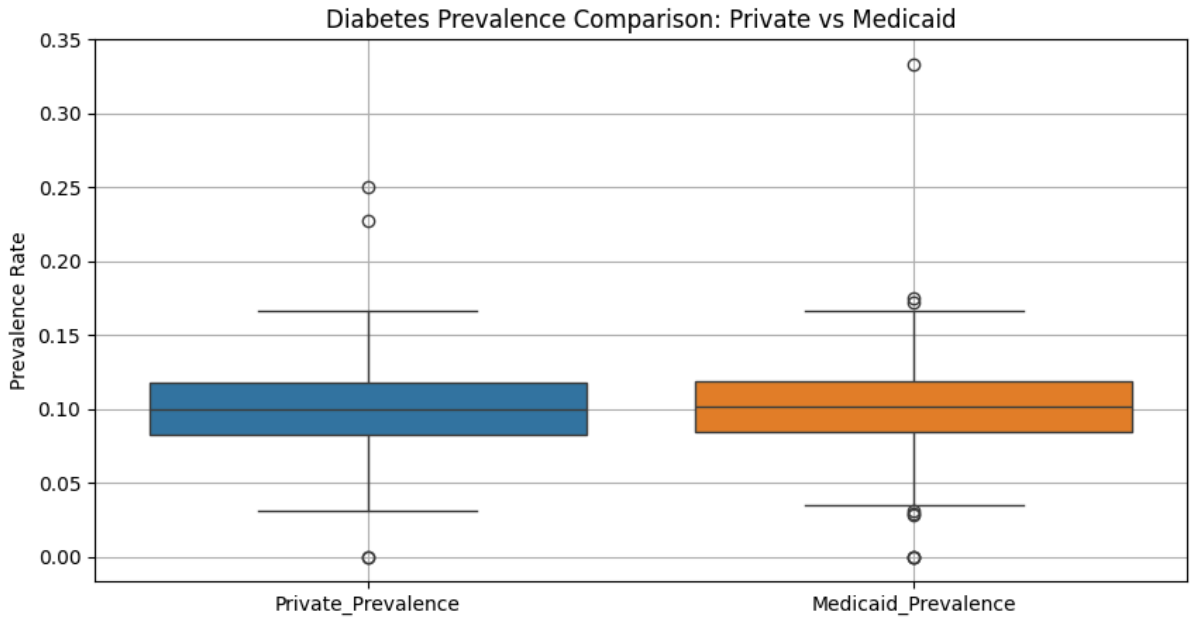

Figure 1. *Boxplot of Diabetes Prevalence by Insurance Type (2016)*

This boxplot shows the variation in diabetes prevalence between Private and Medicaid insurance types in 2016.

- H0-1: $p = 0.1662$ → fail to reject the null hypothesis.
- H0-2: $p < 0.0001$ → reject the null hypothesis.

Descriptive statistics:

- Private insurance prevalence: 10.60% (2015), 9.21% (2016)
- Medicaid prevalence: 10.53% (2015), 9.70% (2016)
- Standard deviations (2016): Private 0.0277, Medicaid 0.0280

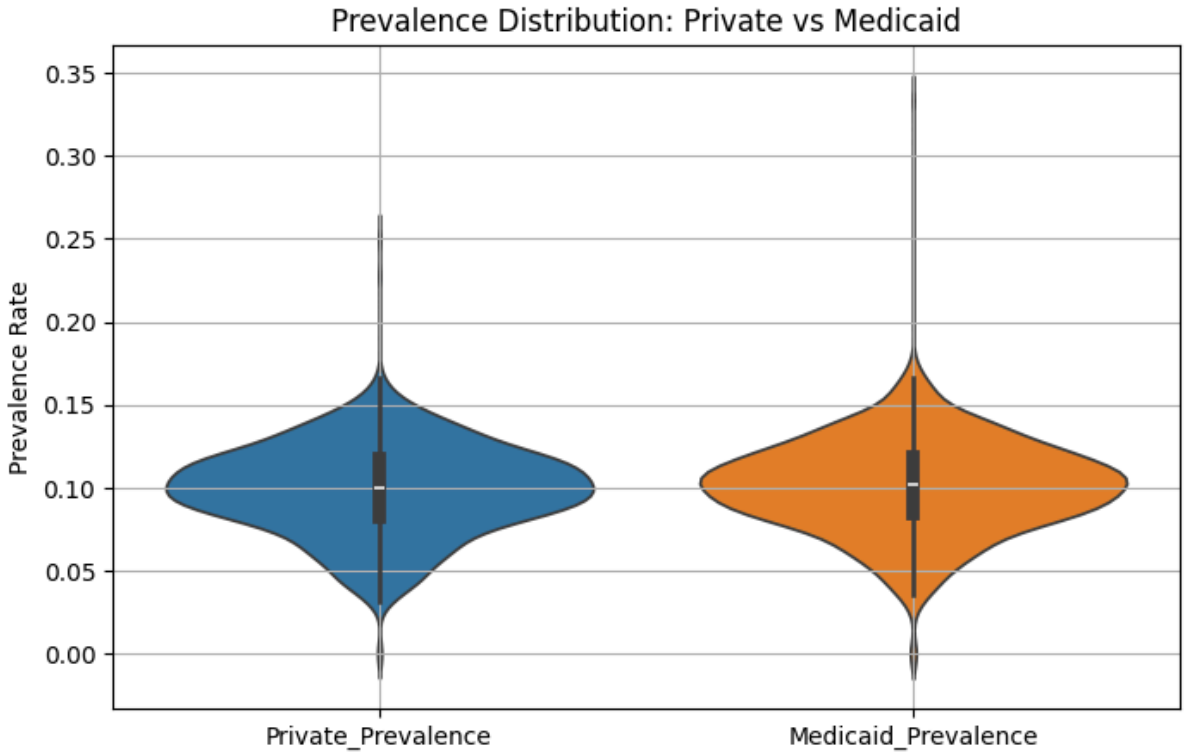

Figure 2. *Violin Plot of Private vs. Medicaid Prevalence (2015 vs 2016)*

This violin plot compares the distribution of Private vs. Medicaid prevalence in 2015 and 2016 across census tracts.

Visual comparisons were provided via boxplots, violin plots, histograms, and barplots. The results showed clear but relatively minor differences between years, especially within the Medicaid group. Some census tracts in 2016 reported zero Medicaid prevalence, which might reflect data capture limitations or true absence of cases.

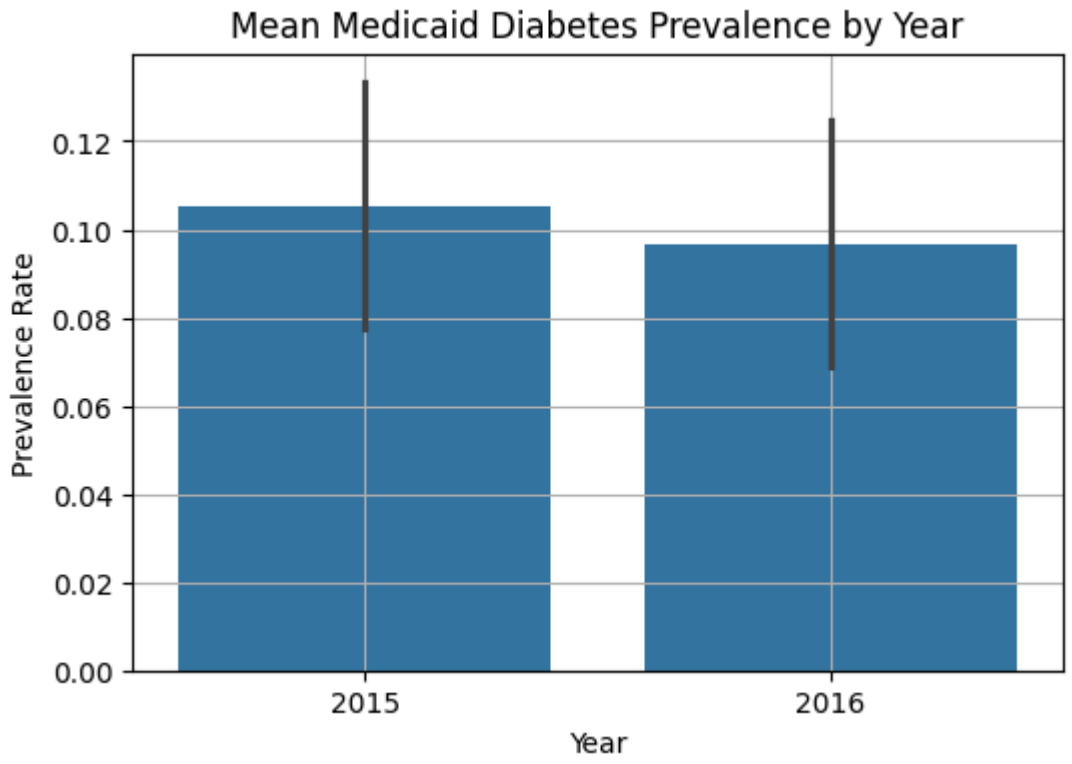

Figure 3. *Histogram of Private vs Medicaid Diabetes Prevalence by Year*

This histogram visualizes the distribution of diabetes prevalence by year, split between Private and Medicaid insurance groups.

4. Feature Engineering

To support classification tasks and improve interpretability, I engineered the following features:

- Private_Prevalence, Medicaid_Prevalence: Prevalence ratios per tract
- Private_Diff, Medicaid_Diff: Differences between loosely and strictly defined diabetic counts
- Year: Encoded as a numeric feature

- Medicaid_RiskLevel: Categorized based on Medicaid prevalence levels: <8% (Low), 8–12% (Medium), >12% (High)

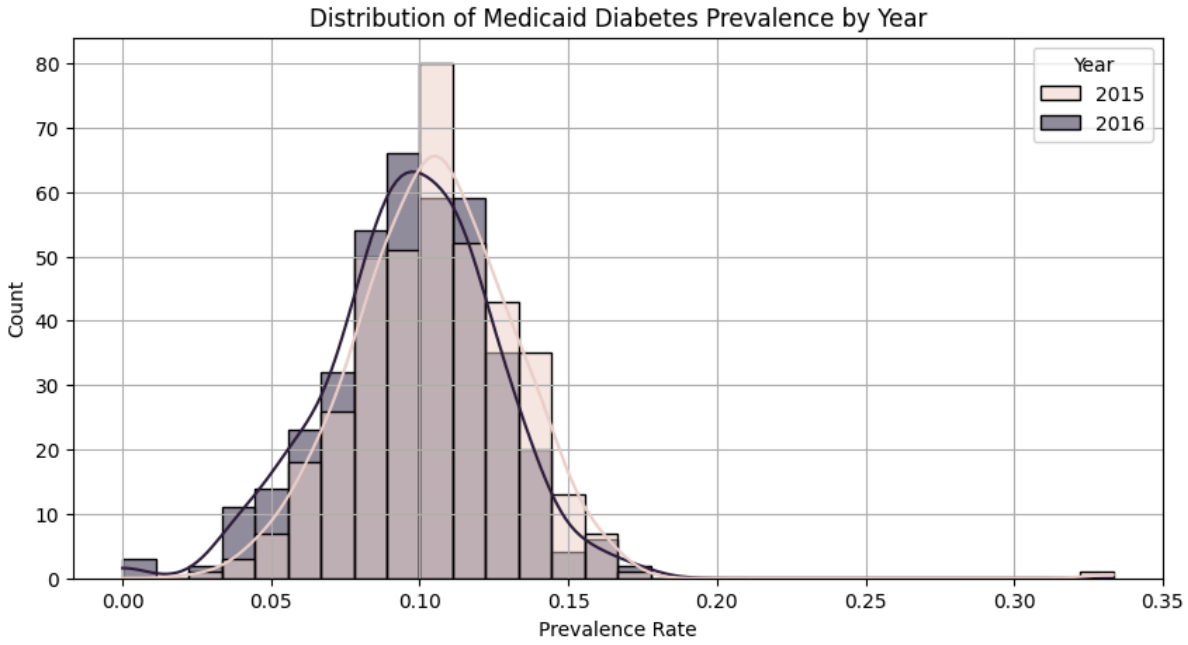

Figure 4. *Distribution of Medicaid Risk Levels*

This bar plot shows the distribution of census tracts across different Medicaid risk levels.

These features allowed me to classify risk levels using machine learning models.

5. Machine Learning Methods

Model Setup:

- **Features:** Private_Prevalence, Medicaid_Prevalence, Private_Diff, Medicaid_Diff, Year
- **Target:** Medicaid_RiskLevel
- **Split:** 80% train, 20% test (stratified)
- **Scaling:** StandardScaler

Models Used and Results:

Random Forest Classifier:

- Test Accuracy: 0.998
- 5-Fold CV Accuracy: 0.999 ± 0.003
- Precision, Recall, F1-score (all classes): 1.00 / 1.00 / 1.00

Logistic Regression:

- Test Accuracy: 0.95
- 5-Fold CV Accuracy: 0.960 ± 0.021
- Precision, Recall, F1-score:
    - High: 0.92 / 0.97 / 0.95
    - Low: 1.00 / 0.91 / 0.95
    - Medium: 0.95 / 0.97 / 0.96

XGBoost (before tuning):

- Test Accuracy: 1.00
- 5-Fold CV Accuracy: 0.996 ± 0.003
- Perfect precision, recall, and F1-score

XGBoost (after tuning):

Best Parameters:

- learning_rate: 0.01
- max_depth: 3
- n_estimators: 50

- Best CV Score: 0.99676
- Test Accuracy: 1.00
- All metrics: 1.00 across all classes

Feature Importances:

- Medicaid_Prevalence was the only feature with a significant importance score in XGBoost (≈1.0)
- Other features contributed negligibly.

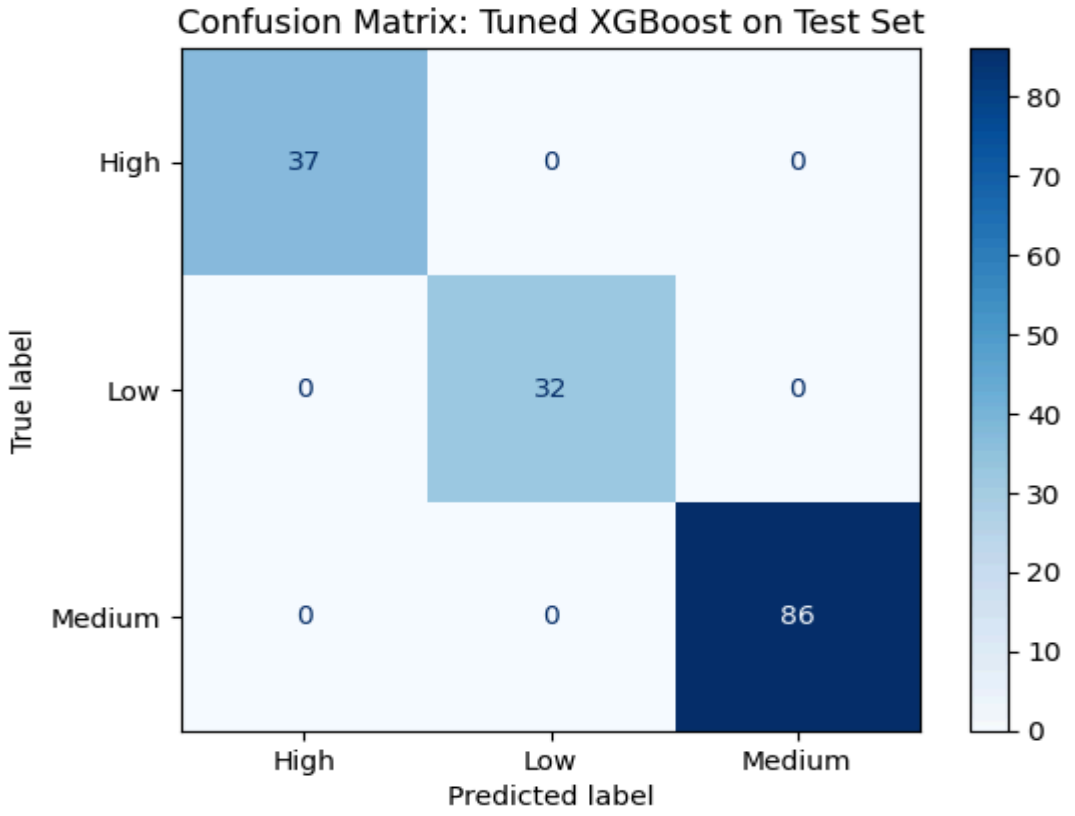

Figure 5. *Confusion Matrix: Tuned XGBoost on Test Set*

The model achieved perfect prediction across all classes (Low, Medium, High), indicating no misclassifications on the test data. Rows represent actual labels, while columns represent predicted labels.

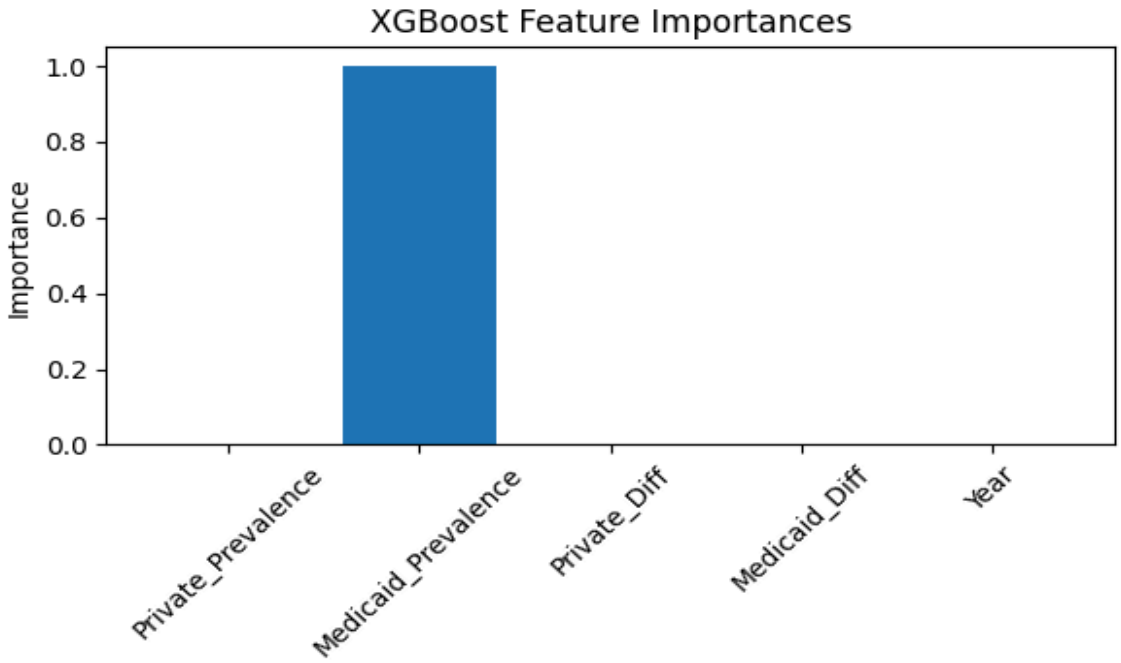

Figure 6. *Feature Importance Plot for Tuned XGBoost Model*

Medicaid_Prevalence emerged as the dominant feature, while the remaining features contributed negligibly to classification performance.

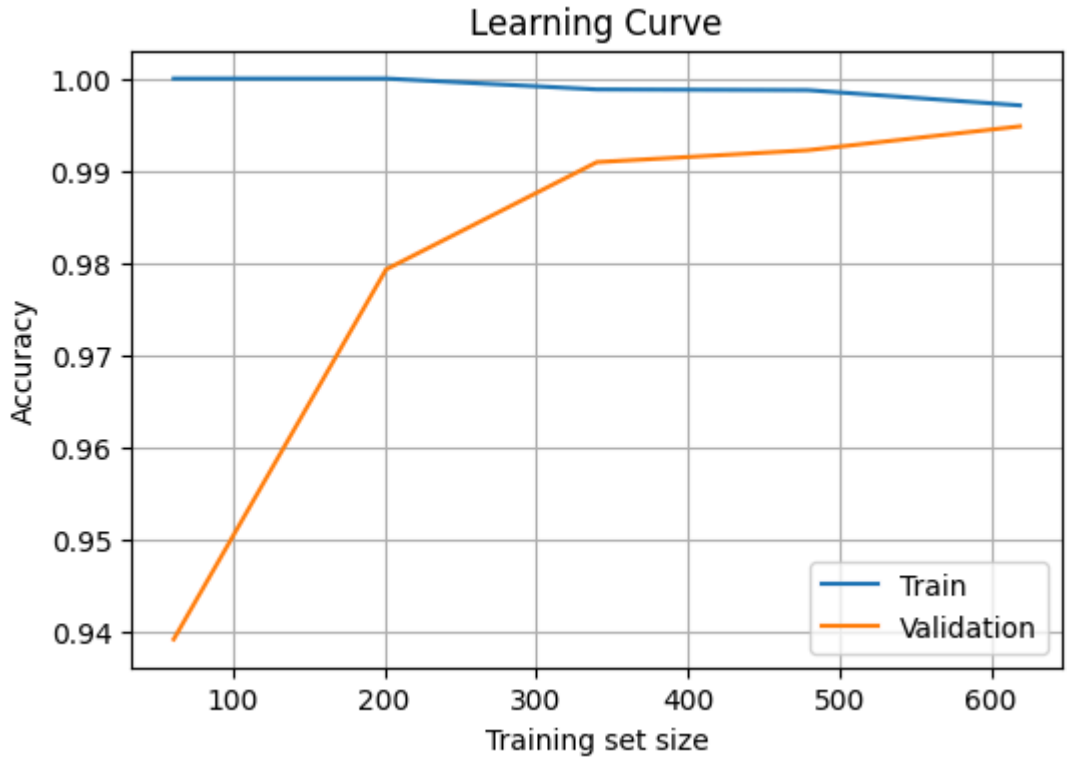

Figure 7. *Learning Curve of Tuned XGBoost Model*

Learning Curve:

- Showed minimal gap between training and validation performance
- Confirmed that the tuned XGBoost model generalized well with no sign of overfitting

6. Conclusion

This project showed how public health insurance claim data can be used to analyze diabetes prevalence and build predictive models. I performed preprocessing, hypothesis testing, visualization, and classification using multiple machine learning algorithms.

Among all models, the tuned XGBoost model showed the best accuracy and generalization. Medicaid prevalence emerged as the most predictive feature, highlighting the value of policy-level metrics in modeling public health outcomes.

7. References

- Western Pennsylvania Regional Data Center. (n.d.). *Diabetes* [Dataset]. Retrieved May ,2025, from https://catalog.data.gov/dataset/diabetes