Kent State University

Ambassador Crawford College of Business and Entrepreneurship

Course: BA-64099-021

Instructor: Prof. Mostafa K. Ardakani, Ph.D.

Capstone Project in Business Analytics (Summer 2024)

Optimizing Diabetes Risk Prediction through Machine Learning and Advanced Data Analysis

Group - 2:

Sania Fatima

Lei Jin

Spandana Sodadasi

Abstract:

Diabetes imposes significant health burdens globally, affecting millions with escalating prevalence rates. This study investigates key predictors of diabetes risk using a comprehensive analytical framework encompassing Exploratory Data Analysis (EDA), Feature Selection, Model Building, and Model Testing. Utilizing MATLAB, nine critical features were identified from lifestyle, health, and demographic factors. The study applied 34 machine learning models across varying testing percentages (10%, 15%, 20%) using the Ohio Super Computer (OSC) to predict diabetes onset. Computational efficiency and predictive accuracy were evaluated, with results detailed through confusion matrix analyses across 9 experimental runs. The findings underscore the efficacy of machine learning in enhancing diabetes prediction, contributing crucial insights for proactive healthcare management strategies.

Key Words:

Diabetes Risk Prediction, Machine Learning, MATLAB, Ohio Super Computer (OSC), IQR, Computational Efficiency, Confusion Matrix etc.

Introduction:

Diabetes is a chronic condition that significantly impacts millions of individuals worldwide, leading to severe health complications and reduced life expectancy. [1] More than 133 million Americans have diabetes or prediabetes. As of 2019, 37.3 million people or 11.3% of the U.S. population had diabetes. More than 1 in 4 people over the age of 65 had diabetes. Nearly 1 in 4 adults with diabetes didn't know they had the disease. About 90% to 95% of diabetes cases are type 2 diabetes. [2] In 2019, 96 million adults, 38% of U.S. adults had prediabetes. [3] About 537 million adults across the world have diabetes. Experts predict this number will rise to 643 million by 2030 and 783 million by 2045.

Understanding the factors that contribute to the onset and progression of diabetes is crucial for developing effective prevention and management strategies. This research aims to analyze various lifestyle, health, and demographic factors to identify key predictors of diabetes risk. The study employs a comprehensive analytical approach that includes Exploratory Data Analysis (EDA), Feature Selection, Model Building, and Model Testing. MATLAB was the primary tool used for this analysis, enabling the selection of the top 9 features. The study also involved performing additional runs with varying testing percentages (10%, 15%, 20%) and different features selection criteria (5, 7, 9 features).

This research involves the application of multiple classification runs, with a list of 34 methods utilized for each type of run. The primary objective is to determine if these runs are sufficiently accurate to predict an individual's likelihood of developing diabetes based on factors such as General Health, Body Mass Index (BMI), High Blood Pressure (High BP), High Cholesterol, heart disease etc. Additionally, the study addresses the time required for each run, incorporating the analysis of all four types of results from the Test Confusion Matrix - True Positives, True Negatives, False Positives, and False Negatives. These results are then cumulatively compared across all 9 runs.

This report aims to elucidate why these results provide insights into the factors that contribute to a higher likelihood of developing diabetes, while also referencing other studies that have explored similar and different methods using deep machine learning. The subsequent section will discuss the data preparation process for training using modern methods and methodologies. Following this, the features, methods, and results will be presented in an organized manner to facilitate easy understanding and interpretation of the findings from this research report.

By leveraging machine learning techniques, we seek to enhance the accuracy and reliability of diabetes classification models, ultimately contributing to more effective diabetes prevention and management strategies.

Literature Review:

Diabetes, a chronic and pervasive disease, presents significant health challenges worldwide. [4] A. Roglic et al. (2016) revealed that diabetes causes millions of deaths annually and significantly increases the risk of heart disease, kidney failure, and other serious complications. Its prevalence has nearly doubled from 4.7% in 1980 to 8.5% in 2014, with low-and middle-income countries being hit hardest. This alarming rise necessitates the development of reliable prediction methods to mitigate the health crisis and economic burden associated with diabetes. [5] A. Negi et al. (2016) addressed the critical need for a robust diabetes prediction method validated across diverse datasets to enhance global applicability and reliability. Highlighting the chronic nature and global prevalence of diabetes, their research developed a novel approach using combined datasets and machine learning techniques, specifically SVM (Support Vector Machine), achieving a significant 72% accuracy in predicting diabetes.

Over the years, multiple studies have used computational intelligence techniques as one of the common strategies for predicting diabetes. [6] Mitushi Soni et al. (2020) highlighted the need for accurate early detection to manage diabetes effectively. The study analyzed the Pima Indian Diabetes Dataset using SVM, K-Nearest Neighbor, Decision Tree, Logistic Regression, Random Forest, and Gradient Boosting, finding that Random Forest outperformed other methods, achieving high prediction accuracy by leveraging key dataset features.

[7] M. K. Hasan et al. (2020) focused on enhancing diabetes prediction using machine learning classifiers, emphasizing the critical role of preprocessing, achieving a high AUC of 0.950 with their ensemble classifier. [8] T. Chauhan et al. (2021) examined various algorithms, noting the high accuracy of decision tree-based methods such as XGBoost and AdaBoost, and discussed the shift to deep learning methods for improved diabetes management. [9] U. Ahmed et al. (2022) discussed a predictive model for early diabetes detection using a fused machine learning approach combining SVM and ANN models, achieving a prediction accuracy of 94.87%. [10] S. A. Shampa et al. (2023) showed the effectiveness of boosting algorithms such as AdaBoost, CatBoost, Gradient Boost, and XGBoost in accurately predicting diabetes, with basic models like Random Forests and Decision Trees delivering promising results.

Predicting diabetes at its early stages is very important and helpful in improving healthcare analytics throughout the world in major ways, but it is also important to predict accurately. [11] K. Vijiya Kumar et al. (2019) aimed to develop a predictive model for early diabetes detection using the Random Forest algorithm, addressing the need for timely diagnosis to prevent complications, achieving over 90% accuracy. [12] A. Mujumdar et al. (2019) aimed to enhance diabetes classification accuracy using big data analytics and ML algorithms, achieving up to 97.2% accuracy by applying logistic regression and AdaBoost. [13] A. Mangal et al. (2022) utilized ML algorithms like Random Forest to achieve 99% accuracy in predicting diabetes based on medical symptoms, suggesting future applications for enhancing predictive accuracy across various health conditions. [14] Menaka. V et al. (2022) developed a framework for predicting Type I and Type II diabetes using machine learning algorithms, achieving up to 99% accuracy in testing with Decision Tree. In addition to achieving accuracy, it is crucial to comprehensively understand the factors influencing predictive accuracy. Researchers like [15] A. Anand et al. (2015) aimed to predict diabetes in Dehradun using lifestyle factors assessed via the Chi-squared Test and CART algorithm, identifying significant predictors such as blood pressure, eating habits, sleep patterns, family history, rice intake, and physical activity, achieving 75% accuracy with the CART model. [16] J. J. Padmini et al. (2020) aimed to develop a non-invasive diabetes diagnosis method using heartbeat rate and temperature sensors, showing higher accuracy and lower false classification rates than traditional methods.

Diabetes is a disease that affects many other organs in the body, making early prediction even more critical to reduce the risk of complications. [17] A. U. Naik et al. (2020) emphasized early diabetes detection to prevent vision loss, reduce ophthalmologists' workload, and enhance accuracy via automated systems, achieving 96.67% accuracy with SVM and 96% with ANN for diagnosing diabetic retinopathy. [18] E. S. Omoora et al. (2023) highlighted XGBoost's effectiveness in accurate diabetes diagnosis to prevent severe complications such as heart disease and kidney failure. Some researchers also focused on specific approaches to reduce diabetes impact, such as developing diet plans and health monitoring apps to enhance life quality.

[19] Peter G. Jacobs et al. (2014) explored machine learning's role in diabetes management, focusing on algorithms like k-nearest neighbors and fuzzy logic to optimize insulin dosing and improve glucose control. [20] C.H. Wu et al. (2014) developed ScasDia, a Java-based system integrating IT tools for diabetes self-care support, aiming to improve management through features like reminders and predictive modeling. [21] B. V. Baiju et al. (2019) aimed to improve diabetes prediction accuracy using the Disease Influence Measure (DIM) based on comprehensive feature analysis. [22] E. Sophiya et al. (2023) developed a machine learning model to predict the suitability of food items for diabetic patients based on their glycemic index (GI), achieving 99.5% accuracy. [23] B. VeerasekharReddy et al. (2023) developed an e-college nurse system using machine learning to monitor student health. The system uses BMI and CNNs to detect Type-2 diabetes and diabetic retinopathy, creating personalized health plans to promote healthy habits and early disease detection. [24] N. L. Fitriyani et al. (2019) proposed a Disease Prediction Model (DPM) for early detection of type 2 diabetes and hypertension using machine learning, achieving high accuracy with iForest, SMOTETomek, and ensemble learning. A mobile app provides real-time health updates, enhancing early health risk management.

Multiple researchers have tried to predict diabetes accurately using different machine learning methods and algorithms with various factors, but our research provides a comprehensive approach. Starting with feature selection, we employed five different methods like MRMR, chi2, ReliefF, Anova, and Kruskal to ensure reliable prediction. We tested 34 machine learning models through the Ohio Super Computer (OSC) to classify diabetes. This exhaustive approach fills the gaps left by prior research by combining the most reliable features with extensive model evaluation to determine the most accurate and efficient prediction methods. By calculating the training time for each model, providing insights into their speed and efficiency in producing results, and leveraging the results from confusion matrices, our study aims to pinpoint the most reliable and fastest method for accurate diabetes prediction.

Data Preparation:

The interest in diabetes dataset arises from the profound global impact of the disease, which affects millions worldwide and results in serious health complications and a reduced life expectancy.

The data is sourced from,
Kaggle: https://www.kaggle.com/datasets/prosperchuks/health-dataset. This dataset aims to identify key lifestyle, health, and demographic factors that influence the risk of developing diabetes. It aims to uncover crucial insights that can inform prevention and management strategies for this chronic disease.

The brief overview of the dataset is as follows:

Number of rows	12,000
Number of columns	13
Number of Categorical variables	11
Number of Numerical variables	2
Target Variable	Diabetes (1: YES, 0: NO)

Given the dataset with 12000 rows and 13 variables, the objective is to determine which factors are associated with diabetes. Here is a brief overview of the variables and how they might relate to diabetes:

1. **Age:** Older age is a well-established risk factor for diabetes.
2. **Sex:** There may be differences in diabetes prevalence between males and females.
3. **High Cholesterol:** Elevated cholesterol levels are linked to diabetes, often due to their association with metabolic (insulin) syndrome.
4. **BMI (Body Mass Index):** A higher BMI is strongly associated with an increased risk of diabetes.
5. **Smoking:** Smoking can contribute to the development of diabetes and its complications.
6. **History of Heart Disease:** A history of heart disease is often associated with diabetes due to shared risk factors.
7. **Physical Activity:** Regular physical activity can reduce the risk of developing diabetes.
8. **Fruit Consumption:** Eating fruit is generally associated with a healthier diet and a lower risk of diabetes.
9. **Vegetable Consumption:** Like fruit consumption, a diet rich in vegetables are linked to a lower risk of diabetes.
10. **Heavy Alcohol Consumption:** Excessive alcohol intake can increase the risk of developing diabetes.
11. **General Health:** Self-reported general health status may correlate with the presence of diabetes.
12. **High Blood Pressure (Hypertension):** Hypertension is a common condition that often coexists with diabetes.
13. **Diabetes (Output Variable):** The dependent variable indicates whether the individual has diabetes or not.

To analyze the factors influencing the likelihood of developing diabetes, statistical methods like classification have been utilized. This approach enables the assessment of the relationship between each independent variable and the probability of having diabetes, while adjusting for the effects of other variables in the model.

Methodology:

The nine runs of these methods were conducted using three different sets of variables, each of which appeared to significantly affect and correlate with the accuracy of predicting and determining the outcome.

Table-1: Feature Selection

S.No	Predictors	MRMR	Chi2	ReliefF	ANOVA	Kruskal	Median
1	Gen_Hlth	1	1	1	1	1	1
2	BMI	4	2	3	2	2	2
3	High_BP	5	3	7	3	3	3
4	High_Chol	3	4	5	4	4	4
5	Heart_Disease	6	6	4	5	5	5
6	Phy_Activity	8	7	11	6	6	7
7	Smoker	10	8	10	7	7	8
8	Veggies	7	9	12	8	8	8
9	Age	9	5	2	11	11	9
10	Fruits	11	10	9	9	9	9
11	Hvy_Alcohol	2	11	6	10	10	10
12	Sex	12	12	8	12	12	12

The table above categorizes variables into three sets for analysis, ensuring a balanced approach to feature selection. Out of 12 variables, they were divided into sets of five, seven and nine each consisting of a mix of numeric and categorical variables. Variables highlighted in the blue were deemed less important and thus excluded from the final set. This decision was based on median values across various feature selection methods, including MRMR, CHI2, ANOVA, RELIEFF, and KRUSKAL.

MRMR (Minimum Redundancy Maximum Relevance) aims to select features that are highly relevant to classification tasks while minimizing redundancy [25]. The chi-square test examines if there is a significant link between two categorical variables using observed versus expected frequencies in a contingency table, without assuming data distribution [26]. ANOVA (Analysis of Variance) tests if there are significant mean differences among three or more independent groups. Its used in healthcare, such as comparing average blood pressure levels across different treatment regimens, unlike the t-test, which compares means between two groups [27]**.** ReliefF feature scoring assesses feature importance by comparing value differences between nearest neighbor pairs. When a difference in feature values occurs between neighboring instances of the same class the feature score decreases [28]. The Kruskal-Wallis method is employed in the proposed approach for selecting significant features due to its computational efficiency and simplicity. It tests whether two or more groups have equal medians [29]. These five methods of feature selection facilitated the selection of variables based on their ranks, optimizing the identification of significant features for the analysis by the median of the given ranks.

Features with higher median values were considered less significant, leading to the removal of the last three variables from the model. This was done to prevent interference during runs on the supercomputer and to ensure accurate representation of results. By examining the results, we can evaluate which features are most influential in predicting outcomes and identify the optimal combination of features for each model. As a result, the third set incorporates the veggies and age attributes specifically for predicting diabetes. By refining the feature set, we aimed to optimize the model's performance and enhance the accuracy of diabetes predictions using OSC. This meticulous selection process helps to identify the most relevant predictors, providing valuable insights into the factors contributing to diabetes risk.

There are several core methods established and discussed in this report, summarized as follows:

- Tree-based machine learning methods have gained significant popularity in statistics and data science, often outperforming traditional methods. They have been successfully applied in areas such as biomarker discovery, estimation of causal effects, healthcare cost prediction, identification of key risk factors, and hospital performance evaluation. However, their adoption in health studies has been slower. To encourage their use in health research, we provide a primer on effectively applying tree-based methods to solve four important statistical problems [30].
- Discriminant analysis is employed when the data is normally distributed, whereas logistic regression is used for non-normally distributed data. An advantage of discriminant analysis is its ability to provide clear classification boundaries. In the context of my analysis, discriminant analysis was applied to datasets with normal distribution, while logistic regression was utilized for those that did not meet this criterion.
- The Naïve Bayes Classifier utilizes Bayes' theorem to combine prior knowledge with new information. This method offers several advantages, including a straightforward algorithm, high accuracy, and efficiency in handling large datasets [31].
- An advantage of Support Vector Machine (SVM) is its strong theoretical foundation and practical effectiveness in handling complex problems such as nonlinearity, high dimensionality, and local minima. Additionally, by using a radial basis kernel function and selecting appropriate parameters, SVM can achieve very high classification accuracy, making it highly suitable for both binary and multi-class classification tasks [32].
- kNN is a simple yet effective classification method and has proven to be one of the most effective techniques for text categorization and classification problems on the Reuters corpus of newswire stories, it motivates to develop a model to enhance kNN's efficiency while maintaining its high classification accuracy. This approach leverages kNN's strengths in classification to achieve reliable results in real world applications.

- Ensemble learning combines multiple classifiers to enhance predictive performance by leveraging diverse models. By aggregating predictions, it reduces generalization errors, particularly when individual models are independent. This approach treats the ensemble as a unified model, boosting accuracy by combining various weak learners into a stronger one. Ensemble methods are widely used in real-world data mining due to their ability to improve accuracy over single classifiers by integrating different algorithms.
- Neural networks excel at sorting data into different groups or features. Classification neural networks, which assign a single output response to each input pattern, are particularly powerful when combined with various predictive neural networks in hybrid systems, enhancing their ability to handle complex tasks and improve accuracy [33].

Table-2: Results of 9 Experimental Runs defining Holdout, Test Size, and Feature Sets

SET-1 [5 Features]			
Runs	Run - 1	Run - 4	Run - 7
Holdout	15%	15%	15%
Test	10%	15%	20%
SET-2 [7 Features]			
Runs	Run - 2	Run - 5	Run - 8
Holdout	15%	15%	15%
Test	10%	15%	20%
SET-3 [9 Features]			
Runs	Run - 3	Run - 6	Run - 9
Holdout	15%	15%	15%
Test	10%	15%	20%

Table 2 presents the execution of nine models, with percentages allocated based on the segregation of features. This detailed breakdown allows for a comprehensive understanding of how different feature sets contribute to the performance of each model. This approach ensures a thorough analysis of feature importance and model efficiency, leading to more accurate and reliable predictions.

Table-3: 34 Executed ML Models

S.No	**Models**
1	Tree (Fine Tree)
2	Tree (Medium Tree)
3	Tree (Coarse Tree)
4	Linear Discriminant
5	Quadratic Discriminant
6	Binary GLM Logistic Regression

7	Efficient Logistic Regression
8	Efficient Linear SVM
9	Naïve Bayes (Gaussian)
10	Naïve Bayes (Kernel)
11	SVM (Linear)
12	SVM (Quadratic)
13	SVM(Cubic)
14	SVM (Fine Gaussian)
15	SVM (Medium Gaussian)
16	SVM (Coarse Gaussian)
17	KNN (Fine)
18	KNN (Medium)
19	KNN (Coarse)
20	KNN (Cosine)
21	KNN (Cubic)
22	KNN (Weighted)
23	Ensemble (Boosted Trees)
24	Ensemble (BaggedTrees)
25	Ensemble (Subspace Discriminant)
26	Ensemble (Subspace KNN)
27	Ensemble (RUSBoosted Trees)
28	Neural Network (Narrow)
29	Neural Network (Medium)
30	Neural Network (Wide)
31	Neural Network (Bilayered)
32	Neural Network (Trilayered)
33	Kernel (SVM)
34	Kernel (Logistic Regression)

Table 3 shows the execution of 34 models on the dataset, offering several advantages. This comprehensive evaluation identifies the best-performing models, enhances understanding of robustness and stability, and aids in feature selection. By comparing multiple models, it increases the likelihood of finding the optimal model that balances bias and variance. Additionally, running numerous models provides performance benchmarks, and enables the creation of ensemble methods for improved accuracy. It also reveals deeper insights into the dataset, ensuring the selected model generalizes well to new data.

Results:

✓ **CPU Time:**

The analysis of CPU processing time for 34 different machine learning models was a critical component of our study to predict diabetes, emphasizing the importance of computational efficiency in practical healthcare applications. The models were evaluated based on their median CPU times to identify the fastest and most resource-efficient algorithms. Notably, Efficient Logistic Regression (Model 7), Tree (Coarse Tree) and SVM (Model 3) demonstrated the lowest median processing times, at 1.18 and 1.28 seconds respectively. Such rapid performance underscores their potential for real-time applications where quick decision-making is essential.

On the other hand, more complex models, such as Neural Networks, SVM and Kernel based methods, exhibited significantly higher median times. For example, the Neural Network (Tri-layered) Model reached a median processing time of up to 95.05 seconds. While these models might offer superior accuracy, their extended processing times can be a limitation in scenarios requiring immediate results. By calculating and comparing the training times for each model, our study provided valuable insights into their speed and efficiency in producing results. This analysis is crucial in balancing the trade-offs between computational time and prediction accuracy. This assists in selecting the most suitable model for diabetes prediction that meets both accuracy requirements and resource constraints.

Models	**7**	**3**	**8**	**4**	**5**	**9**	**2**	**6**
Median (Sec)	1.18	1.28	1.52	1.84	1.87	1.99	2.72	2.78
IQR (Sec)	0.9348	2.53483	1.16724	1.34284	3.6756	1.8766	1.2648	2.5822

1	**10**	**14**	**15**	**11**	**17**	**16**	**18**	**19**
6.69	8.7	28.01	34.01	34.39	39.38	40.61	41.02	41.88
1.9572	4.4673	7.687	15.382	82.623	15.021	5.604	13.799	15.703

20	**22**	**21**	**23**	**24**	**25**	**27**	**26**	**28**
42.58	43.73	43.87	46.15	49.28	49.52	53.54	53.88	62.47
14.139	12.57	12.609	11.831	9.364	13.468	14.602	12.912	16.067

29	**31**	**32**	**33**	**34**	**30**	**12**	**13**
69.76	82.15	95.05	99.35	100.98	103.95	290.99	296.92
26.881	37.724	43.758	31.318	32.189	12.9	112.12	11.91

✓ **Performance Assessment:**

Ensuring the robustness of model results is paramount after conducting rigorous testing across a diverse array of models. Our investigation into CPU times revealed significant variability in processing durations across the models tested. However, to accurately gauge the efficacy of these models, a meticulous examination of their confusion matrices was necessary. This comprehensive evaluation spanned three distinct test scenarios – 10% with 1200 predictions, 15% with 1800 predictions, and 20% with 2400 predictions, encompassing both positive and negative test cases. Consistently, these tests maintained a distribution of approximately 43.5% positive and 56.5% negative outcomes, ensuring a robust evaluation across diverse feature sets.

The median values extracted from these tests, partitioned into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) which were subsequently standardized into Z-scores to assess each model's performance objectively. The weights adopted in this evaluation highlighted the critical importance of accurately identifying false negatives (FN), as misclassifying a diabetic patient as non-diabetic could lead to severe health implications, such as delayed treatment and disease progression. Therefore, FN was assigned the highest weight of 0.4. This was followed by weights of 0.3 for false positives (FP), 0.2 for true positives (TP), and 0.1 for true negatives (TN), ensuring a balanced assessment that prioritized key metrics.

Upon analyzing the aggregated weights derived from the confusion matrix outcomes, Kernel (Logistic Regression) emerged as the top performer with a weight of 0.234. This model, alongside Kernel (SVM) which achieved a weight of 0.197, demonstrated comprehensive and effective performance across all evaluated metrics. In contrast, models such as SVM (Cubic) and Ensemble (Bagged Trees) exhibited lower total weights of 0.160 and 0.131 respectively, indicating less optimal performance balances across the evaluated metrics. Notably, Ensemble (Subspace KNN) registered the lowest weight at 0.073, suggesting potential limitations in accurate diabetes prediction despite potential advantages in other areas.

Evaluating both total weight (indicating predictive performance) and CPU time (indicating computational efficiency) provides a holistic approach to identifying the optimal model. For instance, while Kernel (Logistic Regression) excels with a high total weight of 0.234, signifying robust predictive capability, it does so with a longer CPU time of 100.98 seconds. In comparison, models like KNN (Fine) and Ensemble (Bagged Trees) offer lower CPU times of 39.38 and 49.28 seconds, respectively. However, their lower total weights of 0.088 and 0.131 suggest potential trade-offs in predictive reliability compared to Kernel (Logistic Regression). This comprehensive assessment emphasizes the need to balance accuracy and computational efficiency when selecting the most effective model for diabetes prediction.

	Sec.	Median values from the percentage sheets				Weight	0.2	0.1	0.3	0.4	
Method	**CPU Time**	**TP**	**TN**	**FP**	**FN**		**TP**	**TN**	**FP**	**FN**	**Total_Weight**
Kernel (Logistic Regression)	5.6	24.2	41.9	14.7	19.2		-0.22	-1.15	1.12	0.15	0.234
Kernel (SVM)	5.5	25.6	42.4	14.5	17.9		0.00	-1.00	1.06	-0.05	0.197
SVM(Cubic)	16.5	16.2	49	7.5	29.2		-1.48	1.01	-1.04	1.66	0.160
Ensemble (BaggedTrees)	2.7	28.5	43.4	14.3	15.1		0.46	-0.69	1.00	-0.47	0.131
Ensemble (RUSBoosted Trees)	2.9	32.7	39.8	16.7	11.1		1.12	-1.79	1.71	-1.08	0.126
KNN (Fine)	2.2	10	53	3.8	34.7		-2.46	2.22	-2.14	2.50	0.088
Ensemble (Subspace KNN)	2.9	0.4	56.4	0.1	43.2		-3.97	3.26	-3.25	3.79	0.073
Neural Network (Wide)	5.8	28.7	43	13.4	15.2		0.49	-0.82	0.73	-0.46	0.050
Tree (Fine Tree)	0.4	30.5	42.8	14.1	13.3		0.77	-0.88	0.94	-0.75	0.048
SVM (Fine Gaussian)	1.6	29.9	42.9	13.9	13.7		0.68	-0.85	0.88	-0.69	0.039
Tree (Coarse Tree)	0.1	24.8	44.9	11.5	18.8		-0.13	-0.24	0.16	0.09	0.033
Naïve Bayes (Gaussian)	0.1	26.8	44.9	11.8	16.7		0.19	-0.24	0.25	-0.23	-0.004
SVM (Medium Gaussian)	1.9	29.1	44.1	12.3	15		0.55	-0.48	0.40	-0.49	-0.014
Neural Network (Medium)	3.6	29.6	44.1	12.8	14		0.63	-0.48	0.55	-0.64	-0.015
KNN (Coarse)	2.2	28.1	44.8	11.7	16		0.39	-0.27	0.22	-0.34	-0.018
Neural Network (Trilayered)	5.2	29.6	43.4	13	14		0.63	-0.69	0.61	-0.64	-0.018
Quadratic Discriminant	0.1	25.4	46.3	10.7	18.1		-0.03	0.19	-0.08	-0.02	-0.019
Tree (Medium Tree)	0.1	29.4	43.6	12.9	14.1		0.60	-0.63	0.58	-0.63	-0.021
Neural Network (Bilayered)	4.5	29.3	44.2	12.4	14.3		0.58	-0.45	0.43	-0.60	-0.039
Naïve Bayes (Kernel)	0.6	27.7	45	11.5	15.9		0.33	-0.21	0.16	-0.35	-0.048
KNN (Weighted)	2.4	21.2	48.6	7.9	22.7		-0.69	0.89	-0.92	0.68	-0.054
Neural Network (Narrow)	3.4	29.6	44.4	12.3	13.9		0.63	-0.39	0.40	-0.66	-0.056
SVM (Linear)	2.2	28.8	44.9	11.8	14.8		0.50	-0.24	0.25	-0.52	-0.057
SVM (Coarse Gaussian)	2.1	28.7	45.1	11.6	15		0.49	-0.18	0.19	-0.49	-0.059
Efficient Linear SVM	0.1	28.9	44.8	11.8	14.7		0.52	-0.27	0.25	-0.54	-0.063
KNN (Medium)	2.2	19.5	50.4	6	25.3		-0.96	1.43	-1.48	1.07	-0.065
Efficient Logistic Regression	0.1	28.1	45.7	11.2	15.5		0.39	0.00	0.07	-0.41	-0.066
Linear Discriminant	0.1	28.1	45.6	11.2	15.5		0.39	-0.03	0.07	-0.41	-0.069
KNN (Cubic)	2.4	19.6	50.6	5.9	25.2		-0.95	1.49	-1.51	1.06	-0.071
KNN (Cosine)	2.3	20.2	50.1	6.4	24.4		-0.85	1.34	-1.37	0.94	-0.071
Binary GLM Logistic Regression	0.2	28.1	45.7	11.2	15.4		0.39	0.00	0.07	-0.43	-0.072
Ensemble (Boosted Trees)	2.5	29.6	44.5	12	14		0.63	-0.36	0.31	-0.64	-0.074
Ensemble (Subspace Discriminant)	2.6	26.9	46.2	10.3	16.7		0.20	0.16	-0.20	-0.23	-0.096
SVM (Quadratic)	16.5	26.8	46.8	9.7	17.1		0.19	0.34	-0.38	-0.17	-0.111

Conclusion:

In conclusion, this study underscores the critical importance of accurate diabetes prediction through comprehensive machine learning models. By evaluating 34 different models based on their CPU processing times and performance metrics derived from confusion matrices, we aimed to identify the most efficient and reliable predictors of diabetes risk. Our approach involved meticulous analysis of diverse datasets encompassing crucial health indicators such as BMI, blood pressure, and cholesterol levels. Through rigorous testing scenarios and feature selection techniques, we highlighted Kernal (Logistic Regression) and Kernal (SVM) as top performing models, demonstrating robust predictive capabilities with weights of 0.234 and 0.197 respectively.

The significance of this research lies in its potential to enhance early detection and management of diabetes, thereby mitigating its severe health implications worldwide. By leveraging advanced machine learning methodologies, we not only validated the effectiveness of these models in predicting diabetes onset but also provided insights into optimizing computational efficiency without compromising accuracy. This research contributes to the ongoing efforts in healthcare analytics by offering scalable and precise tools for healthcare providers and policymakers to implement proactive measures against diabetes, ultimately improving patient outcomes and reducing healthcare costs.

References:

[1] CDC. (2024, May 21). National Diabetes Statistics Report. Diabetes. https://www.cdc.gov/diabetes/php/data-research/

[2] CDC. (2024, May 22). Methods for the National Diabetes Statistics Report. Diabetes. https://www.cdc.gov/diabetes/php/data-research/methods.html

[3] Home, Resources, diabetes, L. with, Acknowledgement, FAQs, Contact, & Policy, P. (n.d.). IDF Diabetes Atlas | Tenth Edition. https://diabetesatlas.org/#:~:text=537%20million%20adults%20(20%2D79

[4] Roglic, G. (2016). WHO Global report on diabetes: A summary. International Journal of Noncommunicable Diseases, 1(1), 3. https://doi.org/10.4103/2468-8827.184853

[5] Negi, A., & Jaiswal, V. (2016, December). A first attempt to develop a diabetes prediction method based on different global datasets. In *2016 fourth international conference on parallel, distributed and grid computing (PDGC)* (pp. 237-241). IEEE.

[6] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 9(09), 2278-0181.

[7] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531.

[8] Chauhan, T., Rawat, S., Malik, S., & Singh, P. (2021, March). Supervised and unsupervised machine learning based review on diabetes care. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 581-585). IEEE.

[9] Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. *IEEE Access*, *10*, 8529-8538.

[10] Shampa, S. A., Islam, M. S., & Nesa, A. (2023, June). Machine Learning-based Diabetes Prediction: A Cross-Country Perspective. In *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)* (pp. 1-6). IEEE.

[11] VijiyaKumar, K., Lavanya, B., Nirmala, I., & Caroline, S. S. (2019, March). Random forest algorithm for the prediction of diabetes. In *2019 IEEE international conference on system, computation, automation and networking (ICSCAN)* (pp. 1-5). IEEE.

[12] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.

[13] Mangal, A., & Jain, V. (2022, December). Performance analysis of machine learning models for prediction of diabetes. In *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)* (pp. 1-4). IEEE.

[14] Menaka, V., Likitha, V., Shravya, M., & Pari, R. (2022, August). Accurate Prediction of Type 1 and Type 2 Diabetes. In *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)* (pp. 1117-1121). IEEE

[15] Anand, A., & Shakti, D. (2015, September). Prediction of diabetes based on personal lifestyle indicators. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)* (pp. 673-676). IEEE.

[16] Padmini, J. J., Kavya, R., Ilakkiya, B., Gurupriya, R., & Monika, P. (2020, January). A Non-invasive Way Of Diagnosing Diabetes Based On The Heart Beat Rate. In *2020 Fourth International Conference on Inventive Systems and Control (ICISC)* (pp. 337-341). IEEE.

[17] Naik, A. U., & Kulkarni, R. K. (2020, June). Artificial neural network-based detection of diabetes and its effects on vision-A survey. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1113-1118). IEEE.

[18] Omoora, E. S., Altaweil, H. A., Nagem, T., & Bozed, K. A. (2023, December). Diabetes Mellitus Prediction Based on Machine Learning Techniques. In *2023 IEEE 11th International Conference on Systems and Control (ICSC)* (pp. 225-231). IEEE.

[19] Jacobs, P. G., Herrero, P., Facchinetti, A., Vehi, J., Kovatchev, B., Breton, M., ... & Mosquera-Lopez, C. (2023). Artificial intelligence and machine learning for improving glycemic control in diabetes: best practices, pitfalls and opportunities. *IEEE reviews in biomedical engineering*.

[20] Wu, C. H. (2014, November). A patient-centered self-care support system for diabetics. In *2014 IEEE 11th International Conference on e-Business Engineering* (pp. 298-302). IEEE.

[21] Baiju, B. V., & Aravindhar, D. J. (2019, April). Disease influence measure based diabetic prediction with medical data set using data mining. In *2019 1st international conference on innovations in information and communication technology (ICIICT)* (pp. 1-6). IEEE.

[22] Sophiya, E., & Vidyasekaran, H. (2023, December). Impact of Diet and Nutritional Factors in Controlling Type 2 Diabetes Through Machine Learning. In *2023 International Conference on Next Generation Electronics (NEleX)* (pp. 1-7). IEEE.

[23] VeerasekharReddy, B., Thatha, V. N., Kiran, G. U., Shareef, S. K., RajaSekharReddy, N. V., & Raju, Y. R. (2023, November). Diet Recommendation System for Human Health Using Machine Learning. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1-5). IEEE.

[24] Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *Ieee Access*, *7*, 144777-144789.

[25] Cha, S. (2024, January 2). Unveiling MRMR: Maximizing Relevance, Minimizing Redundancy in Feature Selection. Medium. https://medium.com/@shinkookcha_39651/unveiling-mrmr-maximizing-relevanceminimizing-redundancy-in-feature-selection-b07e21cdd88e

[26] ML | Chi-square Test for feature selection. (2018, December 20). GeeksforGeeks. https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/

[27] ANOVAs with Healthcare Data | Data Science for Health Informatists. (n.d.). Book.datascience.appliedhealthinformatics.com. Retrieved July 12, 2024, from https://book.datascience.appliedhealthinformatics.com/docs/Ch5/anovas

[28] Relief (feature selection). (2021, January 17). Wikipedia. https://en.wikipedia.org/wiki/Relief_(feature_selection)

[29] Ali Khan, S., Hussain, A., Basit, A., & Akram, S. (2014). Kruskal-Wallis-Based Computationally Efficient Feature Selection for Face Recognition. The Scientific World Journal, 2014, 1–6. https://doi.org/10.1155/2014/672630

[30] Hu, L., & Li, L. (2022). Using Tree-Based Machine Learning for Health Studies: Literature Review and Case Series. 19(23), 16080–16080. https://doi.org/10.3390/ijerph192316080

[31] Wibawa, A. P., Kurniawan, A. C., Murti, D. M. P., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. International Journal of Recent Contributions from Engineering, Science & IT (IJES), 7(2), 91. https://doi.org/10.3991/ijes.v7i2.10659

[32] Zhang, Y. (2012). Support Vector Machine Classification Algorithm and Its Application. Communications in Computer and Information Science, 308, 179–186. https://doi.org/10.1007/978-3-642-34041-3_27

[33] Baughman, & Liu, Y. (1995). Classification: fault diagnosis and feature categorization. In Elsevier eBooks (pp. 110–171). https://doi.org/10.1016/b978-0-12-083030-5.50009-6