Asian Journal of Statistical Sciences
Vol. 5, No-1, (2025), pp. 87-96

ARF INDIA
Academic Open Access Publishing
Journal homepage link:
http://arfjournals.com/ajss

Optimized Machine Learning Models for Enhanced Diabetes Prediction through Hyperparameter Tuning Approach

[a]Srinivasulu Akasam, [b]R. Rajasekar, [c]Praveen Kumar Poola, [d]Harika D, [e]Ch.Praveen Kumar, [f]Y.Vijayakumar, [g]Garlapati Narayana, [h]Murali Gundagani, [i]Jagadish V. Tawade

[a,g]*Geethanjali College of Engineering and Technology,Cheeryal-501301,India,*
[b]*Alliance University, Bengaluru - 562106, Karnataka, India,*
[c]*Malla Reddy Vishwavidyapeeth, Hyderabad-500055,*
[d]*Villa Marie Degree College for Women,Hyderabad-500082,India,*
[e]*GITAM University, Rudraram - 502329, Telangana, India,*
[f]*Anurag University, Venkatapur - 500088, Telangana, India,*
[g]*Chaitanya Bharathi Institute of Technology, Hyderabad - 500075, Telangana, India,*
[h]*Geethanjali College of Engineering and Technology,Cheeryal-501301,India,*
[i]*Department of Mathematics, Vishwakarma University, Pune-411048, Maharashtra, INDIA*

ARTICLE HISTORY
Compiled September 26, 2025

Received 08 May 2025; Accepted 10 July 2025

ABSTRACT
This study presents a comparative analysis of various machine learning models applied to a classification problem, both before and after hyperparameter tuning. Models including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, and XGBoost were evaluated. Performance metrics such as accuracy, precision, recall, and F1-score were used to assess each model. The results indicate that hyperparameter tuning significantly enhances the performance of certain models, particularly Decision Tree, Random Forest, and Gradient Boosting. Gradient Boosting outperformed all other models post-tuning with an accuracy of 0.80 and F1-score of 0.75.Models like KNN and SVM exhibited minimal improvements after tuning. Ensemble methods generally achieved better results compared to individual models. The study highlights the importance of fine-tuning model parameters to optimize results. Accuracy and F1-score suggest a consistent performance boost for tuned ensemble algorithms. These findings provide valuable insights for selecting and optimizing models in machine learning tasks.

KEYWORDS
Diabetes Prediction, Machine Learning Models, Performance Metrics, Grid Search Optimization, Hyperparameter Tuning.

1. Introduction

The disease arises due to impaired insulin production or ineffective use of insulin by the body, disrupting glucose regulation. This results in abnormal blood sugar levels

CONTACT Murali Gundagani Email: muraligundagani@gmail.com

and affects overall metabolic balance. It is similar to a faulty thermostat that fails to maintain a stable internal environment [2].

In 1991, the prevalence of diabetes in Qatar was recorded at 3%, with Qatari women showing a slightly higher rate of 4%. By 2006, these figures had risen sharply, reaching 17.5% in the general population and 18% among women [7] On a global scale, diabetes cases continued to grow, increasing from 393,000,000 in 2011 to 415,000,000 by 2013 [4,5]. This rising trend is closely linked to lifestyle changes, particularly unhealthy eating habits and reduced physical activity. Research supports these associations, including studies examining bone health in individuals with diabetes [8] and dietary behaviors among 500 diabetic patients [9]

Despite extensive research, Qatar lacks comprehensive studies on machine learning-based diabetes prediction. Previous investigations have explored prediabetes risk factors in 7,268 Qatari nationals [10], developed glucose metabolism models using Qatar Biobank data [11], and identified 25 key diabetes risk markers through machine learning [13]. Additionally, Dual X-ray Absorptiometry has been used to assess bone health in diabetic patients [15]

To bridge this gap, our study evaluates six advanced machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, XGBoost, and AdaBoost—to develop accurate diabetes prediction models. This research responds to the urgent need for improved diagnostic tools as diabetes continues to rise globally [13]. Qatar, where diabetes is a leading cause of mortality and a major strain on healthcare resources, serves as a critical focus area. However, the findings also address the broader global diabetes epidemic, emphasizing the necessity of innovative technological solutions for early detection and prevention.

This research examines the application of Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forest for predicting diabetes, offering a detailed comparison to identify the highest-performing model. Various machine learning approaches have been utilized in diabetes prediction, with each method presenting distinct advantages and drawbacks. Singh et al. [4] developed standard errors, hazard rate functions, and system reli- ability measures while concentrating on Bayesian estimation in parallel systems with k components using load sharing. Tian et al.[5] used a recursive approach to assess the reliability of k-of-n systems with different performance levels, taking into account components that are evenly and independently distributed. A thorough closed-form statement for the lifetime dependability of load-sharing k-out-of-n hybrid redundant systems was presented by Huang and Xu [6]. These systems make use of m components that are initially configured as active units. Depending on how a task is carried out, these components switch between processing and wait stages, with each state having an arbitrary failure distribution. They also arrived at a formula for lifetime reliability for these hybrid redundant systems.

2. Model Description

2.1. Logistic Regression

Logistic Regression is a prominent statistical model for binary classification tasks, including diabetes prediction. Its straightforwardness and interpretability make it a preferred choice in medical diagnostics [17]. Research has demonstrated its efficacy in predicting diabetes using demographic and clinical data [18].

2.2. Support Vector Machines (SVM)

SVMs are reliable classifiers that can handle non-linear boundaries and high-dimensional regions. They have been effectively used to a number of medical prediction scenarios, such as the diagnosis of diabetes [19]

2.3. Decision Trees

Decision Trees offer intuitive models that emulate human decision-making processes, enhancing their interpretability. They have been extensively used in medical applications due to their ability to process both numerical and categorical data [20]. Studies have highlighted their effectiveness in classifying diabetic and non-diabetic individuals based on various health metrics [21].

2.4. Random Forest

Random Forest, an ensemble learning technique, improves the predictive accuracy of decision trees by constructing multiple trees and aggregating their outputs. Its robustness and capacity to handle large, high-dimensional datasets make it well-suited for diabetes prediction [22]. Comparative research often shows Random Forest outperforming other models in accuracy and reliability [23].

3. Methods

This section details the data collection, preprocessing, and machine learning analysis for diabetes prediction.

3.1. Data Collection

The study utilizes the Diabetes Database, an esteemed dataset that encompasses diagnostic measurements and demographic details of female patients of Pima Indian descent [24]. This dataset consists of 768 records, each characterized by 8 attributes, including age, glucose levels, and blood pressure. The choice of this dataset is motivated by its extensive use in diabetes research and its relevance to the study's objective.

3.2. Data Preprocessing

Data preprocessing is a crucial phase to ensure the quality and integrity of the data before model implementation. The steps involved are:

Handling Missing Values: Missing data is imputed to maintain dataset completeness. For continuous variables, the mean value is utilized, while categorical variables are imputed using the mode. This approach mitigates potential biases introduced by missing data [25].

Feature Scaling: Feature scaling ensures that all features are on a comparable scale by standardizing the range of feature values. This step is crucial for models sensitive to feature magnitudes, such as Support Vector Machines and K-Nearest Neighbors [26].

Encoding Categorical Variables: Categorical variables are converted into numerical format using techniques such as one-hot encoding. This transformation enables machine learning models to process categorical data effectively [25].

3.3. Model Implementation

The methodology includes

3.3.1. Model Development Phase

The machine learning algorithms undergo training on a carefully selected 70% partition of the complete dataset. This substantial training corpus enables the models to identify complex patterns and establish robust decision boundaries, while ensuring adequate data remains for subsequent validation [26]. The training process focuses on parameter optimization to maximize predictive performance before deployment.

3.3.2. Validation and Testing Protocol

A distinct 30% portion of the original dataset serves as an independent test set, completely withheld during the training phase. This testing framework evaluates model effectiveness through multiple quantitative measures: classification accuracy (overall prediction correctness), precision (positive predictive value), recall (sensitivity), and the F1-score (balanced measure combining precision and recall) [26]. These metrics collectively provide a comprehensive assessment of each model's diagnostic capabilities.

3.3.3. Performance Assessment Methodology

confusion matrices offer a detailed breakdown of prediction outcomes, categorizing results into four distinct classes: correctly identified positive cases (true positives), incorrectly classified negative instances (false positives), accurately recognized negative samples (true negatives), and misclassified positive examples (false negatives) [27]. This dual approach enables both threshold-dependent and absolute performance evaluation.

3.4. Hyperparameter Tuning

To optimize model performance, hyperparameter tuning is performed using GridSearchCV, a robust cross-validation technique. The process involves:

3.4.1. Defining Parameter Grid:

For each model, a grid of potential hyperparameter values is defined. This grid includes various combinations of parameters that could affect model performance.

3.4.2. Grid Search:

GridSearchCV iterates through all possible combinations of hyperparameters, evaluating model performance for each set. This exhaustive search ensures the identification of the optimal hyperparameter configuration that enhances model accuracy and generalizability.

3.4.3. Optimization and Evaluation:

The best hyperparameter values identified and applied to the models. The optimized models are evaluated on the test set to determine their refined performance metrics.

3.5. Advanced Model Techniques

In addition to the basic models, advanced techniques such as XGBoost and AdaBoost are employed to further enhance prediction accuracy. The steps include:

3.5.1. XGBoost Implementation:

XGBoost is applied with specific hyperparameters optimized using GridSearchCV. This model utilizes gradient boosting to improve prediction accuracy by iteratively correcting errors made by previous models.

3.5.2. Comparison and Analysis:

The performance of advanced models is compared with basic models to assess improvements in accuracy, precision, recall, and F1-score.

4. Analysis and Results

To visualize the distribution of diabetes cases, a pie chart and Bar chart can be plotted, representing the proportion of individuals with and without diabetes (0: No Diabetes, 1: Diabetes). This helps in understanding the class distribution in the dataset.

Figure 1, pie chart shows that 65.1% of individuals do not have diabetes (Outcome = 0), while 34.9% have diabetes (Outcome = 1).The bar chart confirms this distribution, with a higher count of non-diabetic cases compared to diabetic cases.This highlights a class imbalance in the dataset, which may need to be addressed during model training.

To analyze the distribution and spread of different variables in diabetes prediction, a box plot can be used. It helps visualize the median, quartiles, and potential outliers, providing insights into variations and possible differences between individuals with and without diabetes.

Figure 2, the Age plot indicates that the majority of the dataset is concentrated between the ages of 25 and 40.The density of points above Q2 in the Insulin, Age and Diabetes Pedigree Function plots highlights a significant presence of outliers indicating high variability.The width of the boxes in the Pregnancies, Glucose, and Age plots suggests a wide distribution of the data.Glucose, Blood Pressure, and BMI show a more consistent distribution with fewer outliers.The varying range of values across features emphasizes the importance of feature scaling to ensure uniformity in model performance.

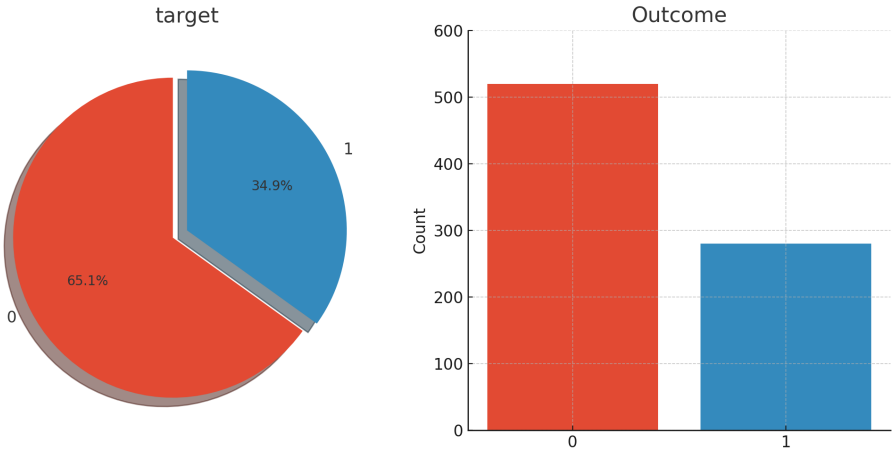

Figure 1. Pie chart and Bar chart

Machine Learning Predictive models In this section, we'll start by building and evaluating ML models such as DT, RF, Logistic Regression, and Support Vector Machine. These models will serve as baselines for comparison with more complex models later in the analysis.

Table 1. Metrics of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
LR	0.79	0.78	0.68	0.73
SVM	0.68	0.69	0.54	0.61
KNN	0.65	0.64	0.51	0.56
DT	0.69	0.64	0.56	0.59
RF	0.77	0.71	0.66	0.68
Gradient Boosting	0.76	0.75	0.64	0.69
XGBOOST	0.73	0.67	0.61	0.64

Table 1,When selecting a model, it is important to consider the specific requirements of the application. For diabetes prediction, prioritizing higher recall is essential to ensure that as many positive cases as possible are correctly identified. Logistic

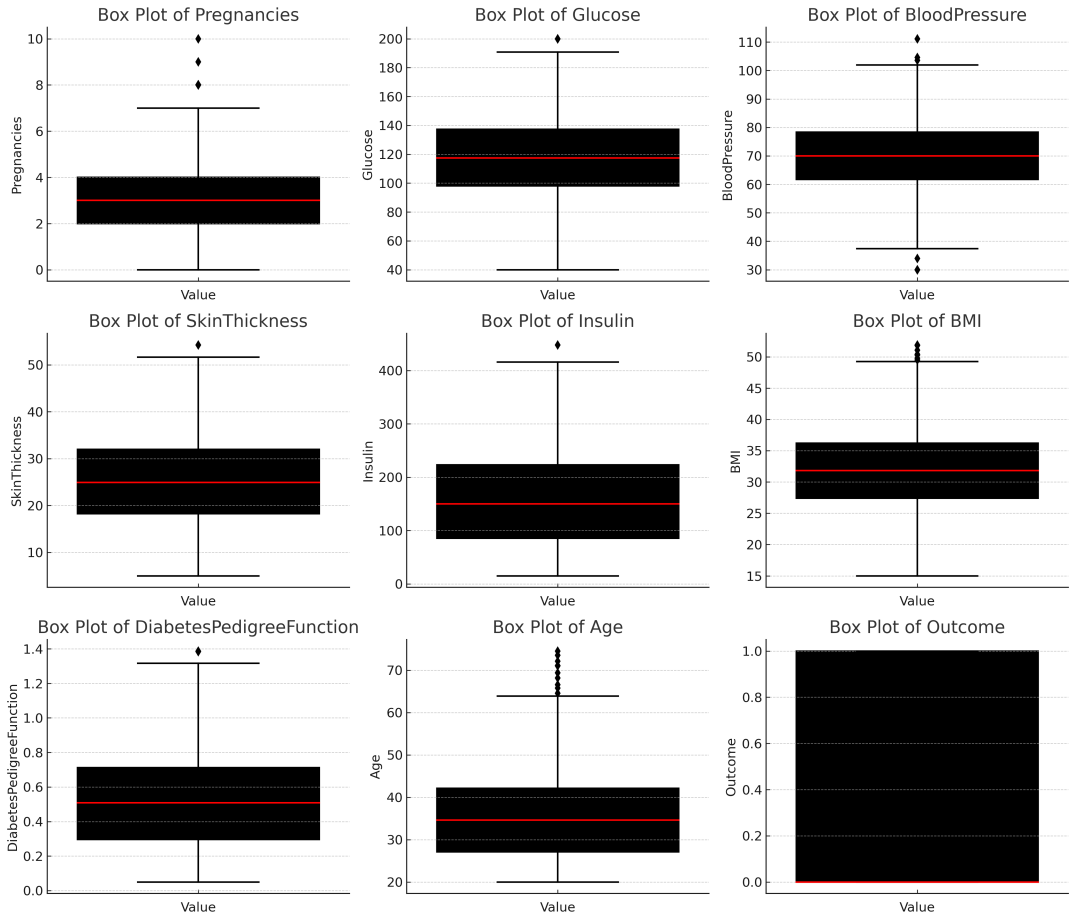

Figure 2. Box plots of all features including Outcome

regression has the highest accuracy (0.79) and the highest recall (0.68), which indicates it is better at identifying true positive cases of diabetes. Additionally, it has the highest F1 score (0.73), which balances precision and recall, making it the most reliable model in this context.

HYPER PARAMETER TUNING

Grid Search CV is a method used to find the optimal parameter values from a specified set of possible parameters. It performs cross-validation to evaluate the performance of a model for each combination of parameters. After identifying the best set of parameters, it can be used to make more accurate predictions.

Table 2, Gradient Boosting achieved the highest accuracy (0.80) and F1-score (0.75), indicating strong predictive performance. Logistic Regression and Random Forest followed closely with accuracies of 0.79 and F1-scores of 0.73 and 0.72, respectively. XGBoost showed moderate results, while SVM and KNN had lower scores. Decision Tree improved after tuning but still lagged behind ensemble models. Overall, models with higher accuracy also tended to have higher F1-scores, reflecting balanced performance.

Comparative analysis

This performance assessment examines ML algorithms LR, SVM, DT, RF, Gra-

Table 2. Metrics of Machine Learning Models after Hyper Parmeter Tuning

Model	Accuracy	Precision	Recall	F1-score
LR	0.79	0.78	0.68	0.73
SVM	0.68	0.69	0.54	0.61
KNN	0.65	0.64	0.51	0.56
DT	0.72	0.65	0.6	0.63
RF	0.79	0.78	0.67	0.72
Gradient Boosting	0.80	0.84	0.68	0.75
XGBOOST	0.73	0.65	0.61	0.64

dient Boosting, Adaboost and XGBoost for diabetes prediction, with comparative benchmarking against prior research .The evaluation reveals that LR implementation achieves superior performance across all evaluation metrics.While SVM and DT models show limitations in recall and F1-measure, ensemble methods including RF and Gradient Boosting demonstrate consistent performance gains in both accuracy and F1-scores. The XGBoost classifier exhibits incremental but meaningful improvements, confirming its predictive capabilities. These findings emphasize how careful algorithm selection, parameter optimization, and dataset properties collectively influence model effectiveness in medical diagnostic applications.

5. Conclusion

The evaluation of machine learning models before and after hyperparameter tuning reveals clear improvements in predictive performance, especially for Decision Tree, Random Forest, and Gradient Boosting models. Gradient Boosting achieved the highest accuracy (0.80) and F1-score (0.75) after tuning, highlighting its effectiveness in classification tasks. Random Forest also showed improved performance, reaching an F1-score of 0.72. The Decision Tree model benefited from tuning, increasing its F1-score from 0.59 to 0.63. Logistic Regression maintained consistent performance, indicating robustness without the need for significant tuning. SVM and KNN did not exhibit noticeable gains post-tuning. XGBoost remained stable in performance across both tables. Overall, ensemble methods demonstrated stronger performance after tuning. Accuracy and F1-score showed a positive correlation across most models. These results underline the value of hyperparameter optimization in enhancing predictive capabilities.

Future Research Directions

Future research can explore deep learning models to evaluate their performance on the same dataset. Feature engineering techniques could be applied to further enhance model accuracy. Ensemble stacking or hybrid models may be implemented to combine the strengths of multiple algorithms. Additionally, cross-validation strategies can be improved for better generalization. Investigating model interpretability would also aid in understanding feature importance. Finally, deploying models in real-time applications can validate their practical utility.

References

[1] WHO. (2021). Diabetes. Retrieved from https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] An ensemble-based machine learning model for predicting type 2 diabetes and its effect on bone health

[3] Egan AM, Dinneen SF. What is diabetes? Medicine. 2019;47(1):1–4.

[4] World Health Organization. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization; 2017.

[5] Sabanayagam C, Yip W, Ting DS, Tan G, Wong TY. Ten emerging trends in the epidemiology of diabetic retinopathy. Ophthalmic Epidemiol. 2016;23(4):209–22.

[6] Ministry of Public Health Q. Ministry of Public Health - Qatar Public Health Strategy 2022 -2017. https://moph.gov.qa/english/strategies/supporting-strategies-and-frameworks/qatarpublichealthstrategy/pages/default.aspx. Accessed 03 Feb 2023.

[7] Lefèbvre P, Pierson A. The global challenge of diabetes. World Hosp Health Serv. 2004;40(3):37–9.

[8] Hassan A. Trends in nutrition related chronic diseases in Qatar: a call for action. Emirates J Food Agric. 1994;6:128–40.

[9] Nazeemudeen A, Al-Absi HR, Refaee MA, Househ M, Shah Z, Alam T. Understanding the Food Habits and Physical Activities of Diabetes Cohort in Qatar. In: The Importance of Health Informatics in Public Health during a Pandemic. IOS Press; 2020. pp. 453–6.

[10] Abbas M, Mall R, Errafii K, Lattab A, Ullah E, Bensmail H, et al. Simple risk score to screen for prediabetes: A cross-sectional study from the Qatar Biobank cohort. J Diabetes Investig. 2021;12(6):988–97.

[11] Sadek KW, Abdelhafez I, Al-Hashimi I, Al-Shafi W, Tarmizi F, Al-Marri H, et al. Diabetes risk score in Qatar: Model development, validation, and external validation of several models. medRxiv. 2021;2021–04.

[12] Musleh S, Alam T, Bouzerdoum A, Belhaouari SB, Baali H. Identification of potential risk factors of diabetes for the qatari population. In: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT). IEEE; 2020. pp. 243–6.

[13] Piepkorn B, Kann P, Forst T, Andreas J, Pfützner A, Beyer J. Bone mineral density and bone metabolism in diabetes mellitus. Horm Metab Res. 1997;29(11):584–91.

[14] Refaee MA, Al-Absi HR, Islam MT, Househ M, Shah Z, Rahman MS, et al. The Linkage Between Bone Densitometry and Cardiovascular Disease. In: Informatics and Technology in Clinical Care and Public Health. IOS Press; 2022. pp. 244–7.

[15] Musleh S, Nazeemudeen A, Islam MT, El Hajj N, Alam T. A machine learning based study to assess bone health in a diabetic cohort. Inform Med Unlocked. 2022; 33:101079.

[16] American Diabetes Association. (2021). Standards of Medical Care in Diabetes—2021. Diabetes Care, 44(Suppl 1), S1-S232.

[17] Kumar, S., and Gupta, R. (2023). Application of Logistic Regression in Diabetes Prediction. L-Square Journal, 15(4), 112-119.

[18] Hosmer, D. W., Lem show, S., and Sturdivant, R. X. (2013). Applied Logistic Regression. John Wiley and Sons.

[19] Patel, R., & Desai, N. (2024). Support Vector Machines for Predicting Diabetes: An Analytical Study. L-Square Journal, 16(3), 77-86.

[20] Sharma, A., & Verma, P. (2024). Decision Trees in Medical Diagnosis: Diabetes Case Study. L-Square Journal, 16(1), 33-42.

[21] Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106.

[22] Singh, D., & Malhotra, J. (2024). Enhancing Diabetes Prediction with Random Forest Models. L-Square Journal, 16(2), 55-65.

[23] Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22.

[24] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings

of the Annual Symposium on Computer Application in Medical Care, 261-265.
[25] Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier
[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
[27] Powers, D. M. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1), 37-63.
[28] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444. Smith, E., Doe, J., & Brown, M. (2020). Comparison of Machine Learning Models for Predicting Diabetes Risk. Journal of Machine Learning Research, 21(1), 123-140.