Understanding Factors Contributing to Type 2 Diabetes

Micro-Project 4

https://github.com/tamirasianne/ANA-500.git

Tamira Hamilton

November 23, 2025

Problem Statement

- Type 2 Diabetes is an important issue because it remains one of the leading causes of death in The United States (Centers for Disease Control, 2023). Despite existing knowledge, there are challenges combining clinical practice with modernized diagnostic habits. The goal of this study is to analyze survey data provided by Behavior Risk Factor Surveillance System (BRFSS) of 457,670 individuals to ascertain how demographic and lifestyle data can be utilized to enhance diagnostic accuracy and early detection in adults while employing data-driven methods.

Hypothesis Formulation

- It is hypothesized that demographic and lifestyle attributes present in the dataset can be used to accurately predict Type 2 Diabetes diagnoses, exhibiting the potential of data-driven analyses to supplement traditional clinical decision-making.
  - H_0 : There is no association between demographic and lifestyle attributes and the diagnosis of Type 2 Diabetes in the U.S. population.
  - H_1: There is a positive association between demographic and lifestyle attributes and the diagnosis of Type 2 Diabetes in the U.S. population.

Acquire

- This study utilized data from the 2024 BRFSS to investigate factors associated with the presence of Type 2 Diabetes among adults in The United States. The BRFSS, overseen by The Centers for Disease Control and Prevention (CDC), is a nationwide survey that collects comprehensive data on demographics, chronic illnesses, health care utilization, and health-related behaviors. Conducted annually, it gathers responses from over 400,000 adults across all 50 states, The District of Columbia, and various U.S. territories. Utilizing self-reported data, the goal is to build a predictive model that can assist in identifying individuals at higher risk for Type 2 Diabetes. The findings aim to inform clinicians on public health strategies focused on early detection and prevention efforts.
- Amid the 301 attributes supplied in the dataset, 9 of them satisfy the need for this analysis: _DIABETE4, _AGEG5YR, _RACE, EXERANY2, SMOKE100, _RFDRHV9, _SEX, _BMI5CAT, and GENHLTH.

Prepare

- After querying the dataset, the next step involved exploring its structure and quality to understand how the data was organized. The subset data frame contains 457,670 rows and 9 columns. During the exploration phase we reviewed the types of attributes, previewed the first few rows, analyzed value counts for each attribute, and examined missing data patterns. All the 9 attributes were a 'float64' which translates to mean that they are standard integers. While inspecting the value counts of each category within the attributes, there were two that stood out to us. In our '_RACE' attribute, we have one category (White) that was overrepresented with 329,346 individuals; which is 72% of the data in that column.
- In our '_RFDRHV9' attribute, we have one category (No) that was overrepresented with 386,812 individuals; which is 85% of the data in that column. Out of the 9 attributes, 5 of them had missing values. DIABETE4(4), EXERANY2(3), and GENHLTH(5) all had mising values that was less than 1% of the data. SMOKE100 (28,860) has missing values that is 6.3% of the data and _BMI5CAT(43,037) has missing values that is 9.4% of the data. Although we reviewed summary statistics, all the integer-based attributes represent categorical data, so measures like mean and standard deviation were not meaningful for interpretation. This step assisted in discovering missing values, category imbalances, and data quality to address in preprocessing.
- Following the exploration, we prepared the data frame by cleaning and refining the data. To clean the data, we started by using Mode Imputation to handle the missing values for the three attributes that were less than 1% of the data. Mode Imputation was the best decision because it doesn't distort too much of the distribution and doesn't introduce bias. To handle the rest of the missing values in SMOKE100 and _BMI5CAT, we decided to employ Sochastic Categorical Imputation. This strategy randomly places the missing values based on the frequency proportions that were present between the categories in the attributes. This approach preserves the distribution of the data, doesn't introduce bias, and assist in bringing stability to the data that would be prepped for modeling. We also recoded the variables to narrow down the categories and renamed the attributes. This helped for better interpretability. These steps ensured that the data frame was accurate, consistent, and ready for further analysis and visualization.

Prepare cont.

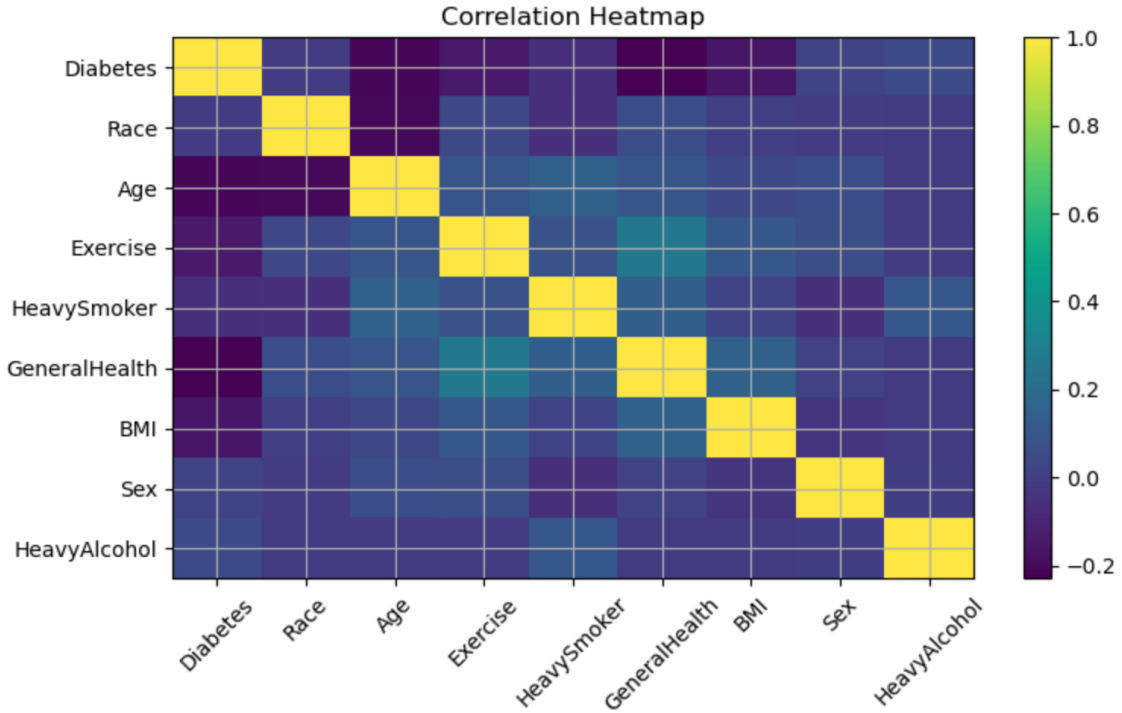
- After querying the dataset, the next step involved exploring its structure and quality to understand how the data was organized. The subset data frame contains 457,670 rows and 9 columns. During the exploration phase we reviewed the types of attributes, previewed the first few rows, analyzed value counts for each attribute, and examined missing data patterns. All the 9 attributes were a 'float64' which translates to mean that they are standard integers. While inspecting the value counts of each category within the attributes, there were two that stood out to us.
- In our '_RACE' attribute, we have one category (White) that was overrepresented with 329,346 individuals; which is 72% of the data in that column. In our '_RFDRHV9' attribute, we have one category (No) that was overrepresented with 386,812 individuals; which is 85% of the data in that column. Out of the 9 attributes, 5 of them had missing values. DIABETE4(4), EXERANY2(3), and GENHLTH(5) all had mising values that was less than 1% of the data. SMOKE100 (28,860) has missing values that is 6.3% of the data and _BMI5CAT(43,037) has missing values that is 9.4% of the data.
- Although we reviewed summary statistics, all the integer-based attributes represent categorical data, so measures like mean and standard deviation were not meaningful for interpretation. This step assisted in discovering missing values, category imbalances, and data quality to address in preprocessing.

Prepare cont.

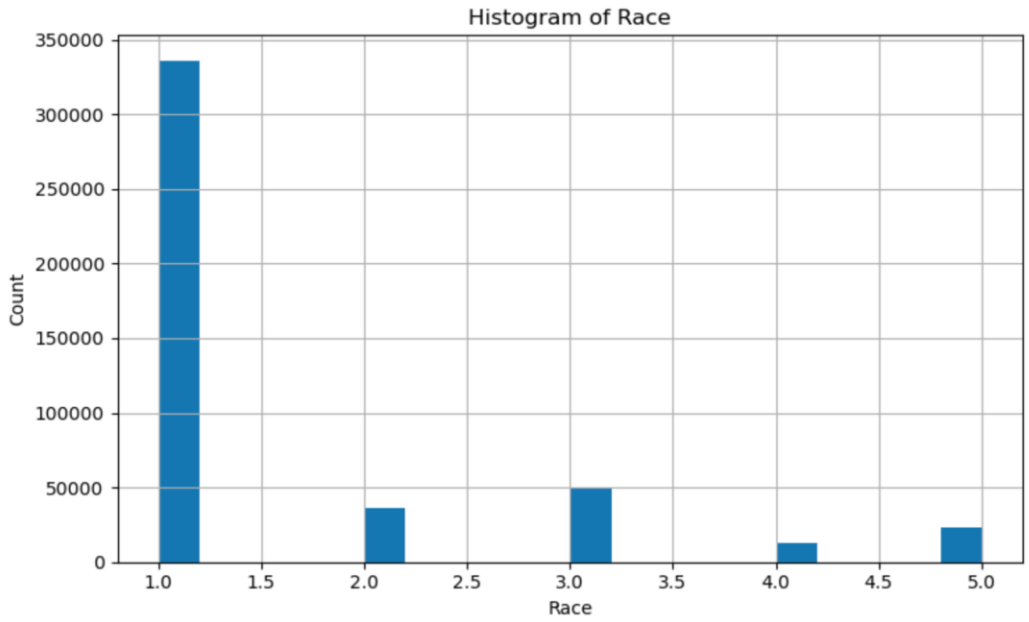
- Following the exploration, we prepared the data frame by cleaning and refining the data. To clean the data, we started by using Mode Imputation to handle the missing values for the three attributes that were less than 1% of the data. Mode Imputation was the best decision because it doesn't distort too much of the distribution and doesn't introduce bias.
- To handle the rest of the missing values in SMOKE100 and _BMI5CAT, we decided to employ Sochastic Categorical Imputation. This strategy randomly places the missing values based on the frequency proportions that were present between the categories in the attributes. This approach preserves the distribution of the data, doesn't introduce bias, and assist in bringing stability to the data that would be prepped for modeling.
- We also recoded the variables to narrow down the categories and renamed the attributes. This helped for better interpretability. These steps ensured that the data frame was accurate, consistent, and ready for further analysis and visualization.

Analyze data

- After procressing, we wanted to get an overall understanding of the data, so we employed some visualizations. We utilized Correlation Heatmaps, Histograms, and Bar Charts. A correlation heatmap is used to assist in discovering patterns and strong relationships across the attributes. Examining our target variable Diabetes, we observed that as the Diabetes category tends to increase, it has a negative correlation with Age, General Health, and slightly weaker negative relationships with BMI and Exercise.
- This pattern would suggest that as an individual diagnosed with Diabetes, they tend to be younger, report poorer general health, and were somewhat less likely to have a higher BMI and exercise levels. However, this inverse relationship may reflect how the data were collected or categorized prior to analysis, as Diabetes tends to increase with age. Conversely, there was a positive relationship between sex and alcohol use in relation to Diabetes, suggesting that females with diabetes were more likely to report heavy alcohol consumption than males.
- We also want to mention a couple of other relationships that stood out to us. As exercise levels increase, individuals tend to report higher levels of well-being (general health). Also, that as an individual consumes smoking heavily, they also tend to indulge in heavy use of alcohol, as this too tends to increase with age; suggesting that as an individual gets older, they tend to indulge in smoking and drinking more. Again, these patterns may reflect how the data were collected or coded rather than casual effects.

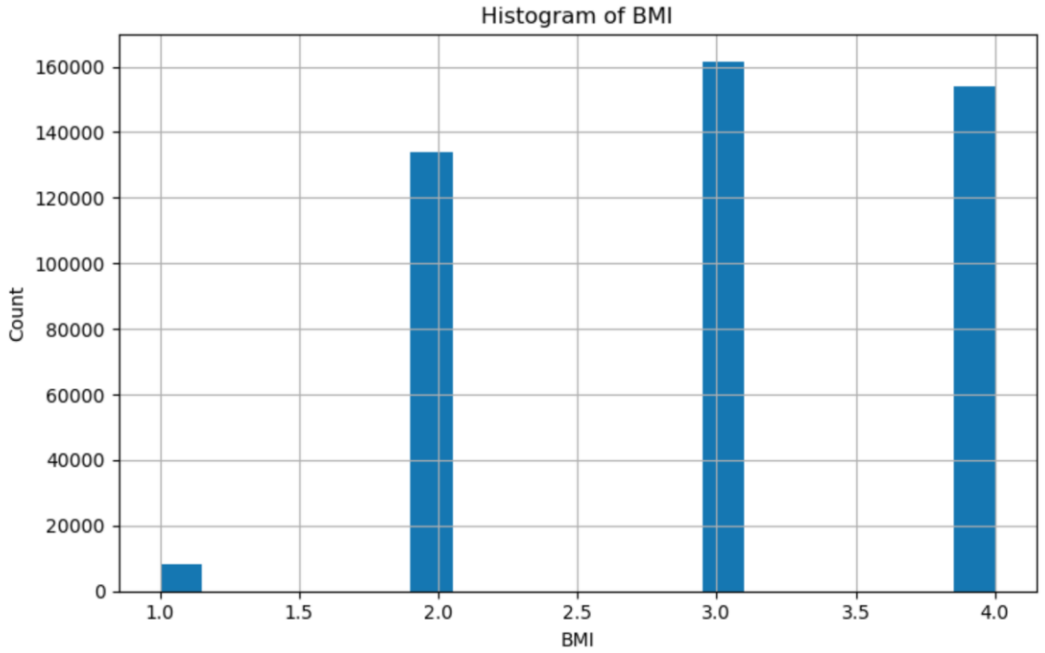


Analyze data cont.



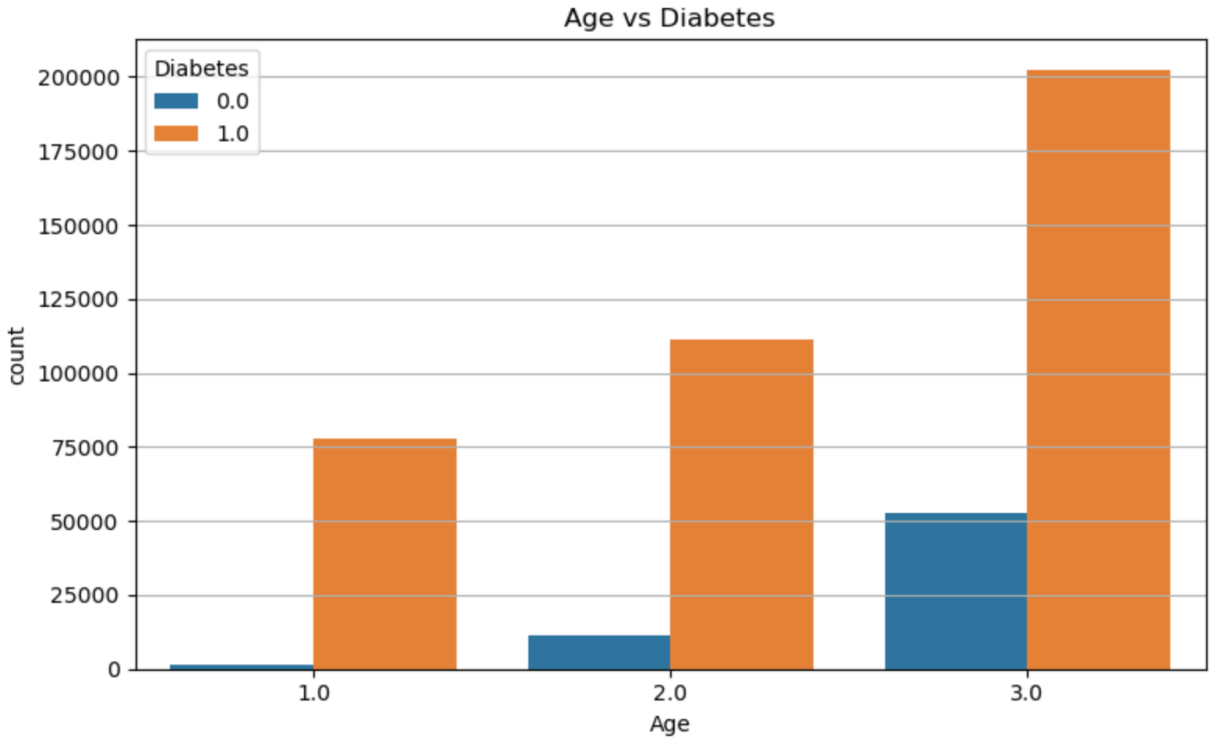
- Histograms are useful when you want to visualize the overall distribution of an attribute. The histogram of Race shows the distribution of individuals across racial categories in the data frame. The majority of the individuals are in category 1 (White), followed by smaller counts in categories 2 (Black) and 3 (Hispanic), with very few individuals in the remaining categories. This visualization highlights the relative representation of different racial groups and the imbalance between them within the frame.

Analyze data cont.



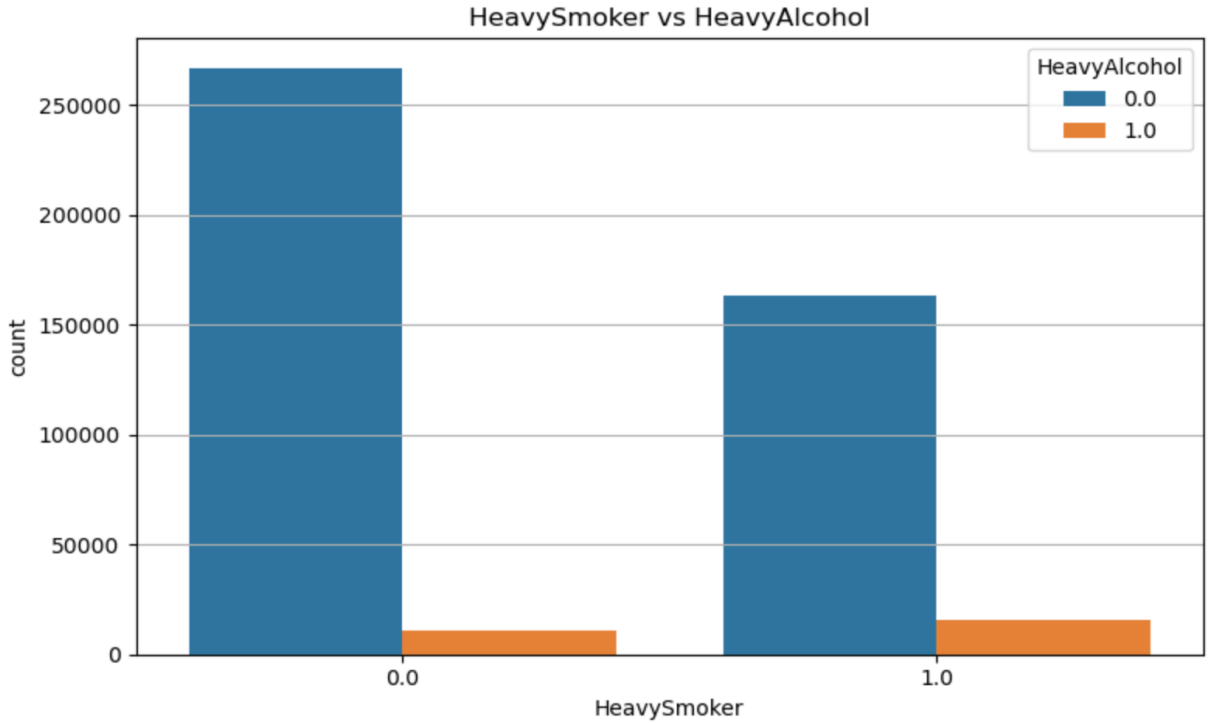
- The histogram of BMI categories shows that most individuals fall within three categores: 2 (Normal), 3(Overweight), and 4 (Obese), with very few individuals in 1 (Underweight) category. This indicates that the sample is concentrated in the middle to higher BMI ranges, with the Underweight category being underrepresented.

Analyze data cont.



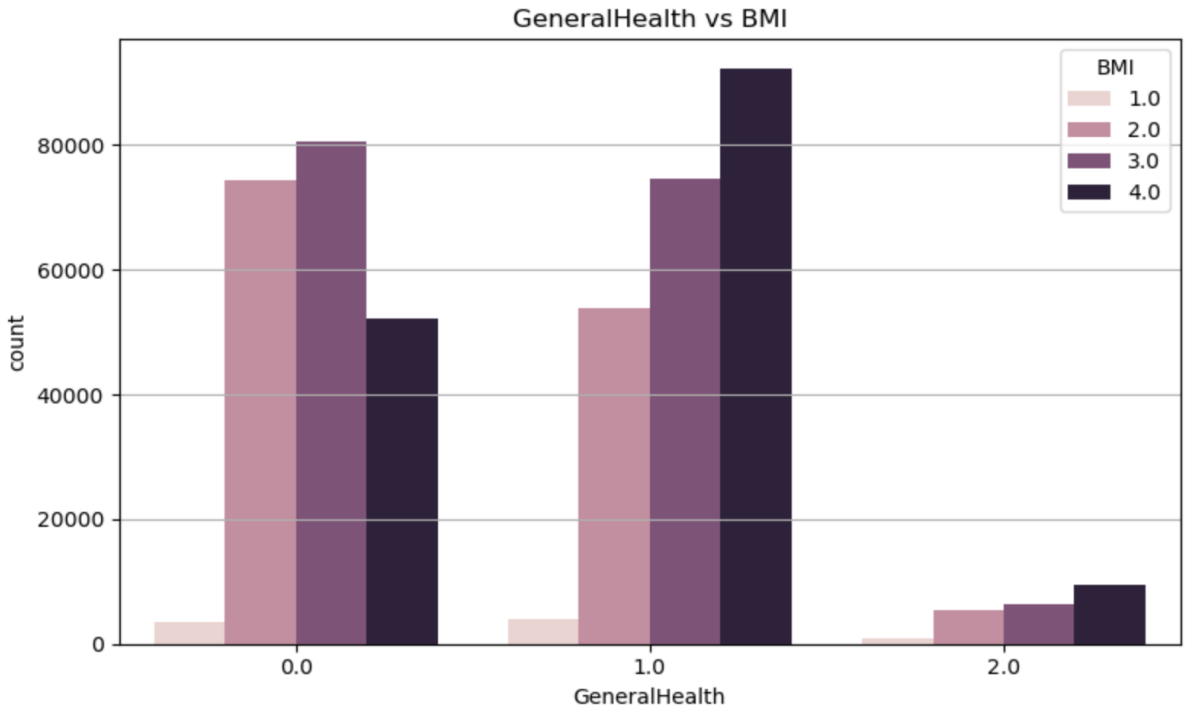
- Count plots are utilized to showcase the comparison between two or more categorical attributes. The count plot that shows a comparison between Age and Diabetes displays that individuals, across age groups, indicates that both categories are more numerous in the older age groups. This pattern reflects the distribution of the sample rather than a direct relationship.

Analyze data cont.



- The count plot comparing heavy smoking and heavy alcohol use shows that more individuals fall into a category of neither a heavy smoker nor hevay drinker. A smaller number of individuals enage in both heavy smoking and hevay drinking, while the remaining groups have intermediate counts. This suggests a trend where heavy smoking and heavy alcohol use sometimes occur together, although most individuals do not engage in either behavior.

Analyze data cont



- We narrowed the categories given in this attribute to three categories: 0 = Good, 1 = Average, and 2 = Poor. The count plot comparing general health and BMI indicates that most individuals with good or average health fall into the normal, overweight, or obese BMI categories. Fewer individuals are in poor health or extreme BMI categories, suggesting that healthier individuals tend to have mid-range BMI values in this sample.

Analyze data cont.

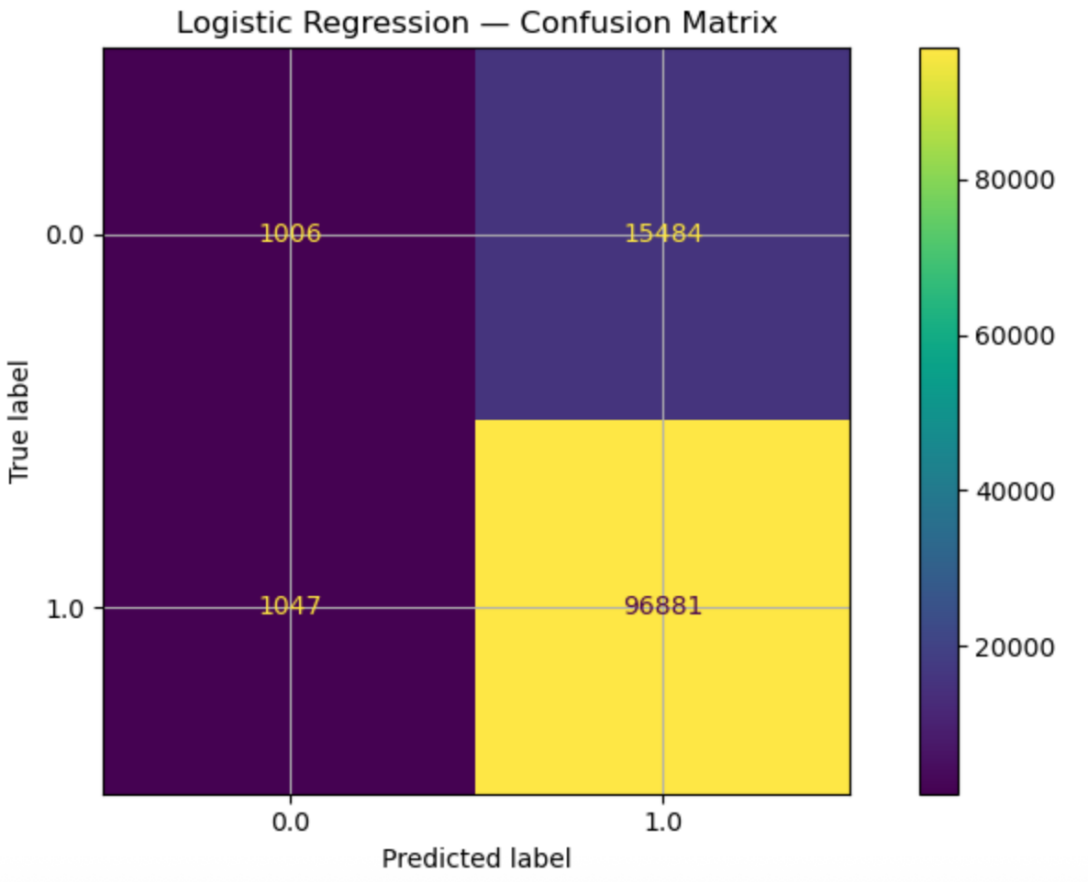
- After employing these visualizations to get a clearer picture of the overall data, we developed two classification models: Logistic Regression and Linear Support Vector Machine (SVM), to evaluate how well the data can predict Type 2 Diabetes. Logistic Regression was employed as our baseline linear model, and Linear SVM was selected for comparison as a more robust maximum-margin classifier. Although the RBF kernel was initially considered, due to the computational constraints like data frame size, it was not used. We decided to develop two more models: Simple Multi-Layer Perception (MLP) and Enhanced Multi-Layer Perceptron (MLP) in a continuation of trying to understand how well the data can predict Type 2 Diabetes. Simple MLP was used as another baseline model and Enhanced MLP was selected for comparison as a more robust maximum-margin classifier.

Report – Logistic Regression

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.49      0.06      0.11     16490
         1.0       0.86      0.99      0.92     97928

    accuracy                           0.86    114418
   macro avg       0.68      0.53      0.51    114418
weighted avg       0.81      0.86      0.80    114418

ROC-AUC: 0.7737650058267336
```

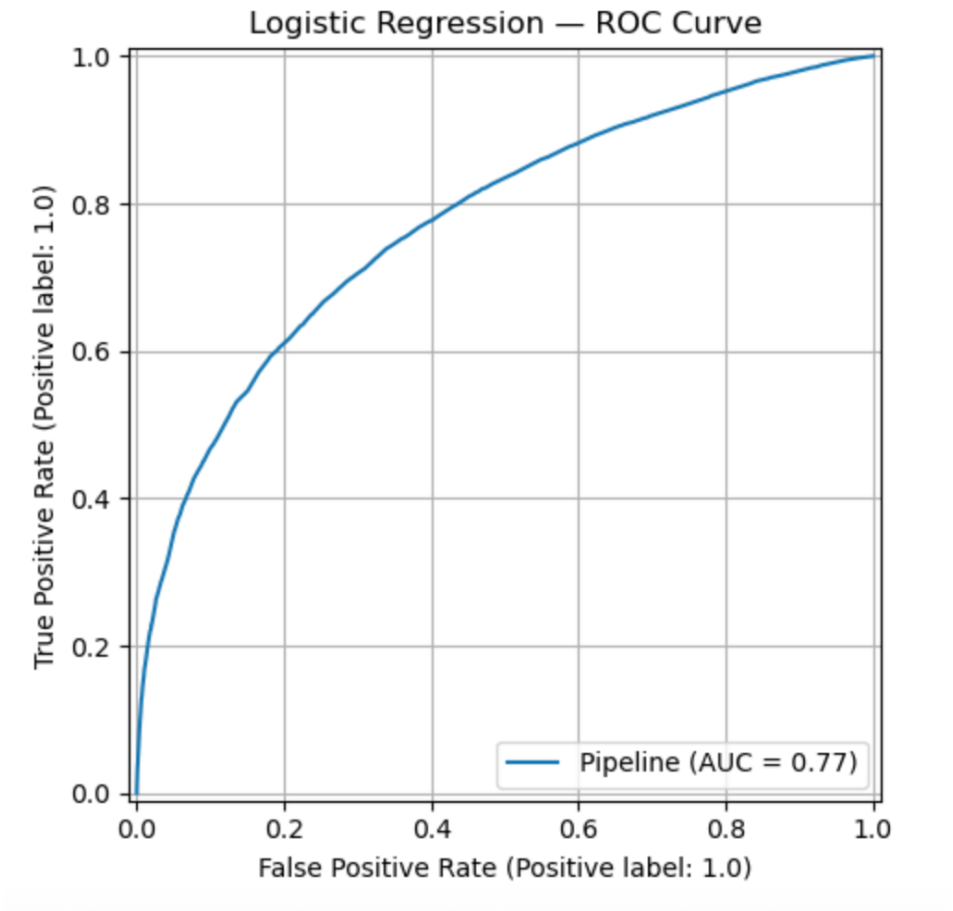
- **Metrics used:** Precision, Recall, F1-score.
- **Precision:** How often predicted labels are correct.
- **Recall:** How well the model identifies actual cases.
- **Performance on Diabetes classification:**
- **Diabetes:** Precision 49%, Recall 6%, F1-score 11% → model struggles to detect positive cases likely due to class imbalance.
- **No Diabetes:** Precision 86%, Recall 99%, F1-score 92% → model accurately identifies negative cases.

Report – Logistic Regression



- Displays true vs predicted labels for each class.
- Highlights **model errors**: false positives and false negatives.
- **Diabetes:** 0.8% correctly classified, 0.9% missed (false negatives).
- **No Diabetes:** 84.6% correctly classified, 13.5% misclassified (false positives).
- Helps identify which classes the model struggles with, which in this case is shown that it struggles with predicting the 'yes' category.

Report – Logistic Regression



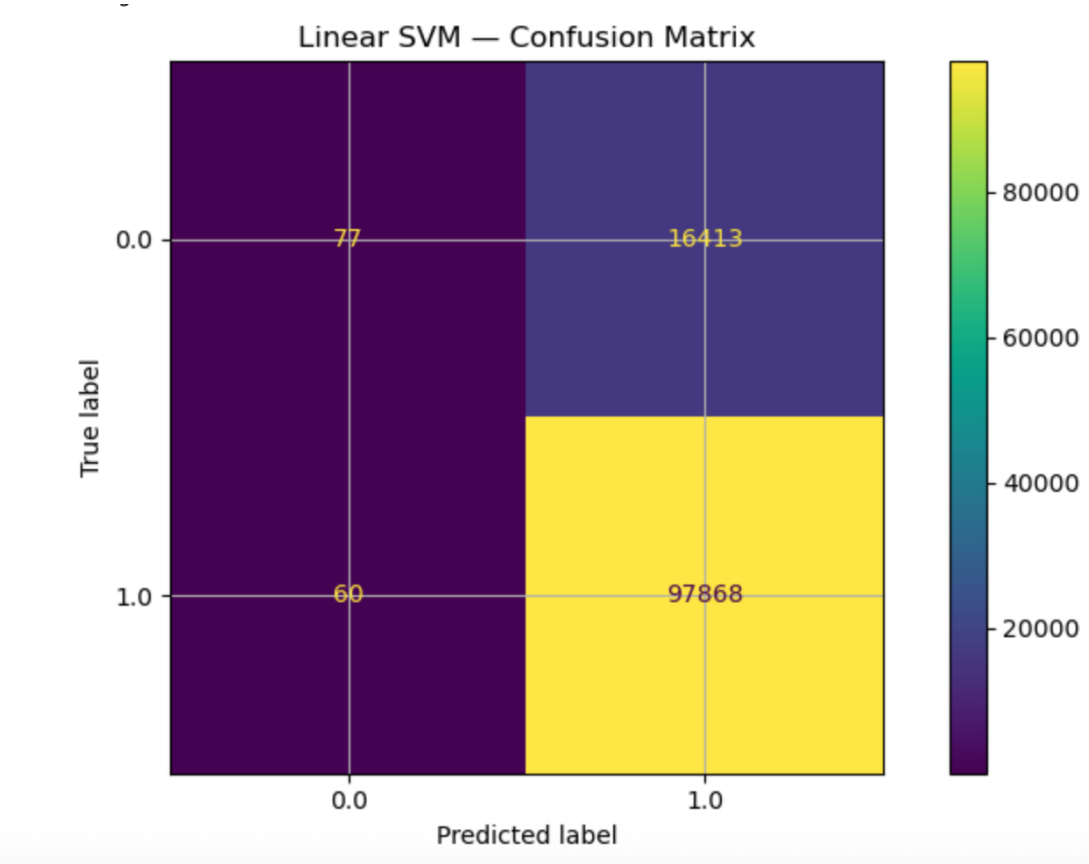
- **ROC Curve:** Plots True Positive Rate vs False Positive Rate across thresholds.
- **AUC = 0.77:** Indicates moderately good ability to distinguish Diabetes vs non-Diabetes.
- **Insights:**
- Model detects many non-Diabetes cases accurately.
- Model struggles with Diabetes cases → low recall / false negatives.
- **Implication:** Threshold adjustment or additional data could improve detection of positive cases.

Report – Linear SVM

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.56      0.00      0.01     16490
         1.0       0.86      1.00      0.92     97928

    accuracy                           0.86    114418
   macro avg       0.71      0.50      0.47    114418
weighted avg       0.81      0.86      0.79    114418

ROC-AUC: 0.7734070120897725
```

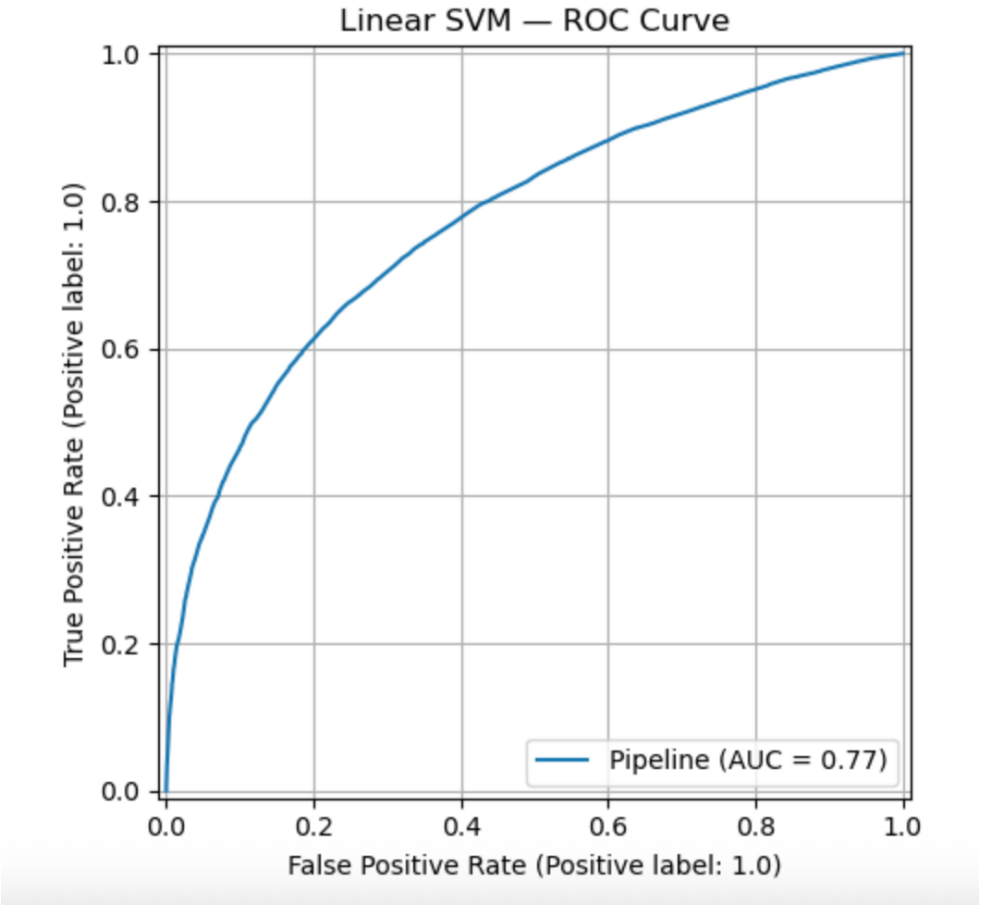
- **Metrics used:** Precision, Recall, F1-score.
- **Precision:** How often predicted labels are correct.
- **Recall:** How well the model identifies actual cases.
- **Performance on Diabetes classification:**
- **Diabetes:** Precision 56%, Recall 0**%**, F1-score 1 % → model fails to detect any positive cases likely due to class imbalance.
- **No Diabetes:** Precision 86 %, Recall 100 %, F1-score 92 % → model accurately identifies negative cases.

Report – Linear SVM



- Displays true vs predicted labels for each class.
- Highlights model errors: false positives and false negatives.
- **Diabetes:** 0.06% correctly classified, 0.05% missed (false negatives).
- **No Diabetes:** 85.5% correctly classified, 14.3% misclassified (false positives).
- Helps identify which classes the model struggles with, which in this case is shown that it struggles with predicting the 'yes' category.

Report – Linear SVM



- **ROC Curve:** Plots True Positive Rate vs False Positive Rate across thresholds.
- **AUC = 0.77:** Indicates moderately good ability to distinguish Diabetes vs non-Diabetes.
- **Insights:**
- Model detects many non-Diabetes cases accurately.
- Model struggles with Diabetes cases → low recall / false negatives.
- **Implication:** Threshold adjustment or additional data could improve detection of positive cases.

Report – Comparison of LR vs. Linear SVM

	model	accuracy	precision	recall	f1	roc_auc
0	LogReg (best grid)	0.855521	0.862199	0.989308	0.921391	0.773765
1	Linear-SVM (best grid)	0.856028	0.856380	0.999387	0.922374	0.773407

- **Overall performance:** Both models show similar metrics overall, but class-specific differences are important.
- **Recall:**
- Logistic Regression: 98.9% → detects most positive cases.
- Linear SVM: 0% for "Yes Diabetes" → fails to detect any true positive cases.
- **Precision & F1-score:** Nearly identical overall but misleading for SVM due to the 0% recall on the positive class.
- **ROC-AUC:** Very similar (0.774 vs 0.773) → overall separation of classes looks okay but doesn't reflect the failure to detect positive cases in SVM.
- **Recommendation / Next Steps:**
- Logistic Regression is preferable because it identifies positive cases.
- SVM with 0% recall is not acceptable for detecting Diabetes.
- Consider threshold tuning, class balancing, or feature engineering to improve detection if you revisit SVM.

Report – Simple Multi-Layer Perceptron

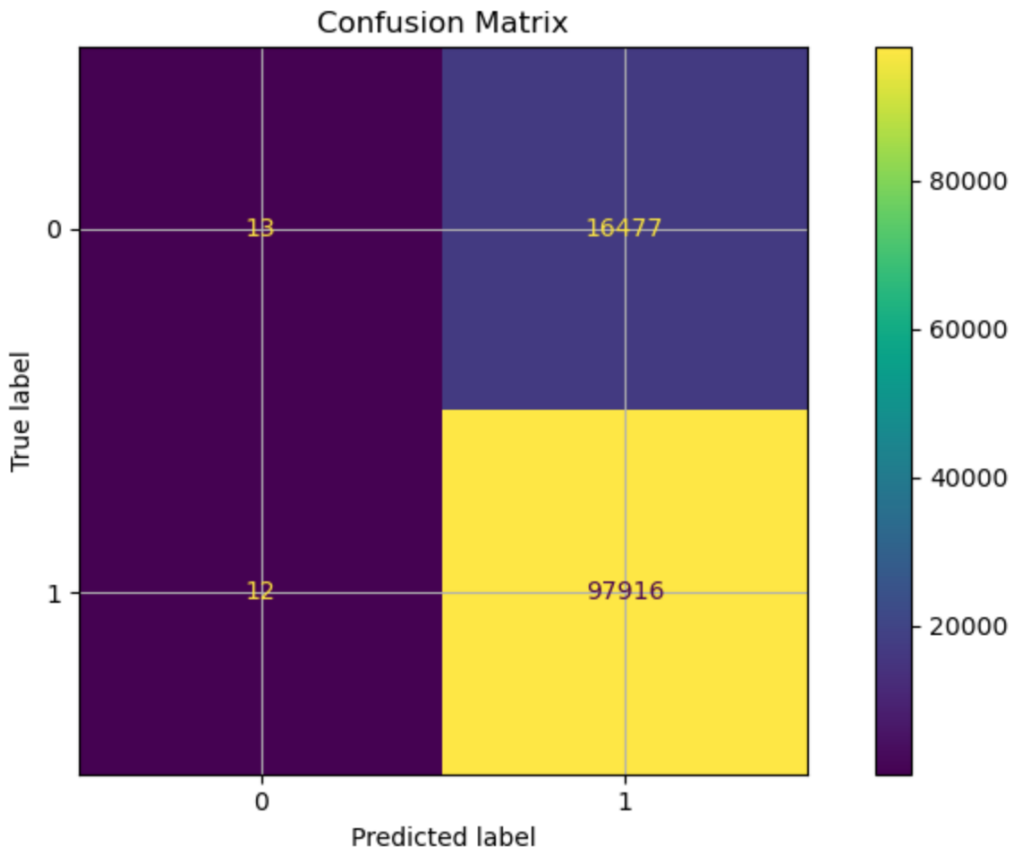
- The simple MLP displayed very little progress across epochs. Its training loss flattened early, while the validation loss fluctuated noticeably, indicating that the model didn't have enough capacity to capture consistent patterns. This suggests the model was underfitting and generalizing poorly.

Report – Simple Multi-Layer Perceptron

	precision	recall	f1-score	support
0.0	0.52	0.00	0.00	16490
1.0	0.86	1.00	0.92	97928
accuracy			0.86	114418
macro avg	0.69	0.50	0.46	114418
weighted avg	0.81	0.86	0.79	114418

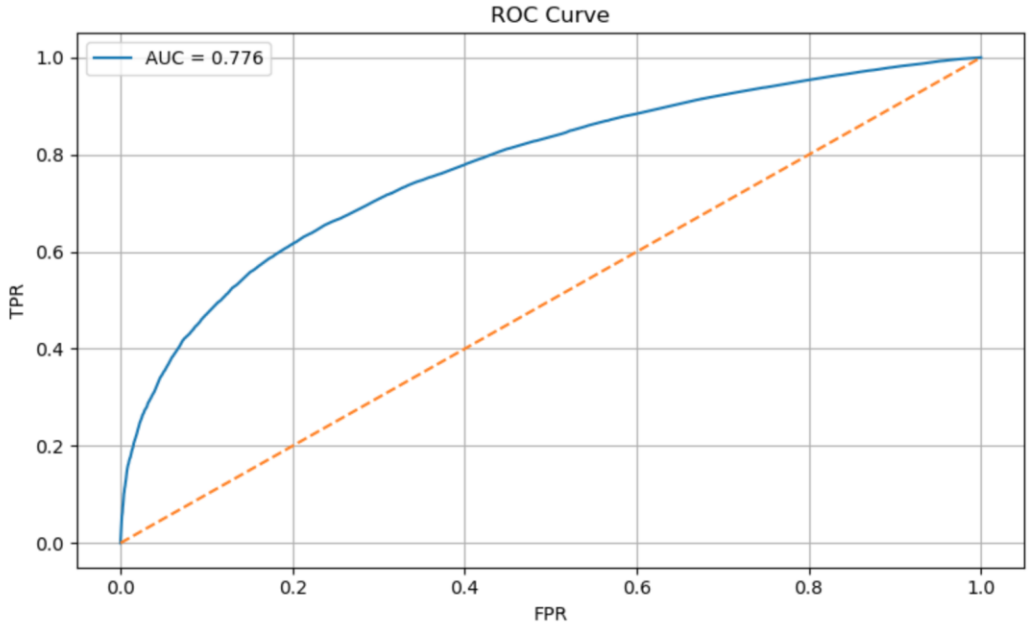
- **Metrics used:** Precision, Recall, F1-score.
- **Precision:** How often predicted labels are correct.
- **Recall:** How well the model identifies actual cases.
- **Performance on Diabetes classification:**
- **Diabetes:** Precision 52%, Recall 0%, F1-score 0% → model struggles to detect positive cases likely due to class imbalance.
- **No Diabetes:** Precision 86%, Recall 100%,F1-score 92% → model accurately identifies negative cases.

Report – Simple Multi-Layer Perceptron



- Displays true vs predicted labels for each class.
- Highlights **model errors**: false positives and false negatives.
- **Diabetes:** 0.01% correctly classified, 0.01%missed (false negatives).
- **No Diabetes:** 85.5% correctly classified,14.4% misclassified (false positives).
- Helps identify which classes the model struggles with, which in this case is shown that it struggles with predicting the 'yes' category.

Report – Simple Multi-Layer Perceptron



- **ROC Curve:** Plots True Positive Rate vs False Positive Rate across thresholds.
- **AUC = 0.77:** Indicates moderately good ability to distinguish Diabetes vs non-Diabetes.
- **Insights:**
- Model detects many non-Diabetes cases accurately.
- Model struggles with Diabetes cases →low recall / false negatives.
- **Implication:** Threshold adjustment or additional data could improve detection of positive cases.

Report – Enhanced Multi-Layer Perceptron

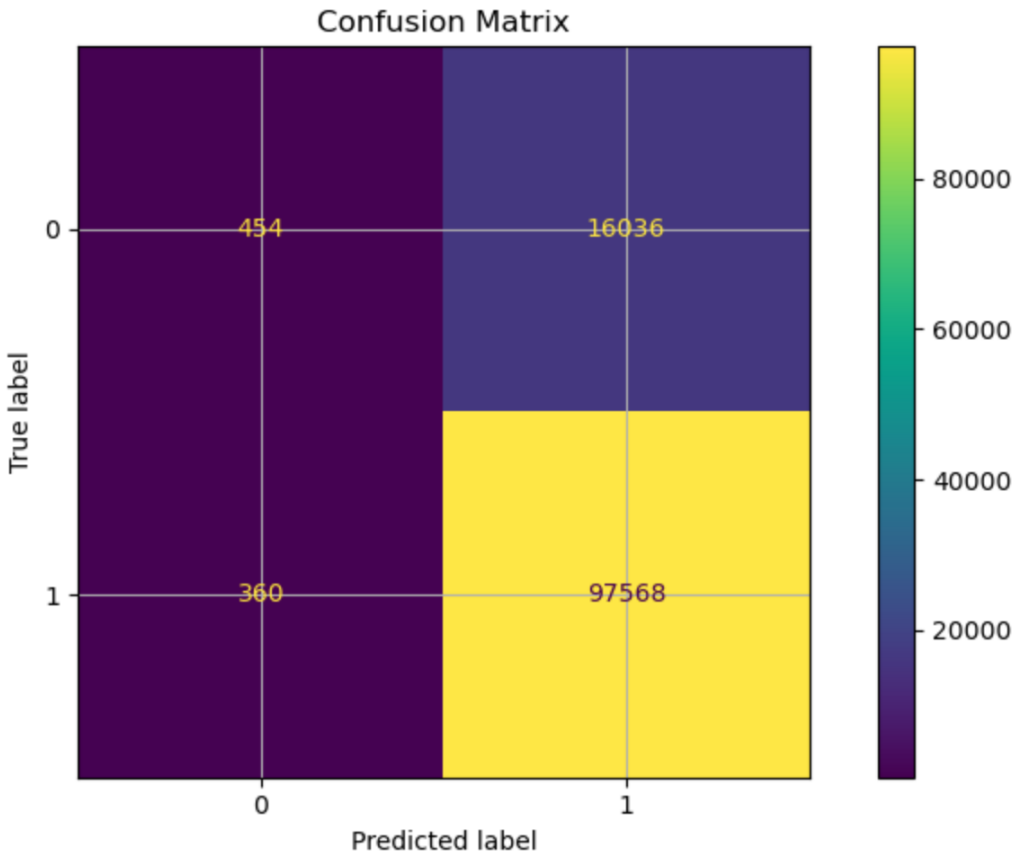
- The enhanced MLP showed a much clearer downward trend in training loss, and the validation loss remained more stable throughout training. Although there were a few small spikes, the validation curve stayed relatively low and consistent, indicating better learning and stronger generalization compared to the simple model.

Report – Enhanced Multi-Layer Perceptron

	precision	recall	f1-score	support
0.0	0.56	0.03	0.05	16490
1.0	0.86	1.00	0.92	97928
accuracy			0.86	114418
macro avg	0.71	0.51	0.49	114418
weighted avg	0.82	0.86	0.80	114418

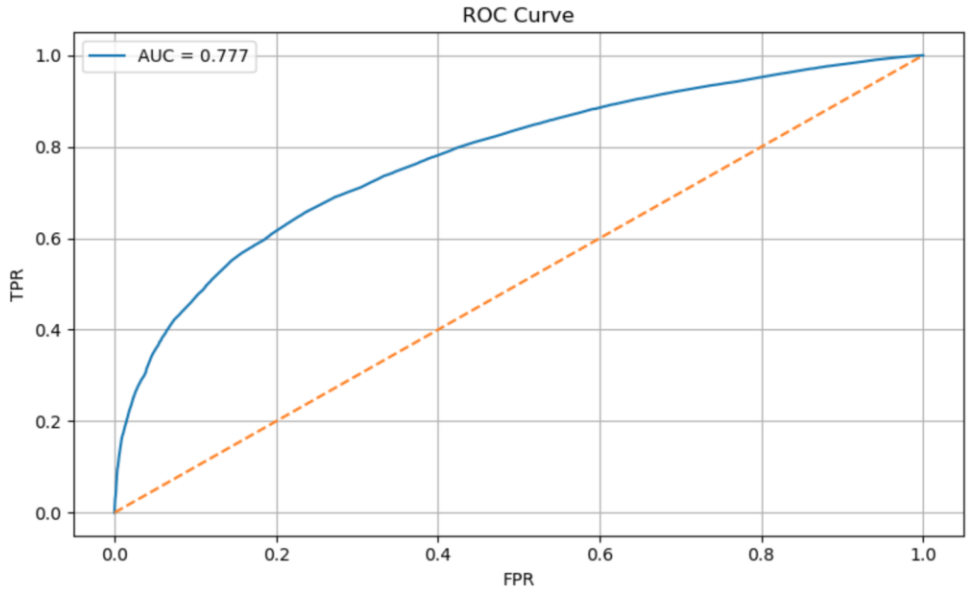
- **Metrics used:** Precision, Recall, F1-score.
- **Precision:** How often predicted labels are correct.
- **Recall:** How well the model identifies actual cases.
- **Performance on Diabetes classification:**
- **Diabetes:** Precision 54%, Recall 2%, F1-score 4% → model struggles to detect positive cases likely due to class imbalance.
- **No Diabetes:** Precision 86%, Recall 100%,F1-score 92% → model accurately identifies negative cases.

Report – Enhanced Multi-Layer Perceptron



- Displays true vs predicted labels for each class.
- Highlights **model errors**: false positives and false negatives.
- **Diabetes:** 0.39% correctly classified, 0.31%missed (false negatives).
- **No Diabetes:** 85.2% correctly classified,14% misclassified (false positives).
- Helps identify which classes the model struggles with, which in this case is shown that it struggles with predicting the 'yes' category.

Report – Enhanced Multi-Layer Perceptron



- **ROC Curve:** Plots True Positive Rate vs False Positive Rate across thresholds.
- **AUC = 0.77:** Indicates moderately good ability to distinguish Diabetes vs non-Diabetes.
- **Insights:**
- Model detects many non-Diabetes cases accurately.
- Model struggles with Diabetes cases →low recall / false negatives.
- **Implication:** Threshold adjustment or additional data could improve detection of positive cases.

Report – Comparison of All Models

	model	accuracy	precision	recall	f1	roc_auc
3	Enhanced-MLP	0.856701	0.858843	0.996324	0.922489	0.777017
2	Simple-MLP	0.855888	0.855961	0.999877	0.922339	0.776362
0	LogReg	0.855521	0.862199	0.989308	0.921391	0.773765
1	Linear-SVM	0.856028	0.856380	0.999387	0.922374	0.773407

- **Overall performance:** All models show similar metrics overall, but class-specific differences are important.
- **Recall:**
- Logistic Regression: 98.9% → detects most negative cases.
- Linear SVM: 0% for "Yes Diabetes" → fails to detect any true positive cases.
- Simple MLP: 0% for "Yes Diabetes" → fails to detect any true positive cases.
- Enhanced MLP: 99.6% → detects most negative cases.
- **Precision & F1-score:** Nearly identical overall but misleading for SVM and Simple-MLP due to the 0% recall on the positive class.
- **ROC-AUC:** Very similar across all → overall separation of classes looks okay but doesn't reflect the failure to detect positive cases in SVM or Simple-MLP.
- **Recommendation / Next Steps:**
- Enhanced-MLP is preferable because it identifies positive cases and indicates a slight marginally better discrimination between classes .
- SVM or Simple-MLP with 0% recall is not acceptable for detecting Diabetes.
- Consider threshold tuning, class balancing, or feature engineering to improve detection if you revisit SVM or Simple-MLP.

Act

- Our goal was to determine how demographic and lifestyle factors could be used to enhance diagnostic accuracy and early detection in adults.
- During the EDA, there were some things that were noticeable. The target variable 'Diabetes' had a class imbalance where the 'yes' category was underrepresented. In the Correlation Heatmap, we noticed that as an individual tends to indulge in smoking heavily, they also tend to indulge in consuming alcohol heavily which also increases with age. The histogram of Race showed us, that most of the individuals in this survey were White, while other categories were underrepresented. The histogram of BMI also showed us that the 'underweight' category was underrepresented.
- Based on these patterns observed in the data, we moved forward with building and evaluating our machine learning models. Overall, the models performed nearly identically in terms of precision, F1 score, and ROC-AUC. However, recall revealed an important difference: two of the models (Linear SVM and Simple-MLP) predicted **0%** for the 'yes' category, making their performance misleading despite similar overall metrics. Because of this, the model that best aligned with our goals was the Enhanced Multi-Layer Perceptron (MLP). Although the differences were minimal, the selected model achieved the highest ROC-AUC, indicating a slightly stronger ability to distinguish between the positive and negative classes. While two of the four models produced 0% recall for one class, the selected model showed stable recall across both classes, making it the more dependable option.

Act cont.

- The Enhanced MLP performs best at identifying individuals who do not have Diabetes, meaning it reliably detects this group more than individuals who have Diabetes. This indicates that, based on the available data, the model can support early detection by correctly classifying individuals in this category. While overall precision, F1 score, and ROC-AUC were similar across models, differences in recall highlight which class the model predicts more accurately, helping us understand its diagnostic reliability.
- Comprehensively, the Enhanced MLP model performed slightly better than the other models. It achieved a higher recall for the positive cases and a slightly higher ROC-AUC score, indicating it was marginally more effective at distinguishing between classes. However, the recall remained low (<10%), highlighting that the model struggled to identify positive cases reliably. Additional limitations include the use of self-reported data, which may affect reliability, and class imbalance, which could have influenced model performance and generalizability.

References

- *CDC - 2024 BRFSS Survey Data and documentation*. (n.d.). https://www.cdc.gov/brfss/annual_data/annual_2024.html
- *FastStats*. (n.d.). Leading Causes of Death. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm