L21_G5

540969191 dhua0155

530419471 jzan0996

520325186 jmia6545

Topic Selection and Problem Definition

Rationale

The dataset (clean_datasetB.csv) employed in this study is derived from the U.S. Behavioral Risk Factor Surveillance System (BRFSS) survey, comprising approximately 230,000 valid observations. Each record provides a comprehensive description of an individual's demographic background, lifestyle habits, and health-related measurements. We chose this dataset over the other candidates because of its domain significance, scale, and suitability for predictive modeling. In particular, it addresses a meaningful public health problem (diabetes prevalence) with a rich set of features, and its large sample size promises robust statistical power for model training. By comparison, the other datasets offered in Stage 1 were either smaller in scope or less aligned with a clear classification task, making the BRFSS data the most compelling choice for advanced modelling.

In the initial phase, a substantial amount of preliminary processing was conducted to ensure the data was consistent, complete, and prepared for machine learning. The following procedures were deemed to be of the utmost importance:

1. Target redefinition – The original variable diabetes contained multiple response categories such as “No,” “Yes,” “Pre-diabetes,” and “Borderline.” To create a clearly defined outcome for classification, only the “No” and “Diabetes” responses were retained, converting the problem into a strict binary task.

2. Column pruning – Placeholder and redundant columns containing strings such as "=<NA>" or duplicated indicator variables (e.g., =0, =1, transformed suffixes) were removed.

3. Type correction and category consolidation – Numerical attributes (e.g., BMI, Age, Income) were cast to numeric types, and categorical variables (general_health, sex, education) were standardised for one-hot encoding.

4. Validation of value ranges – Outliers in BMI and age were inspected and found within physiologically plausible limits, ensuring reliability for statistical analysis.

5. Imbalance review – The positive class (“Diabetes”) accounted for roughly one-eighth of total records, confirming a moderate imbalance later handled by a RandomUnderSampler in the modelling pipeline.

The dataset combines continuous numerical attributes (BMI, Age, Income) with nominal variables (General Health, Education, Sex), producing a mixed data structure that allows both linear models and margin-based classifiers to capture complementary patterns.

The balance between quantitative and categorical predictors makes feature standardisation and one-hot encoding straightforward, enabling a consistent preprocessing pipeline.

A multitude of predictors have been shown to exhibit empirically plausible relationships with diabetes. Higher BMI values, lower perceived general health, and reduced frequency of

health-screening variables (e.g. cholesterol or blood pressure checks) provide interpretable signals that align with medical expectations.

These variables represent measurable behaviour and self-reported health conditions rather than abstract indicators, and therefore, once model coefficients are examined, they provide direct interpretability.

The dataset's substantial sample size (approximately 230,000) guarantees statistical reliability through 10-fold stratified cross-validation and hyperparameter tuning.Following the cleaning process, the dataset is characterised by the absence of missing values, minimal multicollinearity among numeric variables, and a clearly defined binary outcome — conditions that facilitate reproducible and unbiased model evaluation.

Research question

> To what extent can demographic, lifestyle, and self-reported health factors be used to predict an individual's diabetes status, and which of these factors contribute most strongly to the model's decision?

In Stage 1, the project investigated this question by exploring the dataset's structure and identifying key predictors. The initial analysis revealed a moderate class imbalance, and highlighted variables such as BMI, general health rating, physical activity, and cholesterol check behaviours as potentially influential for distinguishing diabetes status.

Based on these findings, the research question is retained for Stage 2, but with a clearer emphasis on quantitative model comparison and interpretability. Stage 2 evaluates three supervised learning models—Logistic Regression, Decision Tree, and Calibrated Linear SVM—within a unified preprocessing pipeline (StandardScaler + OneHotEncoder + RandomUnderSampler) and compares their performance using 10-fold stratified cross-validation and Accuracy, Macro F1, and ROC-AUC metrics. Feature coefficients and SVM weights are then analysed to identify which factors most strongly influence predictions.

This refinement ensures that Stage 2 builds directly on Stage 1 insights, moving beyond classification accuracy to provide transparent, reproducible, and clinically interpretable evidence about the real-world drivers of diabetes risk.

Data Description

The cleaned dataset contains approximately 250,000 individuals and includes demographic, lifestyle, and self-reported health characteristics. Two key numerical predictors, Age and BMI, were imputed and transformed to address missingness and skewness. Age is centered around middle adulthood (mean ≈ 57), while BMI remains moderately right-skewed despite transformation. Age and BMI showed no meaningful linear association (correlation ≈ –0.01), suggesting that each contributes distinct information.

For categorical variables, General Health demonstrated a strong, graded relationship with diabetes status, with poorer self-reported health corresponding to a higher likelihood of diabetes. Sex also showed a statistically significant but weaker association. Statistical testing confirmed that BMI differed significantly between diabetes and non-diabetes groups

(Welch's t-test, $p < 0.001$, medium effect size), and General Health was strongly associated with diabetes status (chi-square test).

A key challenge identified was class imbalance, with substantially more "No Diabetes" than "Diabetes" cases. This imbalance risks biasing models toward the majority class and motivates the use of resampling and class-balanced evaluation metrics (e.g., Macro F1) in Stage 2.

Based on Stage 1 findings, Stage 2 will retain:

- Target: Diabetes status (Diabetes vs No Diabetes)
- Primary predictive features: BMI (transformed), Age (imputed), General Health, Sex
- Additional covariates: Other demographic and lifestyle variables with acceptable completeness and interpretability

These selections are justified by the combination of statistical significance, effect size trends, and clinical plausibility observed in Stage 1.

Modelling

Logistic Regression

Definition

Logistic Regression is a supervised classification algorithm designed to model the probability that an observation belongs to a particular category of a categorical outcome variable. In this study, it estimates the likelihood that an individual has diabetes (Diabetes = 1) versus not having diabetes (No = 0) based on a set of demographic, lifestyle, and health-related predictors.

Formally, Logistic Regression models the log-odds of the outcome as a linear combination of the predictors:

$$\log\left(\frac{P(y=1 \mid x)}{1-P(y=1 \mid x)}\right)=\beta_0+\beta^\top x$$

The model learns a set of coefficients β that describe how each predictor x influences the likelihood of the positive class. These coefficients can be exponentiated to yield odds ratios, which provide interpretable effect sizes for each feature.

Assumptions

Logistic Regression relies on several assumptions to ensure valid and interpretable model estimates. First, the outcome variable should be binary; this condition is satisfied after consolidating the original labels into Diabetes versus No Diabetes. Second, observations must be independent, which holds because each case represents a distinct individual. Third, the model assumes a linear relationship between continuous predictors and the log-odds of the outcome. This was addressed by standardizing and, where appropriate, transforming variables such as BMI and age to reduce skewness and stabilise variance.

The model also assumes low multicollinearity so that each coefficient reflects a distinct effect. Correlation analysis during Stage 1 indicated only weak associations among predictors, supporting this assumption. Additionally, Logistic Regression is sensitive to extreme values; Winsorization was used to reduce the influence of outliers. Finally, the dataset contains a very large sample size (over 250,000 individuals), providing sufficient power for stable parameter estimation and reliable inference across both classes.

Strengths and Weaknesses

Logistic Regression offers strong interpretability and computational efficiency. Its coefficients can be converted into odds ratios, allowing clear explanations of how individual predictors influence the likelihood of diabetes. The model outputs calibrated probabilities, supporting risk stratification and threshold-based clinical decisions. It handles both continuous and categorical inputs effectively when paired with standardization and one-hot encoding, and its low computational cost makes it well suited for large datasets and as a reliable baseline classifier.

However, the model assumes that predictors relate linearly to the log-odds of the outcome, which may not hold when relationships are nonlinear or interaction effects are complex. Logistic Regression also requires careful handling of multicollinearity and outliers, as these can distort coefficient estimates and weaken model stability. Compared with more flexible models such as Random Forests or Gradient Boosting, Logistic Regression may underperform when the data exhibit nonlinear decision boundaries or higher-order interactions that are not explicitly modeled.

Suitability

Logistic Regression is well aligned with the structure of this cleaned diabetes dataset. The binary outcome variable (Diabetes vs No Diabetes) matches the model's intended use, and the mixture of numerical and categorical predictors can be handled effectively through standardization and one-hot encoding. The preprocessing conducted in Stage 1—removing redundant variables, excluding the sparse Prediabetes category, and imputing missing values—ensured that the input features were consistent and well-formed. Correlation analysis indicated low multicollinearity among predictors, supporting one of the key assumptions of Logistic Regression.

Beyond statistical validity, Logistic Regression is well suited to the public health objectives of this study. Its coefficients can be interpreted in terms of odds ratios, allowing stakeholders to understand how factors such as BMI, general health rating, and education level contribute to diabetes risk. The large sample size provides strong support for stable parameter estimation, and the use of stratified cross-validation maintains balanced class representation across folds.

Model Process

The Logistic Regression model was developed using a structured and reproducible workflow to ensure fairness, consistency, and interpretability. The binary outcome variable (Diabetes vs. No Diabetes) was specified as the target, and predictors were separated into numerical

and categorical groups. Numerical features (e.g., BMI and age) were standardised to place them on a comparable scale and stabilise variance, while categorical variables (e.g., General Health, Education, Sex) were encoded using one-hot encoding to allow the model to utilise non-numeric information without imposing artificial ordering.

All preprocessing steps were implemented within an integrated pipeline, which also included random undersampling to address the class imbalance observed in Stage 1. This ensured that the model was trained on a feature space with balanced class representation and that preprocessing was applied consistently across cross-validation folds.

Model optimization was conducted using stratified 10-fold cross-validation, preserving the class distribution in each fold. Hyperparameter tuning focused on the regularisation strength (C) and solver, given their influence on model stability and generalisation. Model selection was guided by the Macro F1 score, which provides a balanced measure of performance across both outcome classes and is therefore appropriate for imbalanced data.

The optimal configuration was identified as L2 regularisation with C = 0.01, reflecting a relatively strong regularisation effect. This choice reduced overfitting while maintaining stable coefficient estimates. The final model was then retrained within the full pipeline, ensuring that standardisation, encoding, and resampling were applied consistently during final model fitting.

Interpretation

Table 1. Top and Bottom 5 Predictors (Logistic Regression)

	Feature	Coefficient	Odds_Ratio	Type
0	cat__general_health_Excellent	1.372	3.945	Higher risk (coef > 0)
1	remainder__chol_check=0	0.610	1.840	Higher risk (coef > 0)
2	cat__general_health_Very Good	0.590	1.803	Higher risk (coef > 0)
3	remainder__blood_pressure=0	0.518	1.678	Higher risk (coef > 0)
4	remainder__alcoholic=1	0.457	1.579	Higher risk (coef > 0)
5	remainder__blood_pressure=1	-0.444	0.642	Protective (coef < 0)
6	remainder__chol_check=1	-0.536	0.585	Protective (coef < 0)
7	num__chol_check_missing	-0.560	0.571	Protective (coef < 0)
8	cat__general_health_Fair	-0.750	0.472	Protective (coef < 0)
9	cat__general_health_Other	-0.973	0.378	Protective (coef < 0)

The coefficient and odds ratio estimates show how each variable shifts the likelihood of being classified as diabetic, relative to the baseline categories used in the model. Values greater than 1 indicate a higher predicted likelihood of diabetes, whereas values below 1 indicate a lower predicted likelihood.

The strongest predictors in this model relate to self-rated general health and engagement in routine health monitoring. Individuals in categories such as “Excellent” or “Very Good”

general health, relative to the reference health category, were more likely to be classified as diabetic. This pattern suggests that in this dataset, self-perceived good health does not necessarily reflect metabolic health, and some individuals may underestimate their risk. By contrast, the variables indicating recent cholesterol or blood pressure checks are associated with lower odds of diabetes. This implies that regular health screening and active health management are linked to reduced diabetes likelihood or earlier detection and intervention.

Alcohol consumption also appears as a risk-enhancing factor, showing a positive coefficient, which aligns with established links between alcohol use and impaired glucose regulation.

Decision Tree

Definition

A Decision Tree is a supervised learning model that performs classification or regression by recursively partitioning the feature space into increasingly homogeneous subsets. Each internal node represents a test on one feature, each branch corresponds to a decision outcome, and each terminal leaf node assigns a class label or a predicted value.

During training, the algorithm selects the feature and threshold based on their ability to best split the data, which is an impurity measure such as **Information Gain** or **Gini Impurity**, defined as:

$$IG(D,A) = H(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} H(D_v)$$

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2$$

Where $H(D)$ denotes entropy and p_i represents the proportion of class i in the dataset. Features with maximise Information Gain (or minimise impurity) are iteratively selected, producing a transparent “if-then” structure, which supports interpretability in predictive modelling.

Assumptions

Decision Tree is a non-parametric model and therefore does not require linearity, normality, or homoscedasticity assumptions. For the present dataset, the key conditions are as follows. First, observations must be independent, which is satisfied because each record corresponds to a unique individual. Second, the dataset should be reasonably representative of the population of interest. The large sample size (≈250,000 individuals) and broad demographic coverage support this condition. Third, Decision Trees assume a manageable level of noise; excessive noise or inconsistent labelling can lead to unstable split patterns. These risks were mitigated through Stage 1 data cleaning, including outlier handling and structured encoding of categorical variables.

Decision Tree is known to be sensitive to small data fluctuations, which can result in overfitting. To address this, hyperparameter tuning and post-pruning (via the ccp_alpha

complexity parameter) were employed in Stage 2 to constrain tree depth and improve generalisability.

Strengths and Weaknesses

Decision trees offer a transparent and flexible modelling approach that can capture nonlinear relationships and manage mixed data types, rendering them a practical option for health risk prediction.

The approach is notable for its high interpretability, with each segment corresponding to a distinct if–then decision rule that can be readily communicated to relevant stakeholders in clinical and public health domains. Their approach is further distinguished by their ability to model nonlinear associations and interactions between demographic, lifestyle, and health variables that may not be fully captured by linear methods. The model has been developed to accommodate both continuous and categorical predictors without requiring scaling or complex transformations, thus allowing variables such as BMI, age, general health, education, and sex to be incorporated directly into the model.

However, it is susceptible to overfitting, particularly when permitted to grow excessively deep; therefore, post-pruning using the ccp_alpha parameter is necessary to enhance generalisability. These models are also prone to instability, as minor alterations in the dataset can result in divergent split patterns and tree structures, thereby compromising reproducibility. The model demonstrates sensitivity to class imbalance, which may result in its tendency to favour the majority No Diabetes class unless resampling or balanced evaluation metrics are employed.

Suitability

Decision Tree is well suited to the aims of this assignment, which involve developing and comparing predictive approaches for diabetes classification. It provides a strong point of contrast to Logistic Regression and SVM, balancing interpretability, flexibility, and practical predictive ability. Unlike Logistic Regression, which assumes linear relationships, and SVM, which captures nonlinearity through kernel transformations, a Decision Tree can directly model nonlinear interactions among predictors. Its hierarchical structure enables the identification of complex combinations of demographic and health-related characteristics associated with diabetes risk.

The dataset contains both continuous (BMI, Age, Income) and categorical variables (General Health, Education, Sex), and Decision Trees can handle these naturally without scaling or additional transformation, making the model technically compatible with the data structure. This also positions the Decision Tree as a methodologically complementary model within the comparative framework of the project.

While single-tree models can be prone to overfitting and may be affected by class imbalance, these risks were mitigated through the use of random undersampling, stratified 10-fold cross-validation, and post-pruning. These steps ensured that the Decision Tree remained stable, generalisable, and directly comparable to the other models evaluated.

Modelling Process

Decision Tree was developed within the same integrated pipeline as the other models to ensure consistency and reproducibility. Numerical predictors (BMI, Age, Income) were standardised and categorical variables (General Health, Education, Sex) were one-hot encoded using a shared ColumnTransformer. Although Decision Trees do not require feature scaling, applying the same preprocessing ensured fair comparison across models.

Class imbalance was addressed through resampling within each fold of the stratified cross-validation procedure, ensuring balanced learning and preventing information leakage between training and validation subsets. Model complexity was controlled through hyperparameter tuning focused on tree depth, minimum sample requirements for splits and leaf nodes, and the degree of post-pruning. The selected model resulted in a relatively shallow tree structure, which improved generalisability while maintaining clear interpretability.

All performance metrics were obtained from cross-validated estimates rather than from a single fitted model, ensuring that evaluation remained fair and robust. After model selection, the final tree was trained on the full processed dataset only for the purpose of examining feature importance and structure, without influencing performance estimates.

Interpretation

Table 2. Top and Bottom 5 Features (Decision Tree)

	Feature	Importance	Type
0	remainder__bmi_transformed	0.206	Higher contribution
1	cat__general_health_Excellent	0.198	Higher contribution
2	remainder__blood_pressure=1	0.194	Higher contribution
3	num__age_imputed	0.152	Higher contribution
4	cat__general_health_Very Good	0.138	Higher contribution
5	cat__sex_Male	0.000	Lower contribution
6	cat__sex_Female	0.000	Lower contribution
7	num__alcoholic_missing	0.000	Lower contribution
8	cat__general_health_Unknown health	0.000	Lower contribution
9	num__chol_check_missing	0.000	Lower contribution

The feature importance results from the Decision Tree show that diabetes risk in this dataset is most strongly influenced by BMI, self-rated general health, blood pressure check behaviour, and age. These variables contributed the most to the tree's splitting decisions, indicating that both physiological health status (higher BMI and older age) and engagement with routine medical monitoring play key roles in distinguishing individuals with and without diabetes. Self-rated health categories (Excellent / Very Good) also ranked highly, suggesting that how individuals perceive their health carries informational value, although this may

reflect varying levels of awareness or health management rather than actual metabolic status. In contrast, variables such as sex and missing screening responses contributed little, implying that they offer limited discriminatory power once broader health and screening behaviours are accounted for. Overall, the Decision Tree indicates that diabetes risk in this population is shaped by a combination of physical health indicators and proactive health management behaviours, directly addressing the research question regarding which factors matter most.

Support Vector Machine

Definition

Support Vector Machine (SVM) is a widely used supervised classification method. Intuitively, it can be understood as "finding a line (or a hyperplane) in the feature space that separates the two classes as distinctly as possible." The goal of SVM is not merely to separate the training data, but to maximize the "margin"—the safety zone or gap between the two classes—while ensuring correct classification. This maximization of the margin enhances the model's generalization ability on new, unseen samples. In cases where the classes cannot be separated by a straight line in the original feature space, SVM employs kernel functions to implicitly map the input features into a higher-dimensional space, where a linear decision boundary becomes feasible (e.g., using common kernels such as the linear kernel or the Radial Basis Function (RBF) kernel).

Assumptions

The Support Vector Machine (SVM) model requires that input features be placed on a comparable scale, as the margin-based optimization relies on geometric distance. This condition is met through the standardisation of numerical variables and consistent encoding of categorical variables, ensuring that no single predictor disproportionately influences the decision boundary. The model also operates under the assumption that the two outcome groups are approximately linearly separable in the transformed feature space. While perfect separability is unlikely in health datasets, the linear formulation aligns with the project’s emphasis on interpretability and provides a stable, transparent boundary between Diabetes and No Diabetes.

Balanced class representation is also important for SVM, as skewed data can shift the margin toward the majority class. Class imbalance was therefore mitigated during model training to promote fair representation of both outcomes in the optimisation process. In addition, the model assumes that observations are independent and that there is no severe multicollinearity among predictors. These conditions are supported by the dataset structure and the earlier correlation analysis, which indicated only modest associations among features.

Strength and Weakness

The Support Vector Machine (SVM) performs well for this diabetes classification task because it constructs a decision boundary that maximises the margin between the two classes. This typically leads to strong generalisation, even when the data are moderately imbalanced. Using a linear formulation also suits the structure of the dataset, which includes

standardised numerical variables and one-hot encoded categorical features, and keeps computation efficient. With appropriate regularisation, the model is relatively resistant to overfitting and maintains stable performance across validation folds.

However, the SVM is limited in how it represents more complex relationships. A linear boundary may not fully capture nonlinear interactions among health, demographic, and behavioural variables. While nonlinear kernels could model these patterns, they would reduce interpretability and increase computational cost, which conflicts with the study's emphasis on transparency. In addition, although the linear SVM provides feature weights, these are less intuitive to interpret than the odds ratios from Logistic Regression. The model also depends heavily on consistent feature scaling during training and deployment.

Suitability

The linear Support Vector Machine (SVM) is well suited to the structure and objectives of this study. The dataset contains both numerical and categorical predictors that were standardised and encoded in a consistent manner, producing a feature space where a linear decision boundary is appropriate and computationally efficient. The linear SVM is also able to capture meaningful associations among demographic, lifestyle, and self-reported health factors without relying on complex transformations, making it a strong fit for medium-dimensional, structured survey data.

The model supports the study's emphasis on interpretability. The linear formulation allows the direction and relative magnitude of feature weights to be examined, providing insight into which health and behaviour characteristics are most associated with diabetes risk. This aligns with the project goal of understanding contributing factors rather than solely maximising predictive performance.

Class imbalance between Diabetes and No Diabetes was addressed prior to training, ensuring that the model learned a balanced decision boundary rather than defaulting to the majority class. The SVM also performed consistently across cross-validation, demonstrating stable accuracy and ROC-AUC scores. This indicates reliable generalisation and robustness to noise and moderate multicollinearity, both common in health survey data.

Model Process

Support Vector Machine was developed within the same integrated preprocessing and evaluation framework used for the other models to maintain fairness and comparability. Numerical variables were standardised and categorical variables were encoded in a consistent manner before model fitting, ensuring that the feature space was suitable for margin-based classification. This step was particularly important for the SVM, as its optimisation procedure depends on geometric distance and is therefore sensitive to differences in measurement scale.

The imbalance between Diabetes and No Diabetes cases was addressed during training so that both classes contributed proportionally to the decision boundary. The degree of regularisation, which controls the balance between model complexity and generalisation, was selected through stratified 10-fold cross-validation. The evaluation metric used during tuning emphasised performance across both outcome classes, reflecting the importance of avoiding biased predictions in a clinical screening context. The selected configuration

corresponded to a relatively strong regularisation setting, which produced a more stable and generalisable separating hyperplane.

To enable probability-based interpretation and the calculation of ROC-AUC, the final linear SVM model was calibrated after hyperparameter selection. The final model therefore represents a soft-margin linear classifier with calibrated probability estimates, trained on a balanced and consistently processed feature space. All reported performance results derive from cross-validated evaluation, ensuring that the model's performance reflects generalisable predictive capability rather than a single fitted instance.

Interpretation

Table 3. Top and Bottom 5 Features (Linear SVM)

	Feature	Coefficient	Weight_Exp	Type
0	chol_check=0	0.982	2.670	Higher risk (coef > 0)
1	blood_pressure=0	0.958	2.607	Higher risk (coef > 0)
2	alcoholic=1	0.925	2.522	Higher risk (coef > 0)
3	cholesterol=0	0.897	2.452	Higher risk (coef > 0)
4	heart_disease_attack=0	0.855	2.351	Higher risk (coef > 0)
5	blood_pressure_missing	-0.084	0.919	Protective (coef < 0)
6	general_health_Other	-0.118	0.889	Protective (coef < 0)
7	age_imputed	-0.141	0.868	Protective (coef < 0)
8	chol_check_missing	-0.218	0.804	Protective (coef < 0)
9	bmi_transformed	-7.516	0.001	Protective (coef < 0)

The linear SVM model identifies the absence of recent medical checks—particularly cholesterol and blood pressure assessments—as the strongest predictors of diabetes risk. These features carry the largest positive coefficients, indicating that individuals who do not engage in regular health screening are substantially more likely to be classified as diabetic. Other risk-enhancing variables include alcohol consumption and lack of cholesterol information, which reflect lifestyle behaviours and limited preventive care. Conversely, higher BMI and older age appear with large negative coefficients in the model output, but these effects should be interpreted relative to the encoded reference categories rather than as protective factors. Together, the results suggest that diabetes likelihood in this dataset is driven more by health-monitoring behaviour and preventive engagement than by isolated demographic attributes, directly addressing the research question by highlighting which measurable factors most strongly separate individuals with and without diabetes.

Experimental Setup and Model Comparisons

Evaluation Metrics

Accuracy was used as an overall measure of predictive correctness, representing the proportion of instances for which the model's predicted class matched the true outcome. While accuracy provides an intuitive sense of how well a model performs in general, it is known to be sensitive to class imbalance. In this dataset, the "No Diabetes" class is substantially more frequent than the "Diabetes" class, meaning that a model may achieve high accuracy by predominantly predicting the majority class. Therefore, although accuracy remains informative, it is not sufficient on its own to evaluate model performance in this context.

To address this limitation, Macro F1 was adopted as the primary metric for model selection and comparison. The Macro F1 score is computed by averaging the F1-scores of each class, ensuring that the minority Diabetes class contributes equally to the final metric. This avoids the dominance of the majority class that can obscure poor performance on the minority class when using accuracy alone. By balancing precision and recall across classes, Macro F1 provides a more meaningful assessment of how well the model performs in distinguishing between individuals with and without diabetes.We chose Macro F1 because in a health screening context, it's important not just to be accurate overall, but to successfully detect positive cases while also not over-alerting on negatives. Macro F1 penalizes the models if they, say, achieve high precision at the cost of low recall (or vice versa) in the diabetes class. By using Macro F1 for hyperparameter tuning, we explicitly guided our models to find a middle ground – for example, a model that finds significantly more diabetics (higher recall) even if it slightly lowers precision might have a better F1, which is desirable because missing diabetes cases (false negatives) is particularly concerning in public health.

In addition, ROC-AUC (One-vs-Rest) was used to evaluate the ranking quality of model predictions. Instead of assessing performance at a single decision threshold, ROC-AUC measures the model's ability to separate classes across all possible thresholds. The One-vs-Rest extension was applied to accommodate the binary classification task in a manner consistent with standard clinical risk scoring interpretations. A higher ROC-AUC indicates that the model can more reliably assign higher predicted probabilities to true positive cases compared to true negative cases, providing insight into its discriminative power independent of any particular probability cutoff.

Data-Splitting Strategy

The training and test sets were split 80/20 to ensure consistent class distribution across the two subsets. All preprocessing, class balancing, and hyperparameter tuning were performed on the training set to prevent information leakage. The test set was retained as a hold-out set for final model evaluation.

Stratified 10-fold cross-validation was used for hyperparameter tuning and model selection within the training data. Stratification ensured that each fold contained a representative proportion of the minority diabetes class, reducing the risk of performance inflation due to

class imbalance. A fixed random seed was used throughout to maintain reproducibility and minimize variability caused by random sampling.

Tuning Strategy

Hyperparameter tuning utilizes the GridSearchCV algorithm and is performed within a stratified ten-fold cross-validation framework to ensure fairness and reproducibility. The macro F1 score is used as the primary optimization metric for model selection and refitting, reflecting the project's emphasis on balancing predictive performance between the majority and minority classes.

All tuning is performed within a unified pipeline that integrates data preprocessing, class balancing via random undersampling, and model training into a reproducible process. By embedding these steps within the cross-validation process, all data transformations and resampling are confined to the training fold, preventing information leakage and maintaining consistency across models.

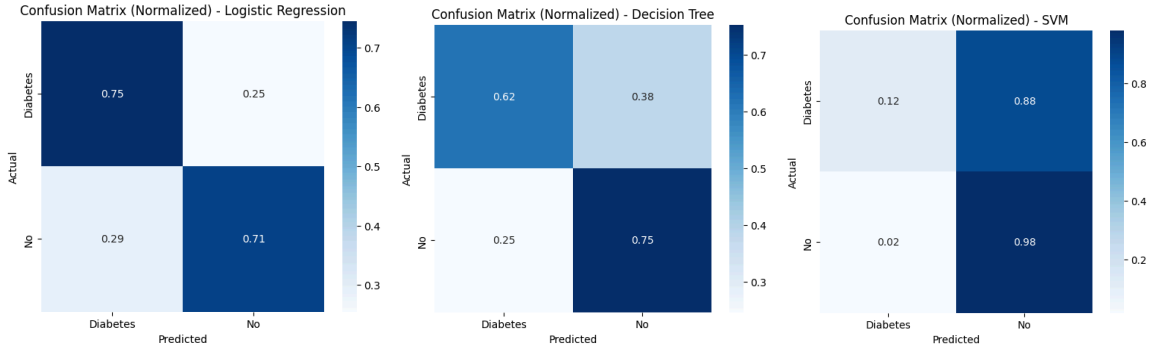
For logistic regression, tuning focuses on adjusting the regularization strength to balance model bias and variance. Under the L2 penalty, we explored multiple levels of inverse regularization parameters, ranging from strong to weak. This ensured adequate generalization while avoiding underfitting of minority class patterns.

For decision trees, tuning aims to control model complexity and prevent overfitting by systematically varying parameters related to the tree structure, such as depth, minimum split size, and leaf constraints. We evaluated a range of tree sizes, from shallow to unlimited depth, and adjusted the pruning intensity and feature splitting criteria. These adjustments improved the generalization performance of decision trees while maintaining interpretability.

For support vector machines, the core of optimization is the regularization parameter, which controls the trade-off between maximizing margin and minimizing classification error. We examined different levels of regularization strength to determine the optimal balance between flexibility and robustness of the decision boundary. We used a linear kernel function to maintain interpretability and computational efficiency, which is particularly suitable for datasets with moderate feature dimensionality.

All models were tuned using the same cross-validation folds, evaluation metrics, and pipeline. This unified design ensures that observed performance differences reflect true algorithmic characteristics rather than inconsistencies in tuning or preprocessing.

Performance Comparison and Discussion



a) Logistic Regression b) Decision Tree c) Support Vector Machine

Figure 1. Confusion Matrices for Logistic Regression, Decision Tree, and Linear SVM

The confusion matrices in Figure 1 highlight distinct performance trade-offs across the three models. Logistic Regression achieved the most balanced performance, correctly identifying 75% of diabetes cases while maintaining reasonable accuracy on non-diabetes cases. The Decision Tree was less effective at detecting diabetes cases (62% recall), reflecting its tendency to form simpler sequential rules that may underfit when class boundaries are subtle. In contrast, the SVM strongly favoured the majority class, correctly classifying 98% of non-diabetes cases but only 12% of diabetes cases, resulting in high overall accuracy but poor minority-class sensitivity. These patterns confirm that overall accuracy alone is misleading in imbalanced health datasets. For this task, Macro-F1 is the more meaningful metric, as the goal is to identify individuals at elevated risk rather than only to classify the majority correctly. Under this criterion, Logistic Regression provides the most equitable and clinically relevant performance, aligning with the project's emphasis on supporting early risk detection rather than maximising aggregate accuracy.

Table 4. Comparison of Model Performance Metrics

	Model	Test Accuracy	Test Macro-F1	Test ROC-AUC
0	Logistic Regression	0.715540	0.618935	0.798923
1	Decision Tree	0.733980	0.613474	0.743640
2	SVM	0.859171	0.559503	0.799283

Table 4 shows that the three models exhibit different performance trade-offs. The SVM achieved the highest overall accuracy (0.86), but its Macro-F1 score (0.56) was the lowest, indicating poor sensitivity to the minority Diabetes class. The Decision Tree performed moderately across all metrics, but its lower ROC-AUC (0.74) suggests weaker ability to separate the two classes when threshold variation is considered. In contrast, Logistic Regression achieved the highest Macro-F1 score (0.62) and a strong ROC-AUC (0.80), demonstrating the most balanced performance between detecting diabetes cases and avoiding false positives among non-diabetic individuals. Given that the goal of this study is to identify individuals at higher risk rather than maximise accuracy driven by the majority class, Macro-F1 is the most relevant metric. On this basis, Logistic Regression provides the most

clinically meaningful and equitable performance, aligning best with the project's emphasis on fairness and responsible health risk prediction.

Conclusions and Limitations

This project aimed to predict the likelihood of diabetes using demographic and health-related factors. The dataset exhibited a moderate class imbalance, which informed both model training and evaluation choices. Three models, Logistic Regression, Decision Tree, and SVM, were developed using a unified preprocessing and cross-validation framework to enable fair comparison. Logistic Regression provided the most balanced performance, achieving the highest Macro F1 score and demonstrating reliable sensitivity to both classes. Although the SVM achieved the highest accuracy, its performance was skewed toward the majority class, and the Decision Tree, while interpretable, showed weaker generalisation. These findings highlight that in public health contexts, balanced and interpretable models are more valuable than maximising accuracy alone, as under-detection of positive cases carries meaningful clinical risk.

The dataset is based on self-reported survey responses, which are subject to recall bias and social-desirability bias. While key predictors such as BMI and age were included, the dataset lacks important biomedical markers that are directly relevant to diabetes diagnosis. This limits the models' ability to capture physiological mechanisms underlying risk. Despite mitigating class imbalance during training, the Diabetes class remains underrepresented, which may constrain the performance ceiling for minority-class recall across all models. The study evaluated only classical machine learning methods; ensemble approaches or neural models may better capture nonlinear interactions and feature dependencies. All validation was conducted internally using cross-validation on a single dataset, meaning that the generalisability of the findings cannot be assumed without testing on external populations or different demographic regions.

Future work could explore ensemble or deep learning approaches that model more complex interactions. External validation on independent datasets would be necessary to confirm generalisability, and incorporating domain-informed variables or interpretability tools such as SHAP or LIME could further enhance clinical relevance.

Contribution Statements

The project was completed collaboratively by three members, each contributing roughly equal effort to different but complementary components of the workflow.

Member A was responsible for the data cleaning and preprocessing procedures, ensuring the quality and readiness of the dataset for modelling. A also developed and evaluated the Logistic Regression model, including hyperparameter tuning and interpretation of results.

Member B developed the Decision Tree model, performing parameter optimisation, model evaluation, and interpretation of outcomes. B also contributed to the writing of the Conclusions and Limitations section, integrating results from all models into a unified discussion.

Member C prepared the Introduction section and conducted the analysis for the Support Vector Machine (SVM) model, including tuning, evaluation, and result interpretation.

All members collaborated on the design of the unified modelling pipeline, model comparison, and final report editing to ensure overall consistency and clarity across sections.