Comparative Analysis of Mental Health & Physiological Predictors of Diabetes Using Machine Learning by Shubh Ghorpade

Author: Shubh Ghorpade

Grade: 11th Grade

School: Cambridge Program at Juanita High School, Kirkland, WA

Mentor: Dr. Kiran Kumar Vejendla (DIT) VejendlaKiran@gmail.com

Abstract

Worldwide, diabetes continues to challenge public health systems due to increasing rates and serious complications. Spotting those early might help guide prevention efforts that ease overall impact. This study evaluates the performance of machine learning models for diabetes prediction using survey-based data. Data was obtained from the 2015 BRFSS survey.

From a balanced sample of 70,692 people, both Logistic Regression and Random Forest approaches took shape through an 80/20 training-testing division. Rather than relying on single metrics, evaluation leaned on Classification Accuracy, AUC, alongside the F1 measure for balance. Feature importance was evaluated using Information Gain to quantify each predictor's contribution. Importance emerged not by assumption, but through calculated shifts in variable impact.

Despite its simplicity, Logistic Regression led in accuracy - scoring an AUC of 0.823 and correct classification at 74.6%. Notably, it surpassed Random Forest on every measured outcome. Key contributors to prediction strength included General Health, High Blood Pressure, and difficulty moving around. In contrast, mental well-being appeared less influential, landing near the bottom when ranked among predictors.

The results indicate that physiological variables provide stronger predictive value than mental health indicators when spotting diabetes. Physical signs demonstrated stronger predictive influence in forecasting illness, even though inner well-being still plays a background part. Mental strain shows up only through its ripple effects on the body.

1. Introduction

One out of every eleven people worldwide live with diabetes, straining medical services across nations (World Health Organization [WHO], 2024). Spotting it early helps avoid serious problems - heart issues, failing kidneys, or daily struggles - that follow unchecked progression (American Diabetes Association [ADA], 2023). For years, studies looked mostly at physical markers like BMI, high blood pressure, along with raised cholesterol readings (National Institute of Diabetes and Digestive and Kidney Diseases [NIDDK], 2016).

Now emerging studies examine how mind well-being ties into long-term illness, pointing to stress and emotional struggles as drivers of metabolic shifts (Kivimäki & Tabák, 2014). Still, whether signs of mental state hold more weight than standard body measurements in forecasting risk are uncertain - especially across broad public health records.

This work tackles a key shortfall through supervised machine learning applied to the 2015 BRFSS data source (Centers for Disease Control and Prevention [CDC], 2015), testing whether signs of mental well-being forecast diabetes likelihood better than standard factors. Aimed at contrasting various forecasting methods (Hassan et al., 2024), it also pinpoints health traits that boost prediction precision most strongly.

2. Definition of Research Terms

2.1 Orange Data Mining (Software)

The Orange Data Mining software is an open-source, visual programming software used for data visualization, machine learning, and data mining. This paper used this software to research and build complex analytical "workflows" through a drag-and-drop interface instead of traditional coding.

2.2 Feature Ranking (Information Gain)

The feature ranking is a mathematical method to calculate the "relevance" of each piece of data. In this study, feature ranking is used to determine which lifestyle factors (like HighBP or BMI) provide the most information about whether a person has diabetes.

2.3 Logistic Regression (The "Winner")

Logistic Regression is a statistical model to predict the probability of a "binary" outcome (like Yes/No, or 0/1).
In this study, Logistic Regression is used because it is the "gold standard" in medical research for predicting disease risk because it is clear, logical, and highly accurate (Bagley et al., 2001).

2.4 Random Forest (The "Comparison")

Random Forest is a complex algorithm that uses multiple "decision trees" and merges them together to get a more stable and accurate prediction. Random Forest is used in this study to see if a complex computer brain could "outperform" the simpler Logistic Regression model.

2.5 The 80/20 Train-Test Split

The 80/20 train-test split is a validation technique in machine learning model development and testing where 80% of the data is used to “train” the model and the rest 20% of the data is used to test the model’s outcome accuracy. This study used an 80/20 train-test split to prove that the developed model can accurately predict diabetes for "new" people it has never seen before, rather than just memorizing the old data.

2.6 Classification Accuracy (CA)

Classification Accuracy (CA) is the measure to identify the percentage of total guesses the model got correct. The CA percentage will determine if the model is right, and if so, how many times. This study used CA to determine the overall success rate of the developed model.

2.7 Area Under the ROC Curve (AUC)

Area Under the ROC Curve (AUC) is a measure to determine how well the model distinguishes between two groups (Diabetic vs. Non-Diabetic). AUC has scores between 0 and 1. Highest score (1 is top score) is considered "Excellent," meaning the model has a very high ability to tell the difference between the two groups.

2.8 Confusion Matrix

Confusion Matrix is a table to visualize the specific types of errors a model makes. This is used to evaluate the practical reliability of the model.This matrix is used in the study to see not just if the model is wrong, but how it is wrong (e.g., did it miss a sick person or give a false alarm?).

3. Materials and Methods

3.1 Dataset

In this study, a large dataset (n=70,692) from Diabetes Health Indicators Dataset derived from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) is used to analyze the top contributors (Physiological factors and psychological indicators) to Diabetes. This dataset has a balance (n0=35,346, n1=35,346) of diabetic and non-diabetic individuals. The data is cleaned public subset of 2015 BRFSS survey responses compiled by Teboul (2021) on Kaggle.

The data were downloaded in the Microsoft excel file (diabetes_binary_5050split_health_indicators_BRFSS2015.csv). This Excel file contained 22 unique fields along with their data.

Field Name	Description
Diabetes_binary	The target variable, Diabetes_binary, is utilized in a consolidated format where '0' represents no diabetes and '1' encompasses both pre-diabetic and diabetic clinical statuses. This binary classification approach is selected to maximize the model's sensitivity in identifying at-risk individuals during preliminary health screenings.
HighBP	Does person have High Blood Pressure 0 = no high BP, 1 = high BP?
HighChol	Does person have High Cholesterol? 0 = no high cholesterol, 1 = high cholesterol
CholCheck	Cholesterol Check done in past 5 years 0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years
BMI	Body Mass Index
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no, 1 = yes

Stroke	You had a stroke. 0 = no, 1 = yes
HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no, 1 = yes
PhysActivity	Physical activity in past 30 days - not including job 0 = no, 1 = yes
Fruits	Consume Fruit 1 or more times per day 0 = no, 1 = yes
Veggies	Consume Vegetables 1 or more times per day 0 = no, 1 = yes
HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no, 1 = yes
AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no, 1 = yes
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no, 1 = yes
GenHlth	Would you say that in general your health is: scale 1-5 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor
MentHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth	Now thinking about your physical health, which includes physical illness and injury, how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk	Do you have serious difficulty walking or climbing stairs? 0 = no, 1 = yes
Sex	0 = female, 1 = male

Age	13-level age category 1 = Age 18 to 24 2 = Age 25 to 29 3 = Age 30 to 34 4 = Age 35 to 39 5 = Age 40 to 44 6 = Age 45 to 49 7 = Age 50 to 54 8 = Age 55 to 59 9 = Age 60 to 64 10 = Age 65 to 69 11 = Age 70 to 74 12 = Age 75 to 79 13 = Age 80 or older
Education	Education level scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
Income	Income scale 1-8 1 = Less than $10,000 2 = $10,000 to less than $15,000 3 = $15,000 to less than $20,000 4 = $20,000 to less than $25,000 5 = $25,000 to less than $35,000 6 = $35,000 to less than $50,000 7 = $50,000 to less than $75,000 8 = $75,000 or more

Table 1. Diabetes Health Indicators Dataset Fields (columns)

The following is a sample data format from diabetes_binary_5050split_health_indicators_BRFSS2015.csv. Please note that the table is transposed to show rows of fields on column1 instead of row1, this is done to accommodate all the fields in the same page.

Column	**<<<---Data--->>>**					
Diabetes_binary	0	0	0	0	0	…
HighBP	1	1	0	1	0	…
HighChol	0	1	0	1	0	…
CholCheck	1	1	1	1	1	…
BMI	26	26	26	28	29	…
Smoker	0	1	0	1	1	…
Stroke	0	1	0	0	0	…

HeartDiseaseorAttack	0	0	0	0	0	…
PhysActivity	1	0	1	1	1	…
Fruits	0	1	1	1	1	…
Veggies	1	0	1	1	1	…
HvyAlcoholConsump	0	0	0	0	0	…
AnyHealthcare	1	1	1	1	1	…
NoDocbcCost	0	0	0	0	0	…
GenHlth	3	3	1	3	2	…
MentHlth	5	0	0	0	0	…
PhysHlth	30	0	10	3	0	…
DiffWalk	0	0	0	0	0	…
Sex	1	1	1	1	0	…
Age	4	12	13	11	8	…
Education	6	6	6	6	5	…
Income	8	8	8	8	8	…

Table 2. Sample Data Format from the Dataset leverages in the Model Development

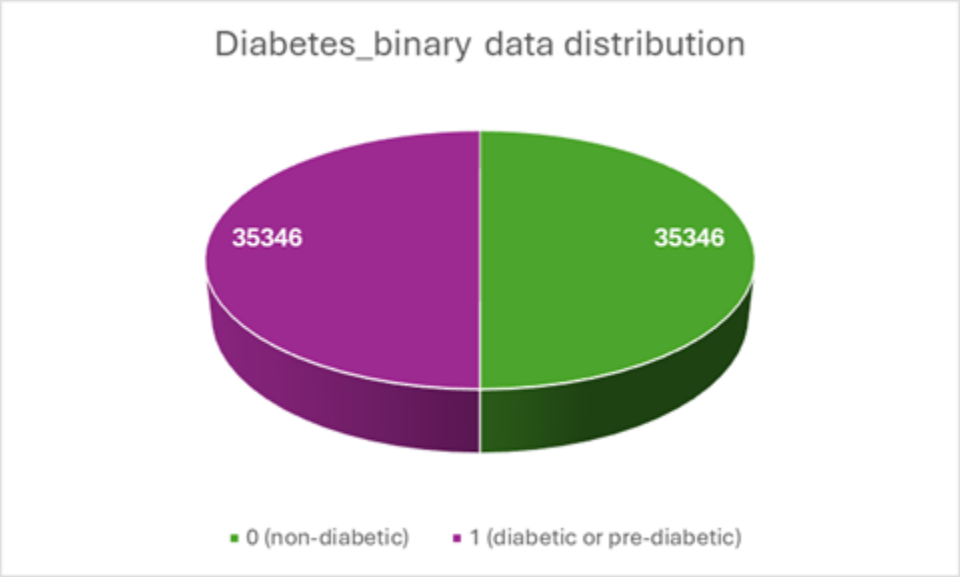


Figure 1. Binary Data Distribution for Diabetic and Non-Diabetic Individuals

3.2 Data Preprocessing and Tools

To conduct this analysis, an open-source visual programming platform for machine learning called Orange Data Mining (version 3.x) is used. This allowed us to build, test and analyze the dataset without having to do any coding.

The variable Diabetes_binary is assigned a role “target” in the Orange Data Mining File component to facilitate supervised learning. This is required so the algorithm can find the correlation between 21 dependent variables with Diabetes_binary and determine the rankings. These dependent variables are classified as follows:

- Physiological factors (e.g., BMI, HighBP)
- Demographic variables (e.g., Age, Education, Income)
- Psychological indicators (e.g., MentHlth)

As the Kaggle dataset was structured and cleaned, no data preprocessing or transformation was required.

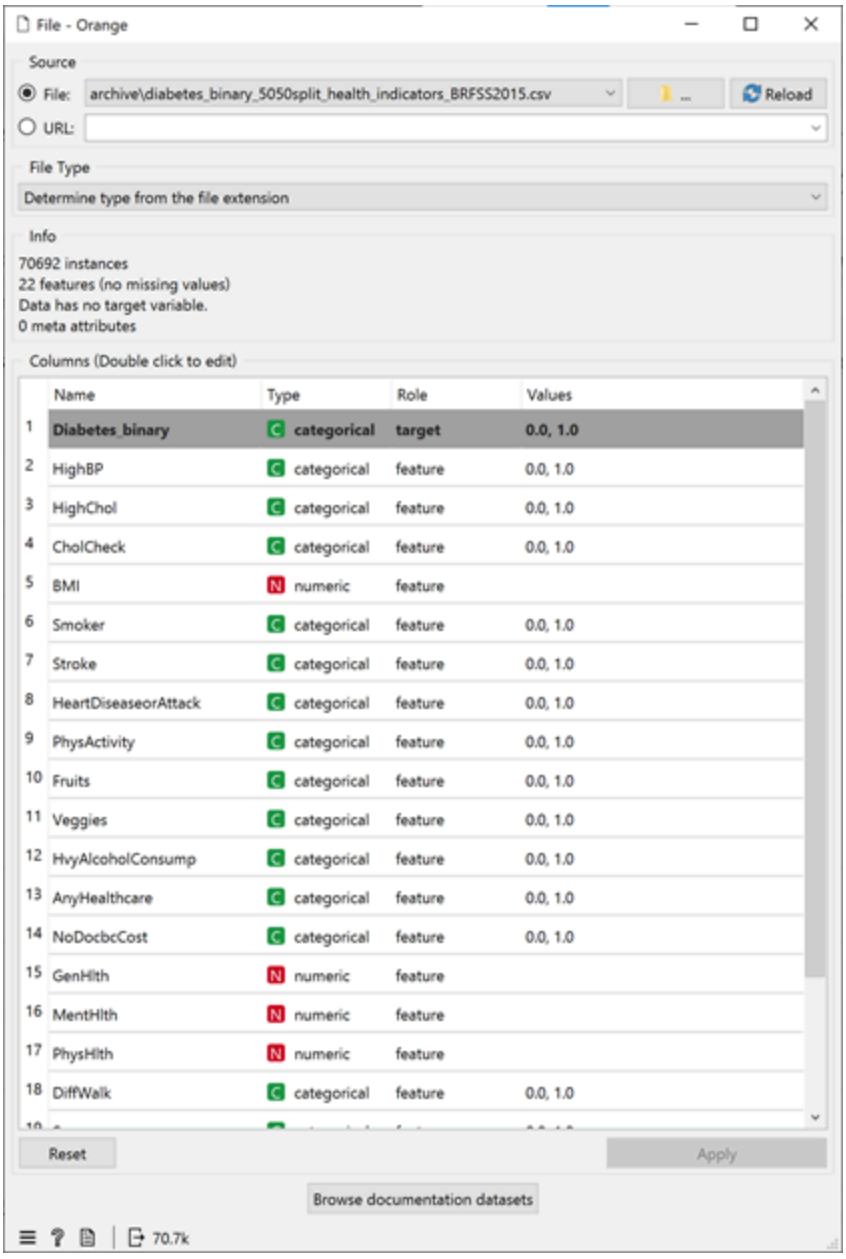


Figure 2. Defined Variable Diabetes_binary with "Target" Role to Facilitate Supervised Learning in Orange Data Mining Software

3.3 Model Development

The objective of model development was to evaluate whether diabetes status could be accurately predicted from survey-based health indicators. To develop a model, the following machine learning algorithms are used. The rationale to use these algorithms is as follows - Logistic Regression for its high interpretability and its status as a foundational standard in clinical predictive modeling (Hassan et al., 2020) and Random Forest to evaluate whether a more complex, ensemble-based approach could identify non-linear relationships within the 21 health indicators to improve classification accuracy (Breiman, 2001).

- Logistic Regression (linear model)
- Random Forest (ensemble-based model)

Machine learning models require a dataset for training and testing purposes to measure its accuracy before it can be used with real-world use cases. The Diabetes Health Indicators Dataset (n=70,692) is divided into two subsets. One for training (80% data; n=56,553) and the other for testing (20% data; n=14,138). Diabetes prediction model is then trained on the training dataset and later evaluated on unseen test data.

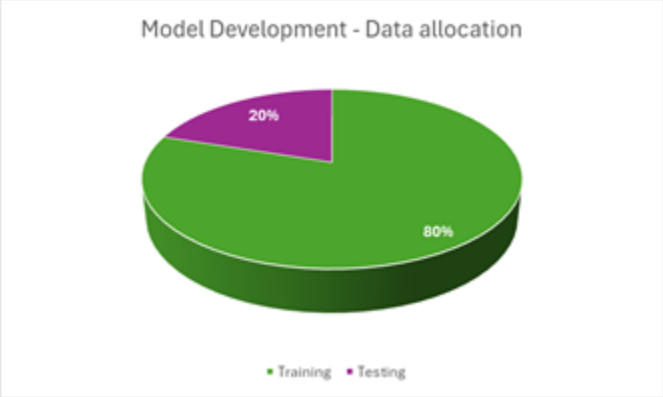


Figure 3. Training and Testing Data Allocation for Model Development

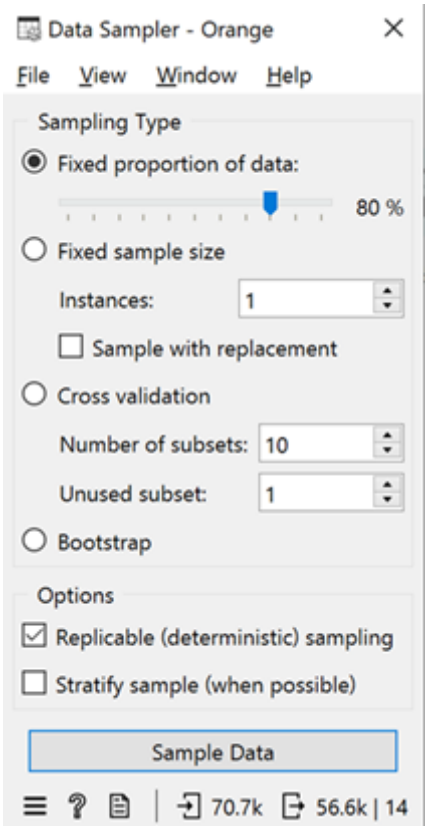


Figure 4. Data Sampler in Orange Data Mining Software – 80% Fixed Proportion of Data to Train the Model

3.4 Evaluation Metrics

After the model is built and trained (n=56,553) using the Orange Data Mining software, the next step is to grade the model to evaluate how it is performing. The 20% dataset (n=14,138) is leveraged to evaluate the models. The model performance is quantified using the following performance metrics.

- Classification Accuracy (CA): The proportion of total instances correctly classified by the model.
- Area Under the ROC Curve (AUC): A measure of the model's discriminatory power, or its ability to distinguish between diabetic and non-diabetic classes.
- F1 Score: The harmonic meaning that provides a balance between precision (exactness) and recall (completeness).

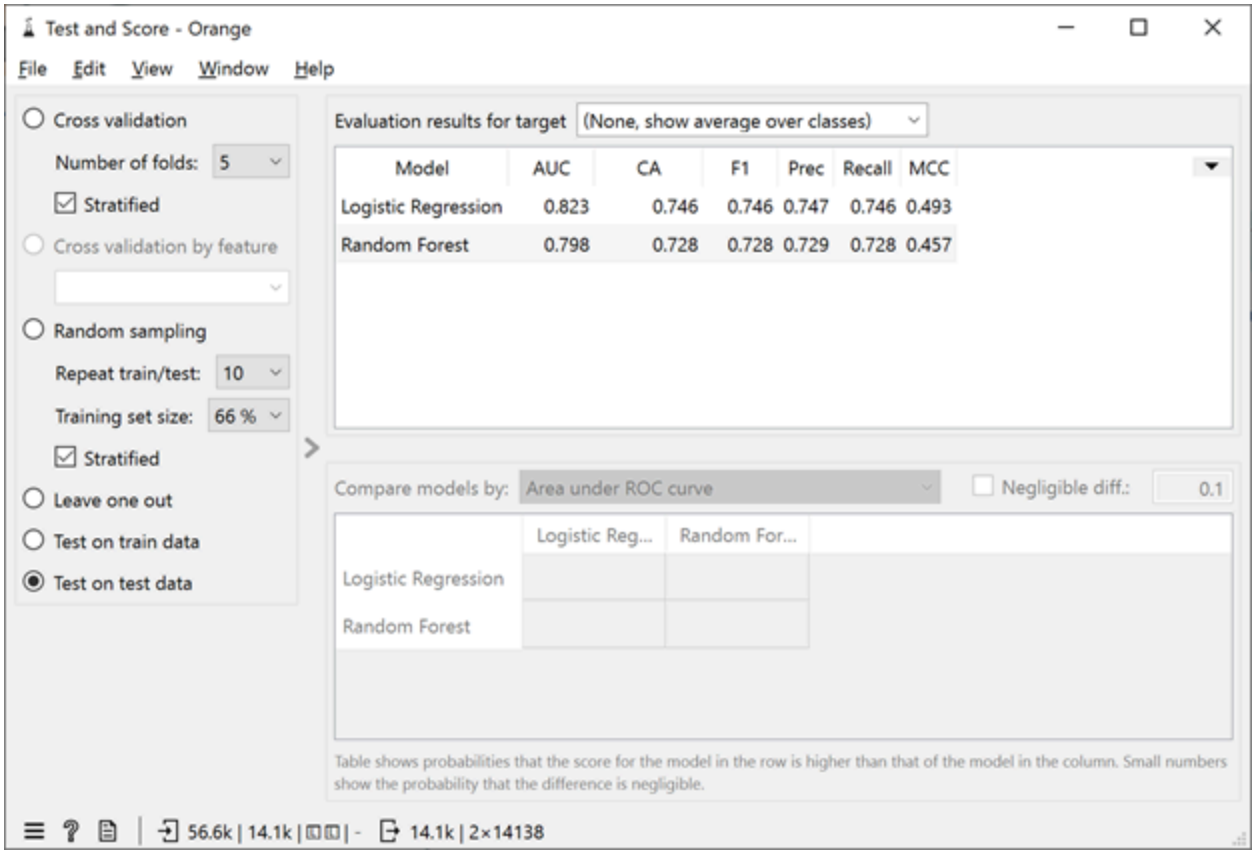


Figure 5. Model Evaluation using AUC, CA and F1 Performance Metrics

3.5 Statistical Analysis

Even though mental health didn't rank highest in the model's predictions, this study further checked if there is a measurable gap in 'poor mental health days between people with and without diabetes. Instead of assuming similarity, this analysis applied an Independent Samples t-test to examine average values across the two distinct groups. To decide importance, study relied on a p-value below 0.05 - meaning results likely reflect real differences rather than coincidence. The threshold helps rule out randomness when interpreting outcomes.

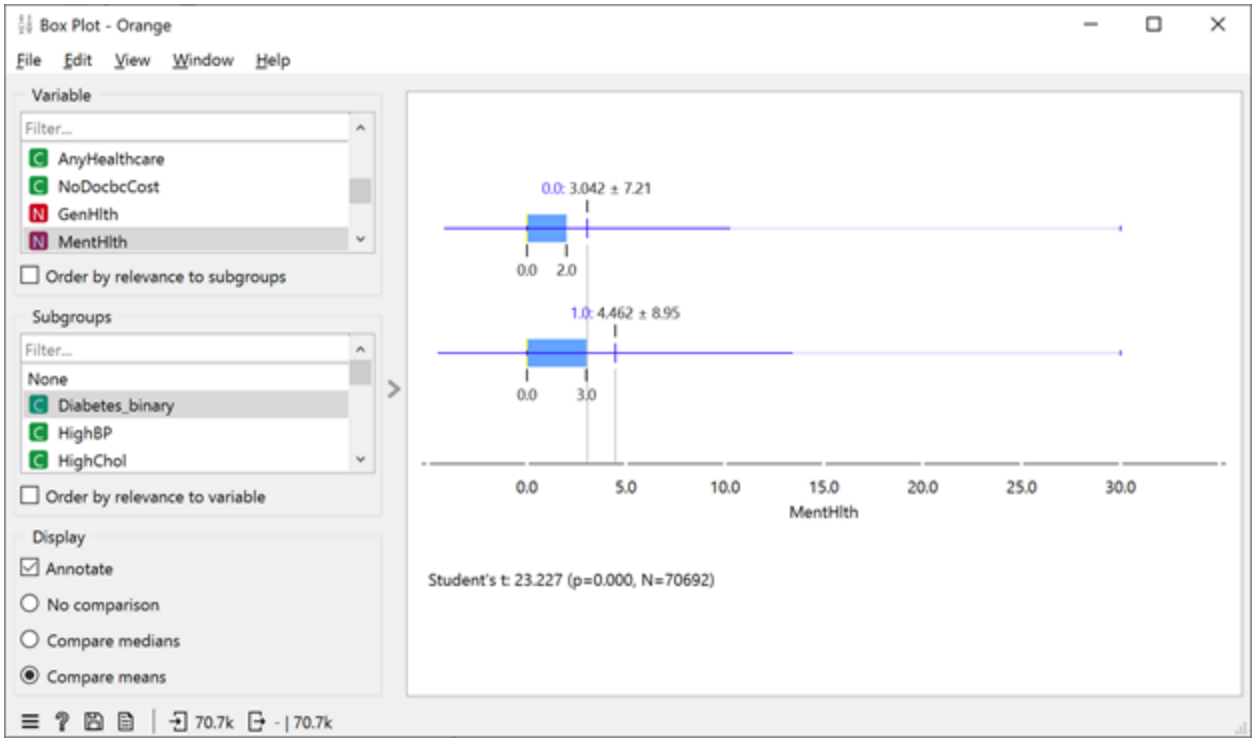


Figure 6. The Box Plot with p-value demonstrating measurable gap in 'poor mental health days' between people with and without diabetes

4. Results

4.1 Feature Importance (The "Discovery")

Using the Orange Data Mining Software, the workflow is defined between the source file (diabetes_binary_5050split_health_indicators_BRFSS2015.csv) and Rank. Diabetes_binary is configured for the "target" role.

Top features linked to diabetes included overall health status, high blood pressure, along with walking trouble - these are identified as the top predictors. Mental well-being appeared much lower in influence, landing near the bottom when sorted by impact strength.

Despite expectations pointing toward mental well-being as key, findings show bodily health markers matter more when sorting outcomes. Yet the initial assumption leaned heavily on psychological factors, the results indicate greater predictive contribution from physiological variables compared to psychological indicators.

The results indicate that though researchers expected mental well-being to play a major role, it placed only 17th among the 21 measured factors. Instead, data from the 2015 BRFSS group show that bodily symptoms and movement difficulties predicted diabetes far more clearly than feelings of emotional strain reported by individuals.

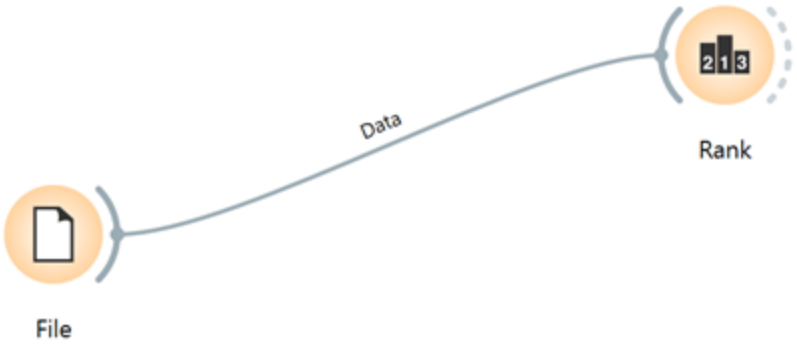


Figure 7. The Workflow Methodology between Dataset (file) and Rank in Orange Data Mining Software

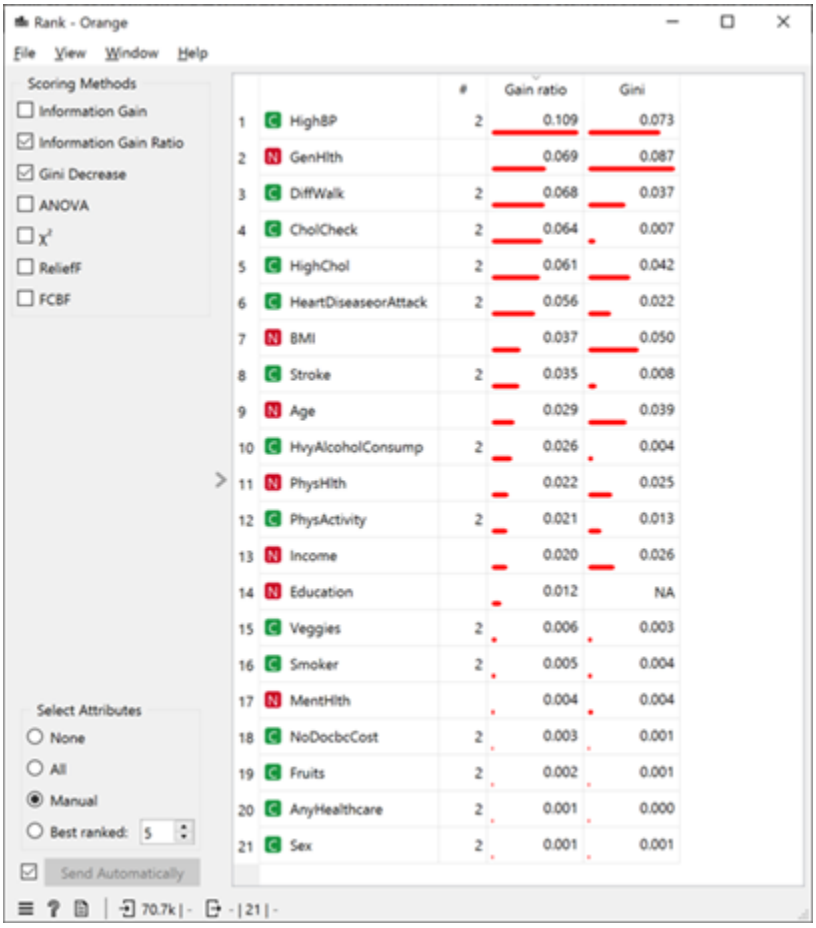


Figure 8. The Rank Table with Top Three Variables (HighBP, GenHlth, DiffWalk)

4.2 Model Performance

Logistic Regression achieved a classification accuracy of 74.6%, outperforming Random Forest. Achieving an AUC of 0.823, it demonstrated higher discriminative performance compared to Random Forest. Despite greater complexity, Random Forest falls short - suggesting simpler relationships might effectively connect biomarkers to diabetes risk. Strong discriminative performance emerges here, indicating usefulness during initial stages of medical evaluation.

A key finding is the strong link between elevated performance scores and the study's main goal. These models may support non-invasive screening approaches. Instead of relying on standard lab tests, this approach uses observable patterns. This shift suggests potential application - one where detection does not require invasive procedures.

Model	AUC	CA (Accuracy)	F1 Score
Logistic Regression	0.823	0.746 (74.6%)	0.746
Random Forest	0.798	0.728 (72.8%)	0.728

Table 3. Logistic Regression outperformed Random Forest in all primary metrics

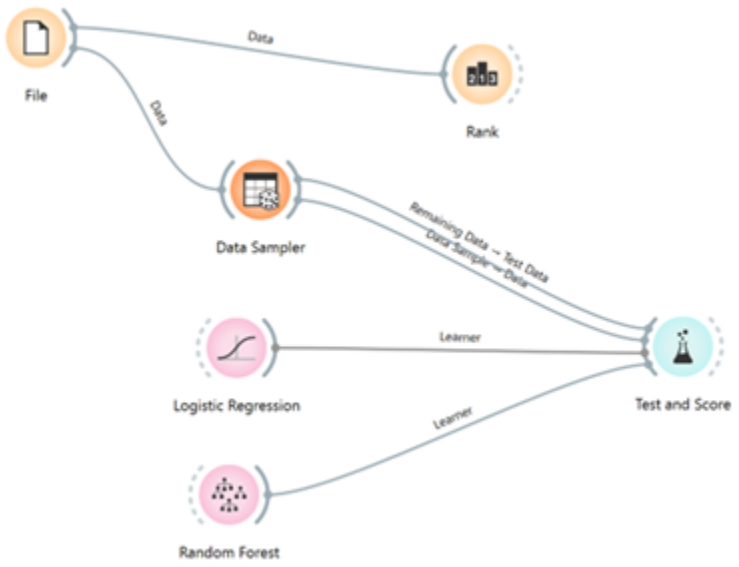


Figure 9. Logistic Regression and Random Forest model Workflow for Model Development in Orange Data Mining Software



Figure 10. AUC, CA and F1 Score Performance Comparison

Supplemental Analysis of Mental Health Indicators

Looking into the reason MentHlth ranks low - 17th among 21 predictors - a comparison of averages is carried out through a Box Plot. Results showed a clear statistical gap (Student's t-test, $p < 0.001$) when examining both groups. On average, those with diabetes noted 4.46 days of poor mental health monthly; without diabetes, the figure stood at 3.04. Though subtle, the pattern holds across samples. That difference appears consistent under testing conditions.

Though the 46% rise in reported distress links diabetes to mental health issues, its position at number 17 in prediction tools comes from broad differences in responses. That scatter - shown by a ±8.95 spread - weakens its role as an early alert compared to steadier signs such as raised blood pressure or trouble moving. High variability reduces its predictive reliability compared to more stable physiological indicators.

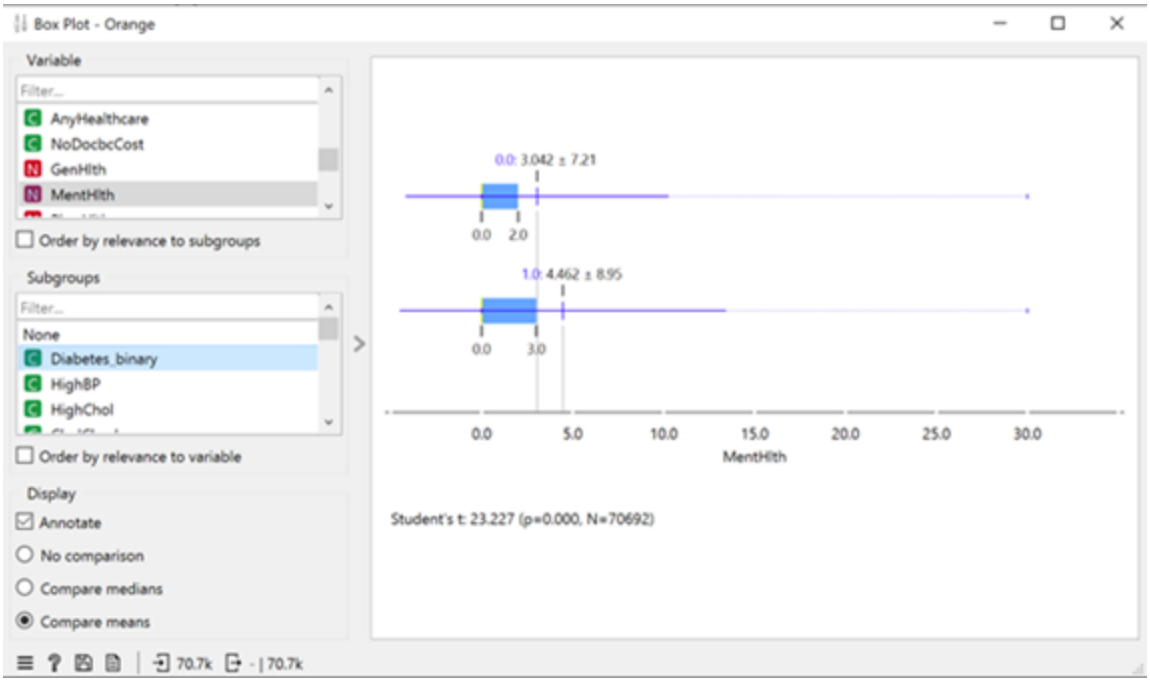


Figure 11. The Box Plot - Comparative Mean Analysis between MentHlth and Diabetes

5. Discussion

This study demonstrates that machine learning tools can effectively forecast diabetes accurately from broad population surveys, delivering solid results even without advanced methods. Despite its simplicity, Logistic Regression did better than Random Forest, hinting that most links between variables and outcomes follow a straight-line pattern. Though ensemble approaches often win, here straightforward math captures the core trends well.

Mental health indicators contributed less to predictive performance compared to physiological variables. Although people living with diabetes showed greater emotional strain, it is physical aspects - like trouble moving or elevating blood pressure - that shaped predictions more strongly. The results show how narrow the influence remained despite clear links to distress.

This may suggest that psychological well-being might shape diabetes risk by shaping daily habits tied to bodily health. Biological markers appeared to offer clearer, faster clues when forecasting disease onset.

Despite its usefulness in screening, the model's false negatives suggest clinical judgment must still guide decisions. A look at the confusion matrix shows where errors occur, reinforcing that algorithms assist rather than take over. While patterns emerge clearly in correct predictions, gaps remain when results miss actual cases. These misses mean human oversight remains essential. Even strong performance cannot remove the need for doctor involvement. These findings highlight the continued importance of clinical judgment alongside model predictions.

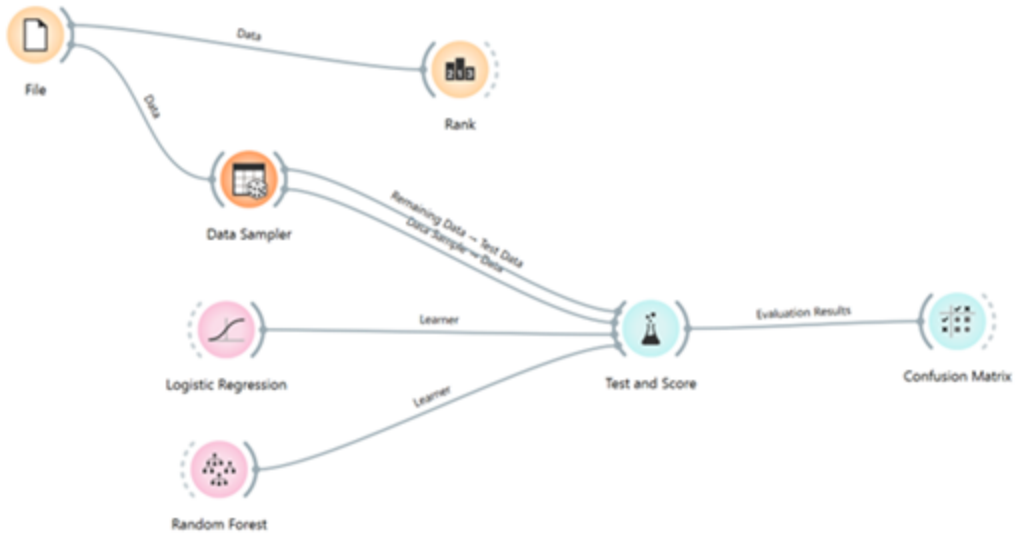


Figure 12. The End-to-End Workflow with Dataset, Ranking, Algorithms and Confusion Matrix

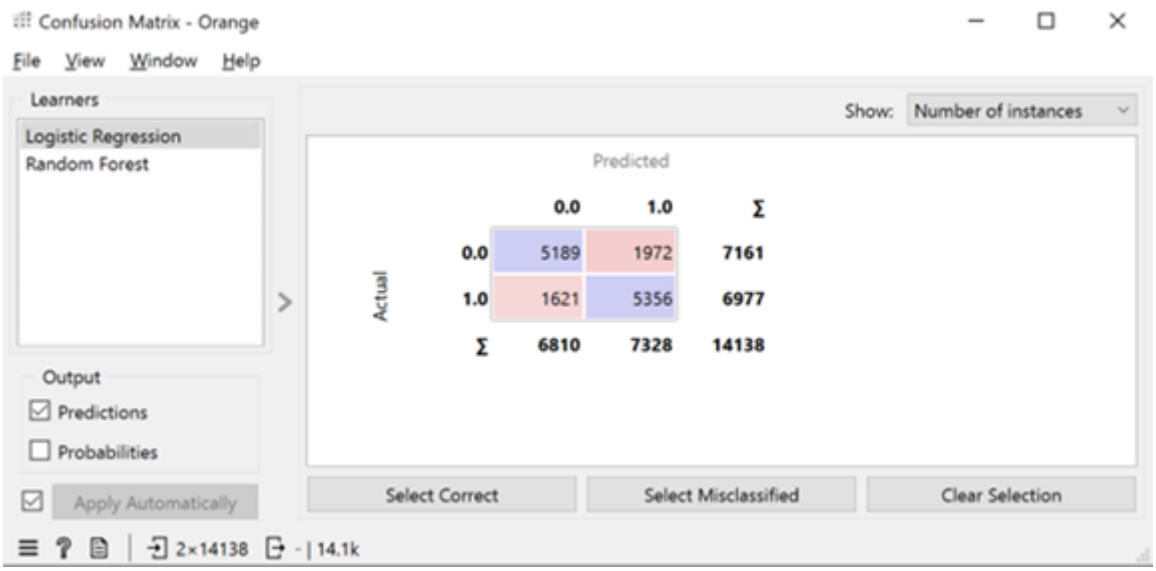


Figure 13. Confusion Matrix Results from Orange Data Mining Software

Check the figures shown in the table under Logistic Regression:

- Healthy individuals are accurately labeled by the system: 5,189 cases where no illness is present and the result matches reality.
- Of those predicted to have diabetes, 5,356 were accurately identified by the model.
- Wrong calls where sickness existed yet prediction claimed wellness - that count as false negatives, 1,621 cases. Though the system labeled them disease-free, truth revealed diabetes present instead.

- Starting off, False Positives (1,972) showed cases where the system mistakenly flagged individuals as diabetic despite them being healthy. At times, results suggested disease where none existed. The mechanism sounded warnings even when blood sugar levels were normal. Those identified did not have the condition at all.

A notable finding is that mental health appears much lower, at 17th place. Its role in diabetes might not be direct; instead, it could shape behaviors linked to higher body mass or overall wellness. This may suggest that emotional well-being influences daily habits more than biological pathways. Looking ahead, studies might explore whether new data from after the pandemic show a different position, given recent shifts in teen psychological health.

The results demonstrate that a no-code method reached notable prediction levels - 74.6% - for spotting long-term illnesses. Despite common assumptions, Logistic Regression did better than Random Forest, hinting at clear straight-line patterns between survey markers and diabetes risk. Because of this trend, intricate algorithms may add little value when using BRFSS data for such assessments.

The results indicate MentHlth ranked 17th. Despite frequent claims connecting stress and mental health to metabolism, DiffWalk appeared stronger here as a warning sign. This may suggest that when diabetes shows up in such markers, bodily strain already outweighs mental burden. Physical limits show earlier than emotional ones.

Looking more closely, one part of the study checked whether mental health mattered more for older individuals (age above 10). Feature rankings stayed largely unchanged, pointing to a steady pattern. Even in varying age segments, physical signs like HighBP and DiffWalk remained stronger predictors than emotional factors. This trend held firm throughout the group.

6. Conclusion

This study built and tested machine learning approaches to forecast diabetes, relying on health information collected in the 2015 BRFSS survey. Out of all methods tried, Logistic Regression performed best - correctly classifying outcomes nearly three out of four times while showing strong discriminatory power at 0.823 AUC. Simplicity here proves effective; straightforward models can handle broad public health tasks well, even beating opaque systems that are harder to interpret.

The results indicated that the results place psychological stress lower than expected in predicting diabetes. Instead of focusing first on mood or anxiety, models pointed toward blood pressure and movement issues as stronger signals. Often beginning with bodily functions, accurate predictions seem tied less to mind-set and more to measurable body changes. Walking difficulty consistently ranked among the strongest predictors. Rather than grouping all risks equally, patterns reveal some matter earlier than others.

Practical Implications and Future Work

This study matters because it opens doors to care that many can reach. Without needing medical procedures, the approach uses simple survey answers, making it suitable for basic smartphone apps. People without strong health systems nearby might check possible risks early.

Starting from concern about symptoms to getting official tests could become faster. Earlier steps toward treatment may follow.

Looking ahead, newer data collected since the pandemic could reveal whether mental health's role in prediction has changed alongside broader shifts in psychological states. Studies down the line might examine how mood and behavior connect - linking emotional patterns to habits like eating or sitting too long - to see if those influence the body signals flagged here as key. Rather than acting alone, mental wellness may shape physical outcomes through these daily behaviors, subtly guiding the pathways that lead to measurable effects.

7. Limitations

One limitation of this study lies in its reliance on 2015 data. Although core biological risks tend to change slowly, cultural attitudes toward mental health have evolved noticeably since then - so findings could shift if newer information were used instead. Because public understanding has grown, how people report or recognize symptoms today may differ, possibly altering outcomes seen earlier.

Acknowledgments

Mentorship: I would like to express my sincere gratitude to my mentor, Dr. Kiran Kumar Vejendla (DIT), for his invaluable guidance, technical review, and support throughout this research process. His expertise helped ensure the statistical integrity of this analysis.

Data Source: This study utilizes data from the Centers for Disease Control and Prevention (CDC) 2015 Behavioral Risk Factor Surveillance System (BRFSS). All data analysis, including model building, training and testing, tabular and charts analysis, is performed independently by the author using Orange Data Mining tool.

AI Disclosure: The author used ChatGPT for initial brainstorming and to refine the manuscript's structure and clarity. All original research, data analysis, and final interpretations were conducted solely by the author, who holds full responsibility for the content

Works Cited

Orange Data Mining Software.
https://orangedatamining.com/

Centers for Disease Control and Prevention (CDC). (2015). Behavioral Risk Factor Surveillance System Survey Data.
https://www.cdc.gov/brfss/annual_data/annual_2015.html

Teboul, A. (2019). Diabetes Health Indicators Dataset. Kaggle.
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

World Health Organization. (2024). Diabetes overview.
https://www.who.int/news-room/fact-sheets/detail/diabetes

American Diabetes Association. (2023). Classification and Diagnosis of Diabetes.
https://diabetes.org/newsroom/american-diabetes-association-2023-standards-care-diabetes-guide-for-prevention-diagnosis-treatment-people-living-with-diabetes

National Institute of Diabetes and Digestive and Kidney Diseases.
https://www.niddk.nih.gov/health-information

Kivimäki & Tabák et al. (2014). Psychological Distress and Incidence of Type 2 Diabetes in High-Risk and Low-Risk Populations: The Whitehall II Cohort Study.
https://pmc.ncbi.nlm.nih.gov/articles/PMC4113169/

Hassan, et al. (2024). Prediction of Diabetes Using Data Mining and Machine Learning Algorithms: A Cross-Sectional Study
https://pubmed.ncbi.nlm.nih.gov/38359851/

National Institute of Diabetes and Digestive and Kidney Diseases. (2016). Risk factors for type 2 diabetes. U.S. Department of Health and Human Services, National Institutes of Health.
https://nih.gov/

Breiman, L. (2001). Random Forests. Machine Learning.
https://link.springer.com/article/10.1023/A:1010933404324

Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain.

https://pubmed.ncbi.nlm.nih.gov/11576808/