Exploring the weightage of correlates of diabetes prediction using Machine learning

Aarush Raheja

Abstract

There is an increase in the prevalence of Type 1 and Type 2 Diabetes/diabetes around the world. This alarming trend highlights the need for prediction tools that can help identify risk factors associated with these chronic diseases, so that interventions can be implemented at a much earlier stage in their development. This study investigates two distinct datasets drawn from Kaggle, focusing on clinical and lifestyle factors, respectively. We constructed machine learning models, such as **Logistic Regression L1 penalty, Logistic Regression L2 penalty, Random forest and a dummy classifier model** to contrast basic accuracy, to first make predictions of the likelihood of diabetes occurrence given various factors. We demonstrate that the models have high predictive accuracy, with the logistic regression L2 penalty model achieving 95.297% accuracy, the logistic regression L1 penalty model achieving 96% accuracy, and the random forest model achieving 97% accuracy. However, the key contribution of this study is to provide interpretation of these models to determine the most important drivers of the models' predictions. We find that even though well-known factors such as HBA1C level, hypertension, and heart disease have high associations with diabetes, **factors such as mental health even though below BMI and HBA1C level do have a moderate predictive power.**

Introduction

Diabetes is a disease that affects 500 million people worldwide and constitutes a major public health crisis. To predict individual propensity toward developing diabetes, traditional methods include HBA1c testing and fasting glucose levels. However, these methods are time consuming, invasive, resource intensive and often impractical for large scale screening (Perez & Molano, n.d., 2025) . Identifying the key factors contributing to diabetes risk at the population level is important, as this will enable policy makers and public health experts to determine the most effective interventions.

To address this challenge, machine learning methods can be used to identify specific factors that determine the incidence of diabetes. Many ML models have shown promising results in the prediction of diabetes, such as studies by (Lugnar, n.d., 2023) health administration data models (Ravaut, 2021, #) , and detailed lifestyle-only modeling (Qin, n.d., #) . However, these studies only focus on a minimal number of variables and often fail to include lifestyle factors – such as smoking, physical inactivity, alcohol use, and dietary patterns – that have a known relationship to diabetes risk. In addition, these studies commonly focus on optimizing a performance metric but provide little information about the most important factors driving the model's performance (Jyoti, n.d., 2020). For example, the study by (Noh, n.d., #) explores factors such as BMI, HBA1C but does not explore basic lifestyle factors such as the

physical activity of the individual. The study by (Farida, n.d., #), while incorporating similar factors does not provide insight into which factors are most determinative of the model's predictive power. Additionally, there are several examples such as (Zhou, 2023, #), (Negi, n.d., #) and (Alhussan, n.d., #) that encompass less than 10 features while predicting diabetes thus not being able to access the full range of factors that affect it.

To address these limitations, this study analyzes a broad range of lifestyle factors and also provides interpretation of each of the models used to determine the most important variables. Here, we test a wide variety of common machine learning frameworks,
including logistic regression(with both L1 and L2 regularization) and random forests on two distinct datasets. The first one focuses on clinical factors, whereas the second one includes primarily lifestyle factors to expand the range of factors of the study. Rather than aiming for optimization, this study utilizes the factors to answer a more nuanced question of which range of factors are better predictive of whether the patient has diabetes or not, which gives insight into which factors are most influential with respect to diabetes prediction.

The results show that the random forest model achieved the highest accuracy - 97% on the first dataset. While the logistic regression with L1 regularization did not perform as well it still offers interpretable insights into the key contributing factors in both datasets. In addition to well known factors such as HBAIc level and BMI, factors such as smoking, alcohol consumption and mental health also emerged as important drivers. These findings open a pathway to diabetes prediction in the future, ensuring that diabetes prediction in the future is based on a vast range of factors that can be accounted for and changed easily.

Motivation

The global prevalence of diabetes has risen sharply over recent decades. Even though as individuals we are aware of the factors that affect us and cause diabetes the global rate still has yet to take a hit. With blood glucose levels and BMI being at the top of our list there are several other factors that aren't considered. Thus this study aims to delve deeper into these factors and identify how vital they are when it comes to predicting whether a patient has diabetes or not through numerous machine learning models. Alongside this as a personal motivation, both my parents are diabetic as well putting me as an individual also at high risk, knowing this research allows me to understand what other factors there are that can push me towards developing this disease whether it be smoking, alcohol consumption, mental health,etc.

Methods

Datasets for this study were obtained from publicly available sources. The first dataset contains factors such as: gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level and blood_glucose_level to predict a final outcome of whether the patient has diabetes or not. The second dataset contains other factors such as : HighBP(Blood Pressure), HighChol(Cholesterol),

CholCheck(Cholesterol Check), BMI (Body mass Index), Smoker, Stroke, HeartDiseaseorAttack, PhysActivity(Physical Activity), Fruits, Veggies, HvyAlcoholConsump(Heavy Alcohol Consumption, AnyHealthcare, NoDocbcCost (Was there a time in the past 12 months when you needed to see a doctor but could not cause of cost? 0 = no 1 = yes), GenHlth(General Health), MentHlth(Mental Health), PhysHlth(Physical Health) , DiffWalk(Difficulty to walk), Sex, Age, Education, Income.

Notably many of dataset 2's variables were binary(eg: - smoker vs non-smoker, Heavy alcohol consumption vs not heavy alcohol consumption). These coarse categories could reduce the predictive abilities of models to capture variation within behaviour, thus limiting the predictive performance of this dataset.

Data Preprocessing is an integral part of any machine learning model.During the course of this research after reviewing the entire data set thoroughly little data was removed; data values that were missing were marked with "no info".

No features and no samples were removed as the number of instances of this occurring were minimal and had a negligible effect on the machine learning model. However, encoding these values as "no info" in a few places may have introduced bias by treating absence of data as a valid category which could introduce noise since the model could interpret the absence of data as predictive information. While this,in turn, did enable the preservation of most of the dataset for training, more rigorous imputation techniques (eg: - mean/mode imputation or k nearest neighbors) could be explored in the future to assess whether they improve performance and reduce noise.

Categorical variables were encoded numerically to allow machine learning model process data Variables that held a string value were encoded into different integer categories, a process which is described in the latter part of this paper. The Target variable for this research was whether the patient had diabetes – for the first dataset healthy patients were encoded with "0" while as diabetic patients were encoded with a "1" . Other variables that were encoded were gender where "0" represented female, "1" represented male and "2" represented another category. Furthermore, smoking history was also encoded where "0" represented never, "1" represented no info, "2" represented current, "3" represented former, "4" represented ever and "5" represented not current. For the second dataset however there was a change. Even though the target variable was the same, diabetes, there was an additional category for pre - diabetes which was removed to allow the target variable to be classified into a binary outcome, allowing comparison with data set 1 which only had this binary target variable. "0" was encoded as non - diabetic and "1" was encoded as diabetic. For this particular model the split between training and testing data was 70:30.

Alongside this a simple cross - validation was used to evaluate model performance. Each model was trained on 4 folds and tested on the remaining fold, rotating across all folds. This approach provided an estimate of mean accuracy across all folds. However, stratification by class labels was not applied meaning that folds may have contained imbalance distributions of diabetic vs non - diabetic cases.

Also, other evaluation metrics such precision, recall, F1 score and ROC - AUC were used to provide a whole assessment of the models' performance. Precision quantified how many patients predicted as diabetic were actually diabetic, while recall measured the proportion of true diabetic cases correctly identified. F1 score balanced these two metrics into a single measure of performance while the ROC - AUC allowed to understand the models' discrimination ability across thresholds.

To assess the variability and assess the robustness of the model performance, we computed 95% confidence intervals for all evaluation metrics using bootstrap resampling. For each model we sampled with replacement from the test set over 1000 iterations and recalculated accuracy, precision, recall, F1 score and ROC-AUC. This procedure provided estimates of the stability of each metric, ensuring that reported performance was not due to random variation is a single train-test split. 2

Dataset 1 in this instance exhibited class imbalance, with a larger proportion of non - diabetic cases to diabetic cases. No balancing techniques such as SMOTE oversampling, undersampling or class weights were applied in this study. Instead, the models were evaluated using cross validation to mitigate overfitting, but future work should directly test balancing strategies to assess their impact on predictive stability.

To predict an outcome this study uses models like Decision tree, Random forest, Logistic regression and logistic regression with an L1 penalty to train and develop the model to a higher accuracy.

Logistic Regression : -
Logistic Regression model is a well - known and widespread model that makes a binary classification. It is somewhat similar to the linear model although the logistic regression model generates a probability value ranging between 0 and 1 on how likely the data point is to fit into the classification of a success or a failure. This probability is then compared to a threshold value, and based on the probability deciphered a final output is given whether the model predicts true or false.

Logistic Regression (L1 Penalty) : -
Logistic regression using an L1 penalty applies a penalty proportional to the absolute value of the coefficients. This, in turn, drives non predictive factors to zero, resulting in a sparse model. Such sparsity enables effective feature selection, allowing only the highest weighted correlates to affect the model's accuracy.

Logistic Regression (L2 Penalty) : -
Logistic regression using the L2 penalty applies a squared penalty to the coefficients although not to zero. Unlike L1 it does not perform feature selection but instead stabilises the model, while retaining all predictive features.

Random Forest : -
A random forest model is a combination of decision tree models that work on the concept of a series of questions that classify an input into categories. Each decision tree has several nodes, nodes that represent question making classifying the input further. Although this model is sometimes prone to

overfitting which can hamper the results. The model employed consisted of 100 trees ; Gini impurity as the splitting criterion, and no maximum depth restriction, meaning trees were allowed to grow until the leaves were pure. The choice of 100 trees was a reflection of prior successful application of random forest models in medical datasets. Additionally, allowing unrestricted depth minimized bias, though at risk of overfitting which was mitigated by cross - validation.

Dummy Classifier : -
The dummy classifier included as a baseline, highlighted dataset imbalance by achieving high accuracy(91.67%) while failing across recall and precision.

In current clinical practice, diabetes risk is frequently assessed using established scoring systems such as the ADA risk test, FINDRISC or Qdiabetes. These models rely on simple additive rules based on demographic and lifestyle factors, which makes them interpretable but also limited in their ability to capture complex non-linear relationships. Our study did not directly compare ML models to these clinical scores, instead we employed a dummy classifier as a baseline to establish the minimum threshold of predictive value. This approach allowed us to quantify the relative improvement achieved by different ML methods However future work should benchmark machine learning models against clinical scoring systems to provide a more direct assessment of their added value in practice. Prior studies have shown that ML approaches often outperform traditional scores in recall and ROC-AUC, suggesting their potential to complement and enhance clinical decision making.

Data Set 1 : -

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Figure 1 : A snippet of the first dataset showing it’s header

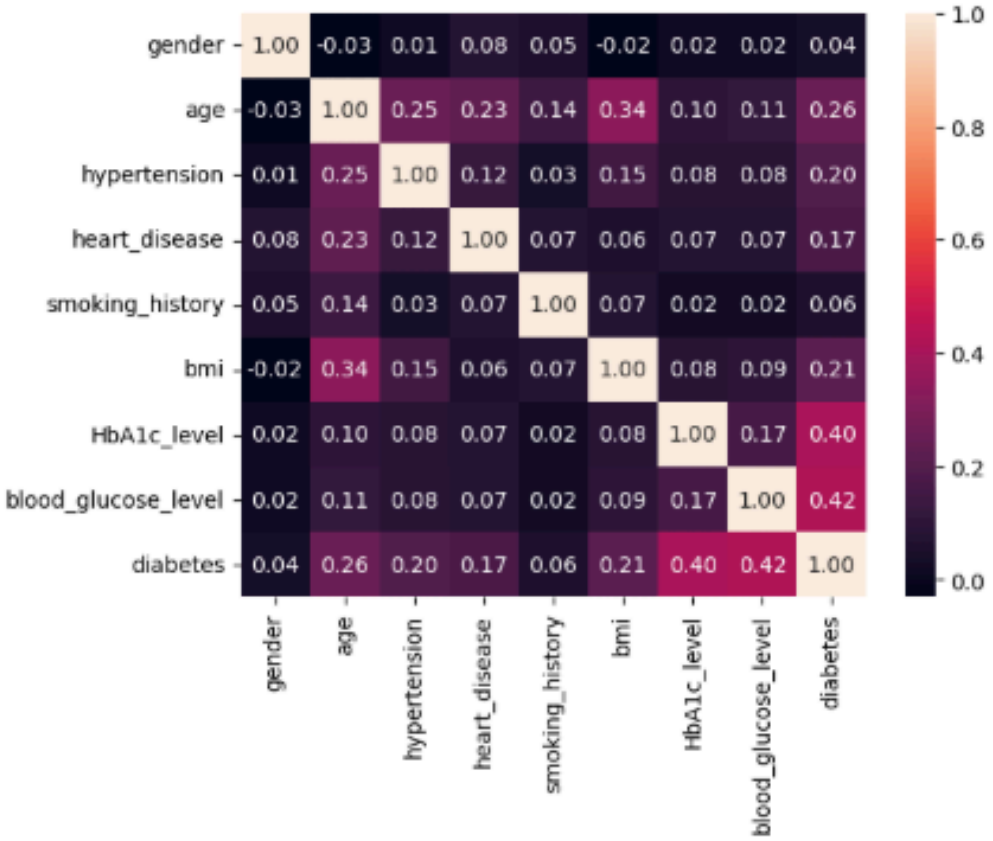

Figure 2: - Correlation between different factors affecting diabetes

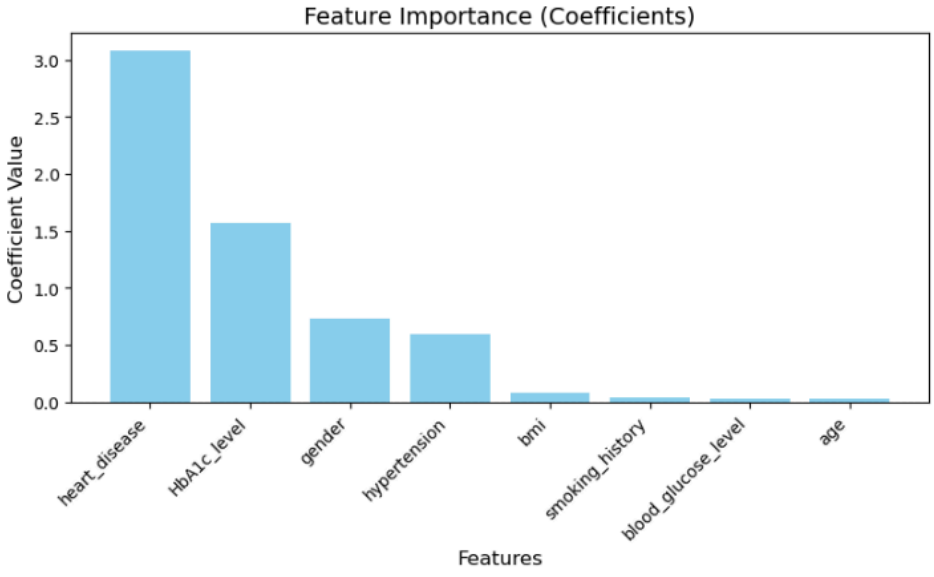

Figure 3: - Feature Importance from coefficients of a logistic regression model with the training method - L2 penalty

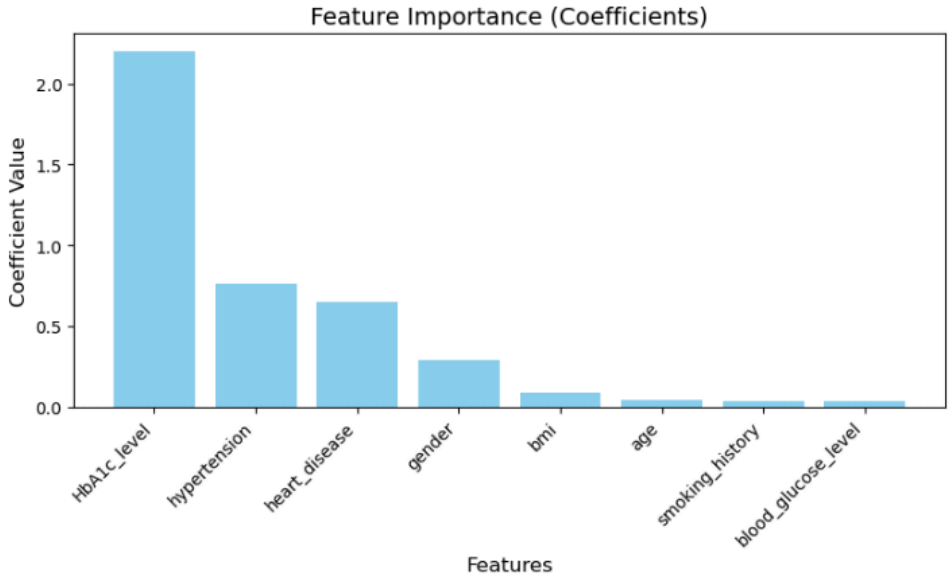

Figure 4 : - Feature Importance L1 penalty

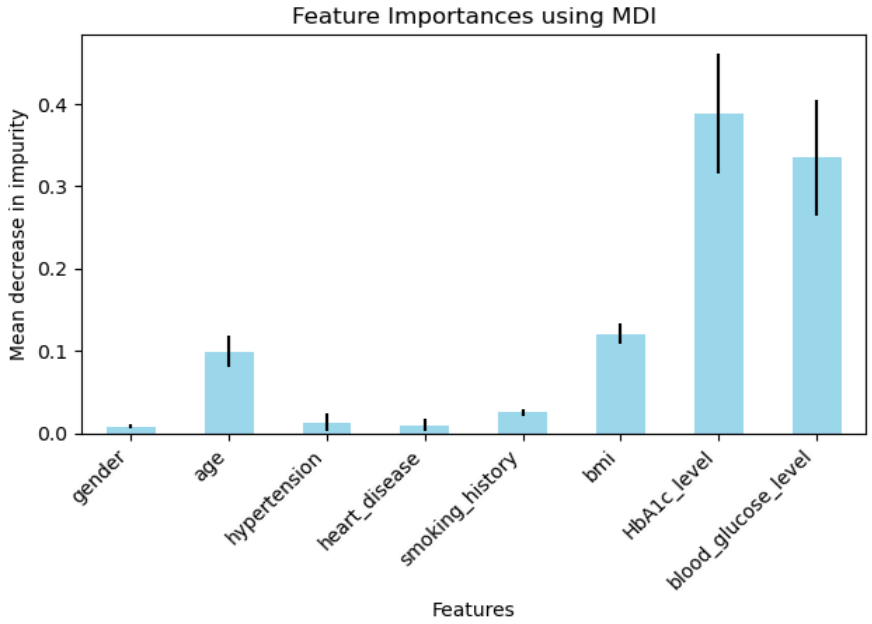

Figure 5 : - Feature Importance Random Forest MDI

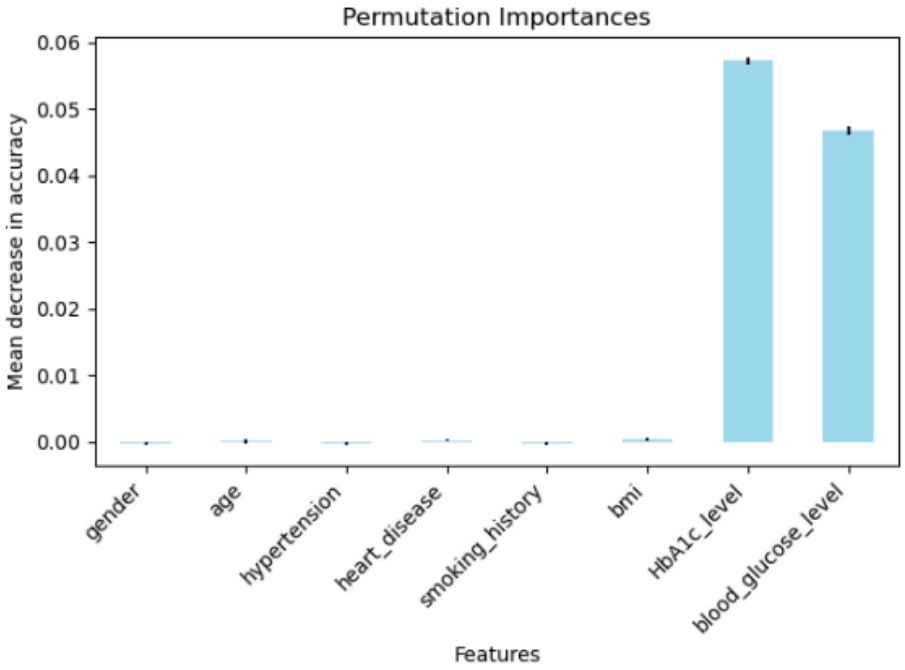

Figure 6 : - Feature Importance Random Forest permutation Importance

Results : -

These results should be interpreted cautiously however, as the class imbalance of the first dataset may have contributed to inflating the accuracy of these models. The dummy classifier achieving 91.667% highlights this imbalance.

Dataset 1

MODEL	Accuracy	Precision	Recall	F1 - score	ROC - AUC
Logistic Regression L1	96%	0.88	0.60	0.72	0.961
Logistic Regression L2	95.297%	0.74	0.48	0.58	0.902
Random Forest	97.003%	0.96	0.68	0.79	0.964
Dummy classifier	91.667%	0	0	0	0.5

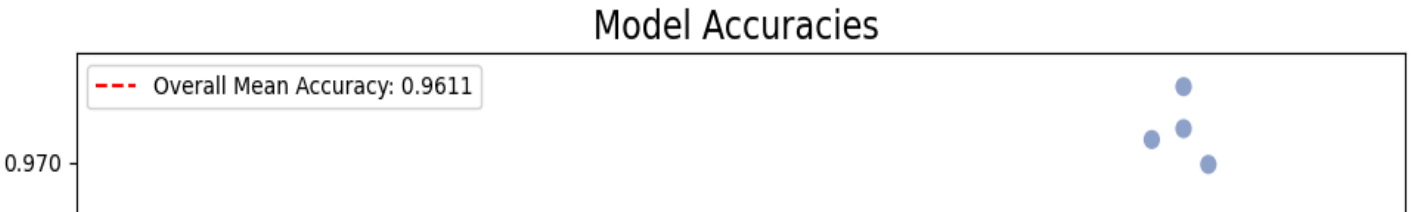

Figure 7 : - Model accuracies represented on a strip plot

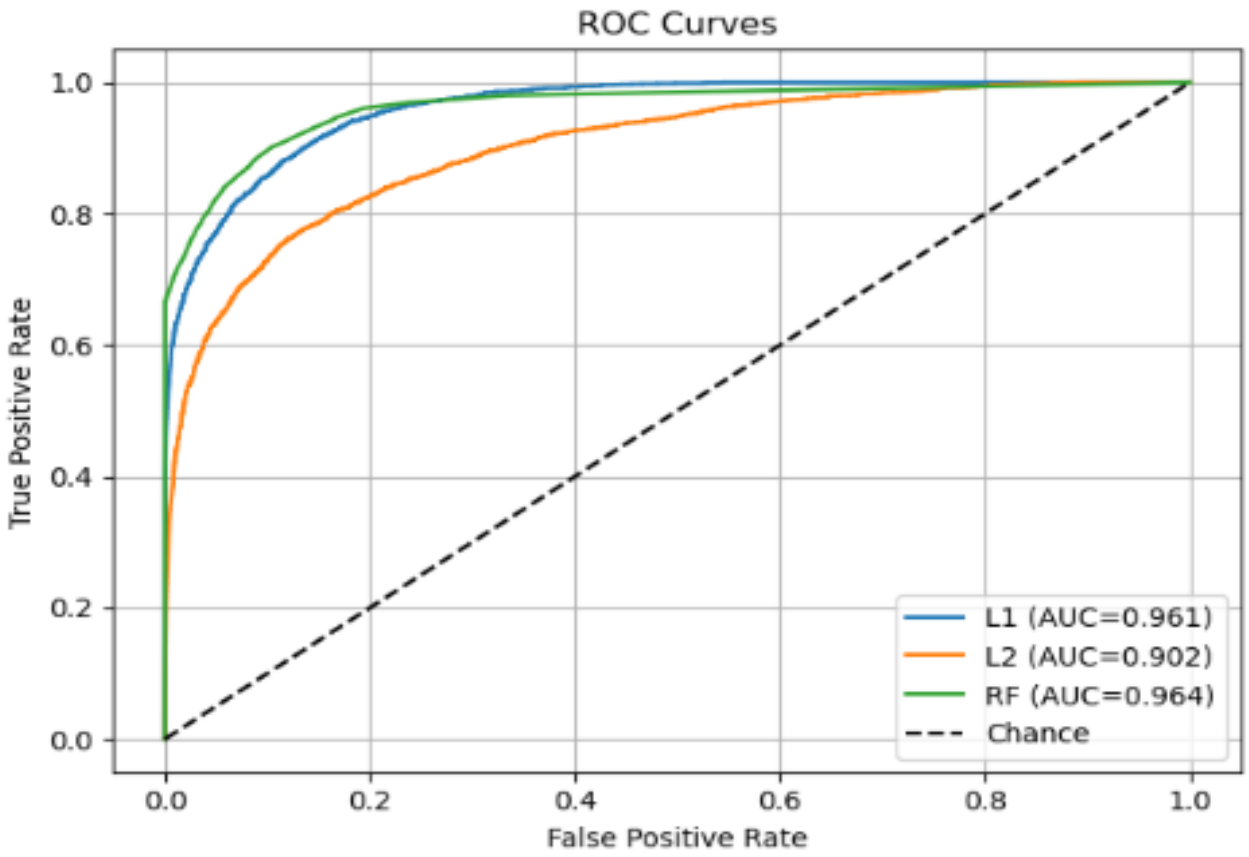

Figure 8 : - Representation of the ROC values for each model

Data Set 2 : -

[28]:

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	Me
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	

Figure 9 : - A snippet of the header of the dataset

Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0
0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0
0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0

Figure 10: - A continuation of the snippet of the header of the dataset

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	M
253675	0.0	1.0	1.0	1.0	45.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	
253676	2.0	1.0	1.0	1.0	18.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	4.0	
253677	0.0	0.0	0.0	1.0	28.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	
253678	0.0	1.0	0.0	1.0	23.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	
253679	2.0	1.0	1.0	1.0	25.0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	2.0	

5 rows × 22 columns

Figure 11 : - A snippet of the tail of the dataset

HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	5.0	0.0	1.0	5.0	6.0	7.0
0.0	0.0	0.0	...	1.0	0.0	4.0	0.0	0.0	1.0	0.0	11.0	2.0	4.0
0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	5.0	2.0
0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	0.0	0.0	1.0	7.0	5.0	1.0
1.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	9.0	6.0	2.0

Figure 12 : - A continuation of the snippet of the tail of the dataset

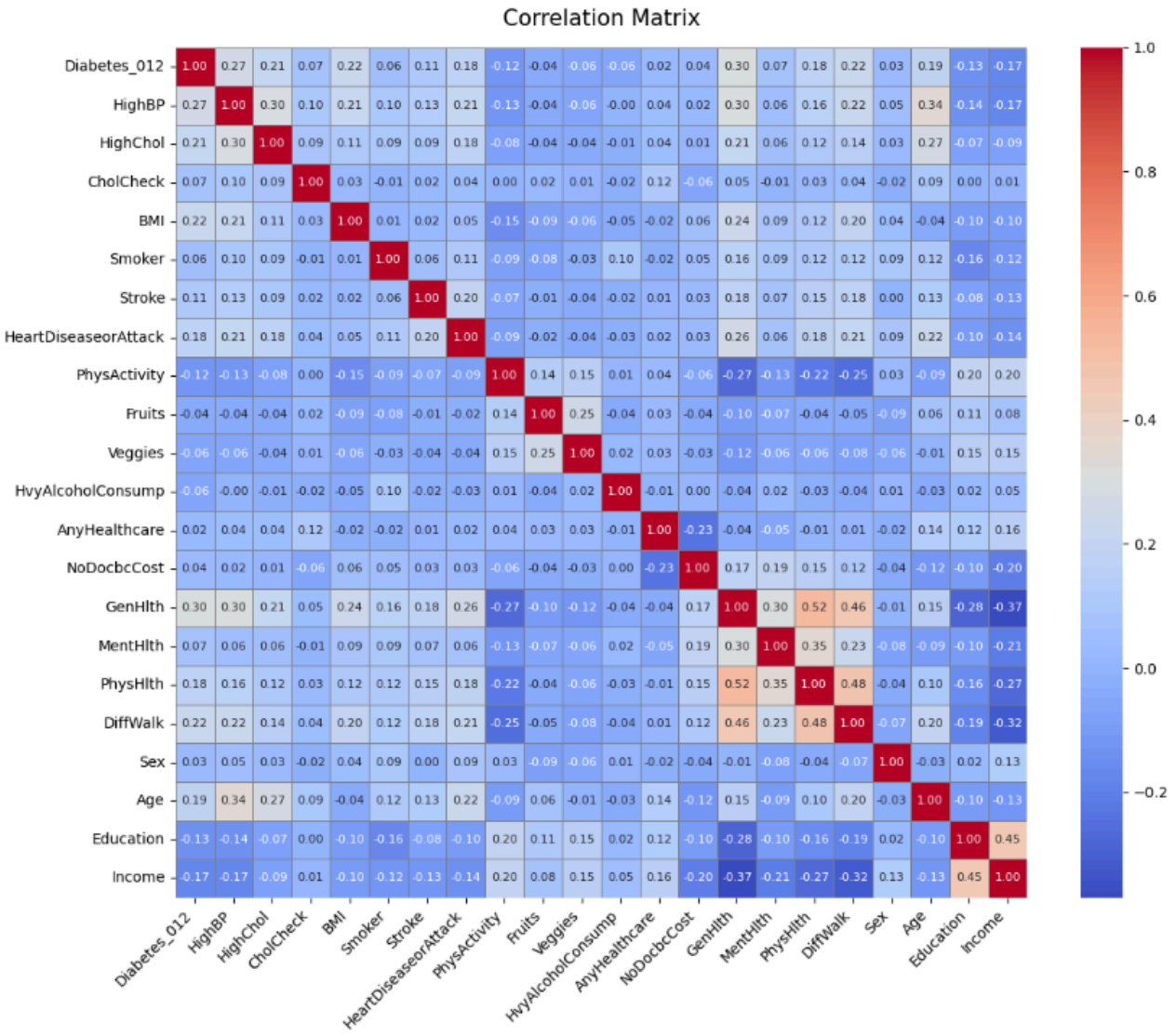

Figure 13 : - Correlation between different factors affecting diabetes

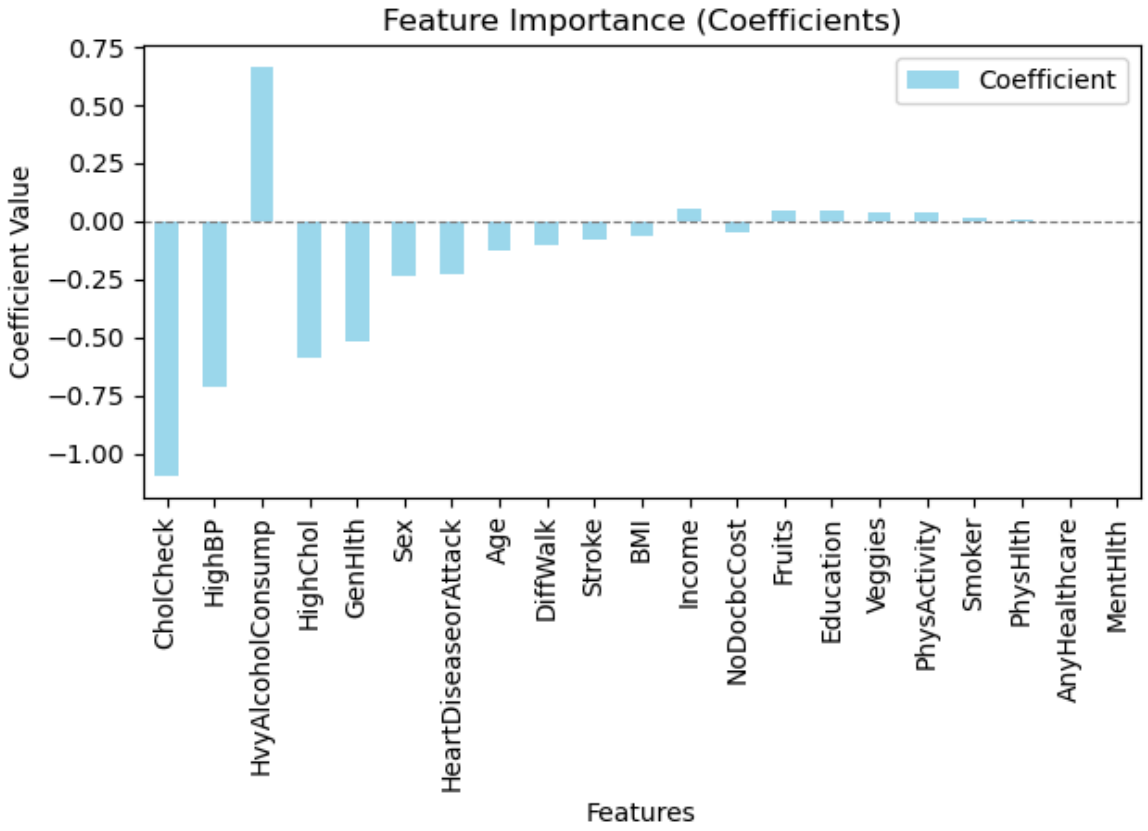

Figure 14 : - Feature Importance L2 Penalty

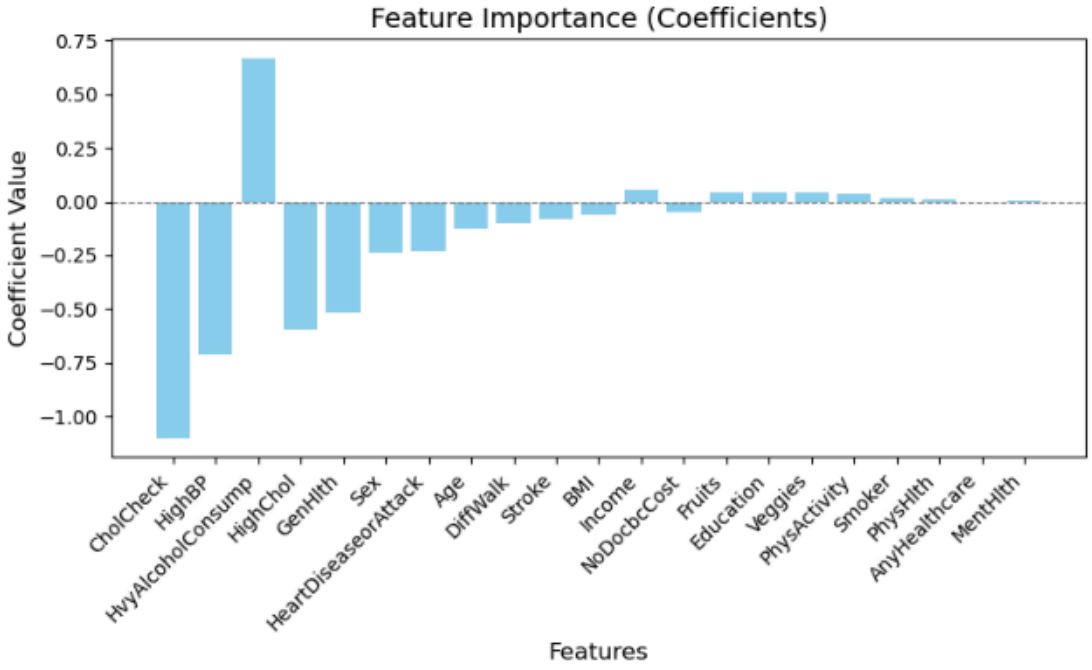

Figure 15 : - Feature Importance L1 Penalty

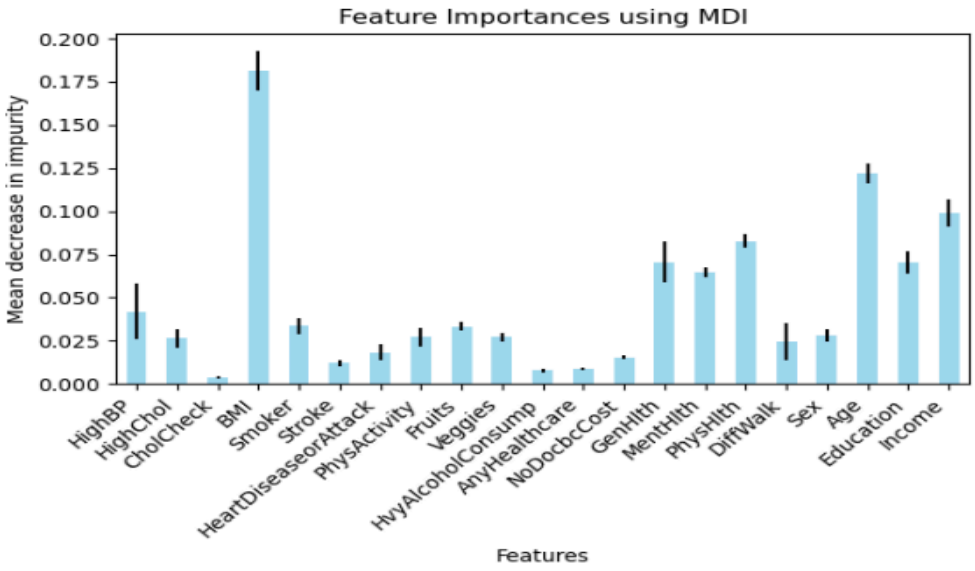

Figure 16 : - Feature Importance Random Forest MDI

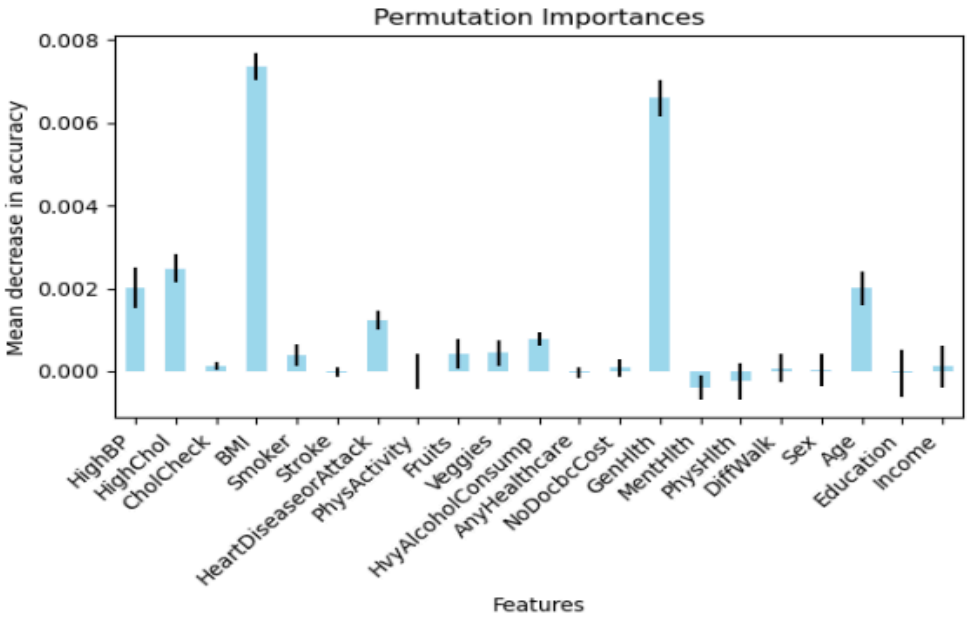

Figure 17 : - Feature Importance Random Forest Permutation Importance

Results : -

These lower accuracies are likely due to the categorical and binary structure of the second dataset, which restricted the model's predictive ability leading to worse results when contrasted with the first dataset.

Data Set 2

MODEL	Accuracy	Precision	Recall	F1 - score	ROC - AUC
Logistic Regression L1	84.65%	0.54	0.17	0.26	0.853
Logistic Regression L2	84.33%	0.54	0.17	0.26	0.853
Random Forest	84.175%	0.51	0.19	0.28	0.803
Dummy classifier	84.274%	0	0	0	0.5

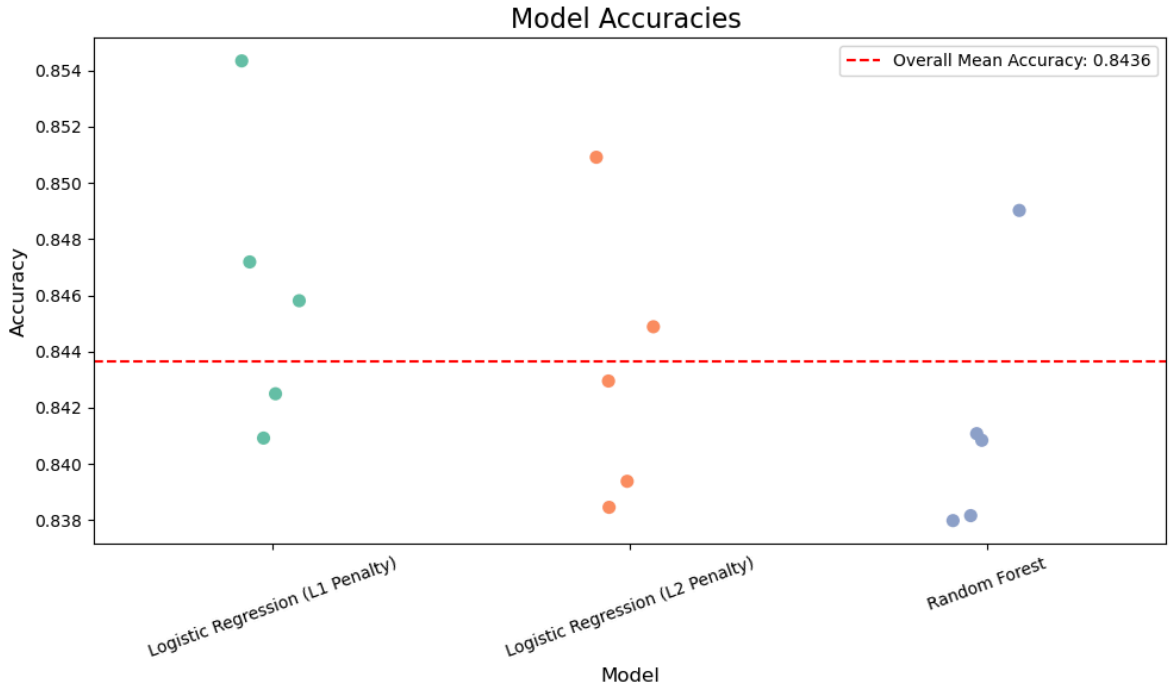

Figure 18 : - Model accuracies for data set 2 represented on a strip plot

Results

The aim of this was to evaluate multiple datasets comprising clinical and lifestyle variables to assess their predictive value for diabetes.. This was done through building two different models that iterated through publicly available datasets, with numerous machine learning models such as Random Forest, Logistic Regression (L2 penalty) , Logistic Regression (L1 penalty). We measured performance using accuracy on a held-out dataset. After multiple iterations, multiple accuracies for different models were obtained and plotted to compare and contrast with one another.

Across Dataset 1, the random forest classifier achieved the highest prediction accuracy of 97.003% (Figure 7), outperforming logistic regression models with both L2 (95.297%) and L1 (96%) penalties. The dummy classifier, included as a baseline for comparison, scored an accuracy of 91.667%,The dummy classifier in this instance was used to show contrast between models who actually take into account input data and predict and models who take simple rules and predict on the basis of those. Ex : - predicting all values as true or false, or randomly basing your prediction based off of one variable,etc.The dummy classifier in dataset one has an accuracy of up to almost 91% which depicts a representation of how the data set is actually structured it shows how the data set is imbalanced and a simple method of predicting all false or all true in this case has a high percentage probability of predicting the actual value. In data set two this value is lowered showing how due to the boolean nature and the percentage of values being true against the whole dataset is greater in this case then with dataset 1 it is harder to predict using simple rules developed by the dummy classifier.

Evaluation on Dataset 1 demonstrated that accuracy alone overstated performance due to class imbalance, as the dummy classifier achieved a 91.667% accuracy. Incorporating additional metrics provided a more complete view showing that the dummy classifier achieved a recall and precision of 0.00 - depicting that it failed to identify any diabetes cases. The random forest model achieved the best results, with a precision of 0.96 indicating that most patients predicted as diabetic were indeed diabetic. While its recall of about 0.68 showed it correctly identified a majority of actual diabetic patients, its F1 score of 0.79 showed the highest balance and depicted it to be the most reliable model of all of them. Logistic regression L1 regularization

In contrast, models trained on Dataset 2, which consisted mostly of binary and categorical variables, achieved lower accuracy overall. Evaluation on dataset 2 revealed weaker model performance across all metrics compared to dataset 1. While logistic regression for L1 regularization achieved the highest accuracy at 84.65%, its precision (0.54) and recall (0.17) indicate a significant struggle in correctly identifying diabetic patients. This low recall means that the majority if diabetic cases were misclassified as non-diabetic, and the F 1 score of 0.26 confirms the imbalance between precision and recall. Logistic regression with L2 regularization performed almost identically(accuracy 84.33%, F1 score 0.26), while the random forest model despite achieving the best recall among the three(0.19), only marginally improved the F1 score to 0.28 and produced the lowest ROC-AUC of 0.803, suggesting weaker discriminatory power. The dummy classifier once again highlighted the inadequacy of accuracy as a sole metric, recording an accuracy of 84.27% but failing entirely on precision, recall and F1 with an ROC-AUC of 0.5.

In comparison to dataset 1, where Random forest achieved strong results (F1 score- 0.79, ROC-AUC - 0.964) and logistic regression L1 maintained a solid balance(F1 score - 0.72, ROC-AUC - 0.961), Dataset 2 demonstrates a substantial decline in both discrimination and robustness. The dramatic drop in recall (from ~0.60-0.68 in Dataset 1 to ~0.17-0.19 in Dataset 2) and corresponding fall in F1 scores shows that the model struggled to generalize when applied to the second dataset. This contrast highlights Dataset2, which contained more categorical and binary features, posed a great challenge for all models, reinforcing the importance of incorporating multiple evaluation metrics rather than relying on accuracy alone.

The dummy classifier in this case performed at 84.274%, reinforcing that Dataset 2 was more balanced, and that basic heuristics alone were less effective in this context (Figure 18).

Confidence intervals indicated that performance estimates were highly stable across bootstrap samples. The random forest model consistently outperformed others, with its accuracy(0.971,95% cl:0.969-0.973) and F1 score(0.798, 95% cl:0.785-0.810) showing narrow intervals, confirming robustness.

In contrast Dataset 2 produced wider intervals and lower F1 scores across all models. Logistic regression (L1) achieved accuracy of 0.846 (95% cl:0.839-0.865), but its recall remained poor at 0.174(95% cl:0.167-0.181). These overlapping intervals highlight the difficulty of discriminating diabetic cases from categorical lifestyle data alone.

The accuracies of data set 1 models in comparison to dataset 2 models as dataset 2 contains a huge problem. The problem lies within how the dataset is structured and what type of values it stores. In this case dataset 2 stores mostly its values in the form of binary data where either the person is a smoker or non smoker, or whether the person consumes fruits or doesn't. The problem with this kind of data is that it doesn't explore the extent to which people are consuming fruits or smoke or any other variable for that matter. Thus the lack of precise data creates a similarity between individuals and narrows the line between variables that actually affect diabetes or not, as an individual who smokes once a month may not have diabetes but may still be listed as a smoker with the person who smokes everyday, with this person having diabetes thus making it harder for models to predict.

A major addition to the research was the addition of using a logistic regression model trained with an L1 penalty. An L1 penalty is when a logistic regression model weighs down certain variables with a negligible effect on the target variable to zero. This allows more emphasis on variables that play a significant role in determining whether the target variable is true or false thus being able to make a more accurate and precise prediction. However an L2 penalty is where these negligible factors aren’t weighed down to zero but are given a minimal value. In this research using an L1 penalty model did have a positive effect on the accuracy, being 96% in the first data set and 84.65% in the second data set showing a marginal improvement then when an L2 penalty for a logistic model was utilized. This shows how weighing down negligible factors allowed the logistic regression model to predict keeping into account only the main variables and it showed a higher precision in it’s prediction.

In the data set - 1 the variables that affected the target variables the most were HBA1C level, whether or not the patient had heart disease, hypertension and their bmi(figure 3 - 6). Although in this dataset that was unexpected to have this high of an impact on the final outcome which was gender(figure 4). Even through prior research however we know that gender may not have any correlation to diabetes this dataset showed different results a possible explanation for this could be due to the size of the dataset or the dataset being skewed in such a way that there were more women with diabetes than women with

non - diabetes and the opposite was true for the men in the dataset. This problem could be rectified by taking a more balanced and larger dataset, thus allowing us to explore the possibility of gender being an important correlate in the presence of diabetes.

In data set 2 there were variables like alcohol consumption, BMI, generational health, Physical Health , age that had a larger impact on the presence of diabetes(figure 14 - 17) although there were a few unexpected variables such as education,income, Mental Health(figure 16) that also had an influence on the target variable. Again a possible explanation for this could be the size of the dataset and its imbalance with its nature of being a boolean dataset making it more difficult to determine a correlate between two variables

Discussion

Dataset 1 contained few but more continuous variables, while dataset 2 comprised a large set of binary variables. Due to the distinct structuring of both datasets there were challenges while iterating through it to find the accuracies, for example due to the boolean type values in the second dataset, it was difficult for any machine learning model to identify key variables to predict whether the test sample had diabetes or not thus yielding a low accuracy and a varied feature importance graph which wasn't optimal for this research.

Additionally, another key limitation amongst datasets was the class imbalance that existed within dataset 1. This may have inflated the predictive accuracy, as suggested by the high accuracy of the dummy classifier.

Similar studies conducted, for instance - [7] showcase a similar approach and their results showcase similar results with factors such as smoking, BMI , hypertension,etc. This study also emphasizes that even though Type - 2 diabetes is predominantly affected by genetic factors, multiple lifestyle factors also shape the way on how severely diabetes can affect an individual.

The handling of missing values by encoding them as "no info" was another such limitation, which may have introduced noise into the models. Future studies should evaluate the usage of imputation techniques to improve predictive performance and reduce bias.

Also the exclusion of the "pre diabetes" individuals from data set - 2 reduced the overall clinical applicability since pre - diabetes is an important stage for early prevention. Future work should consider predicting all 3 outcomes.

Because simple K fold validation was used rather than stratified cross - validation, some folds may have had imbalance class distributions thus influencing accuracy values in dataset 1. Stratified K fold validation should be employed in the future to ensure class proportions are preserved across folds.

For future improvements I would consider trying out ensemble methods as a strategy to yield higher accuracy values and more accurate feature importance graphs. Moreover, I would consider a hybrid approach to this research, where using feature engineering would be key to making a dataset suitable for this project. Lastly, I believe taking a boolean dataset was the wrong approach as the tradeoff between discrete, precise values for a range of variables wasn't as suitable as I had presumed it would be.

Moreover, while model performance metrics were averaged across folds to ensure stability, feature importance values were reported from single model runs. This means the relative ranking of predictors may vary with different random seeds or cross validation splits, particularly for less influential variables. Future iterations of this work should address this by averaging importance scores across folds or by using permutations-based methods and reporting confidence intervals, to ensure that conclusions about the relative influence of the predictors remain robust.

References

Ahmed, N. (n.d.). Machine learning based diabetes prediction and development of smart web application. *Science Direct*. https://doi.org/10.1016/j.ijcce.2021.12.001

Alhussan, A. (n.d.). Classification of Diabetes Using Feature Selection and Hybrid Al-Biruni Earth Radius and Dipper Throated Optimization. *PUBMED Central*. 10.3390/diagnostics13122038

Farida, M. (n.d.). A scoping review of artificial intelligence-based methods for diabetes risk prediction. *PUBMED*. 10.1038/s41746-023-00933-5

Jyoti, R. K. (2020). Diabetes Prediction Using Machine Learning. *IJSCSEIT*. https://doi.org/10.32628/CSEIT206463

Khokar, P. B. (2025). Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. *Science Direct*. https://doi.org/10.1016/j.artmed.2025.103132

Lugnar, M. (2023). Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific Reports*. https://doi.org/10.1038/s41598-024-52023-5

Negi, P. (n.d.). Evaluating Feature Selection Methods to Enhance Diabetes Prediction with Random

Forest. *ACM*. https://doi.org/10.1145/3647444.364793

Noh, M. J. (n.d.). Diabetes Prediction Through Linkage of Causal Discovery and Inference Model with Machine Learning Models. *MDPI*. https://doi.org/10.3390/biomedicines13010124

Perez, E. R., & Molano, B. A. (2025). Learning from the machine: is diabetes in adults predicted by lifestyle variables? A retrospective predictive modelling study of NHANES 2007-2018. *PUBMED*. 10.1136/bmjopen-2024-096595

Qin, Y. (n.d.). Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. *PUBMED*. 10.3390/ijerph192215027

Ravaut, M. (2021). Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes. *PUBMED*.

Zhou, H. (2023). A diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC*. https://doi.org/10.1186/s12859-023-05300-5