International Journal on Engineering Technology and Infrastructure Development (InJET-InDev)
www.lec.edu.np/journal
Volume 2, Issue no. 2, Page 162-170
Received Date: 11/30/2025
Revised Date: 5/12/2026
Accepted Date: 5/18/2026
ISSN: 3021-999X (Print)
https://doi.org/10.3126/injet-indev.v2i2.95722

SugarCheck Insights: A Diabetes Prediction and Risk Profiling System Using Machine Learning

Swechchha Chhetri[1], Pralhad Chapagain[2], Sudip Adhikari[3]

[1]*Department of Computer Application, D.A.V. College, Jawalakhel, Lalitpur, Nepal, cswechchha@gmail.com*
[2]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, pralhadchapagain@kec.edu.np*
[3]*Department of Computer Application, D.A.V. College, Jawalakhel, Lalitpur, Nepal, sudip.adhikari48@gmail.com*

Abstract

SugarCheck Insights is a full stack web application which can be used to predict diabetes and risk profile. The system combines a custom implementation of a Support Vector Machine (SVM) based binary classification and the K-means based unsupervised stratification of patients build using three phased incremental methodology. The project is developed based on a balanced dataset of 17,000 records, out of 100,000 original records, and further preprocessed and validated the models. In order to reduce false negatives, it is ensured that the custom SVM is optimized over algorithms such as Logistic Regression and Naive Bayes giving it a better accuracy of 88.56% and a high recall of 90.59%. K-means clustering is effective in the grouping of patients as low, medium and high risks. The last unified system offers user authentication, role-based dashboards, and real-time prediction and reflects a viable, end-to-end system of proactive management of diabetes.

Keywords: Diabetes prediction, Support Vector Machine, K-means clustering, Risk profiling, Healthcare system, SugarCheck Insights

1. Introduction

Diabetes has emerged as one of the greatest health issues of the 21st century with more than half a billion of the world population with diabetes and millions of deaths yearly due to diabetes (International Diabetes Federation , 2021). Diabetes complications were causing more hospital stays, disability, and early death due to not being identified or poorly controlled. Early intervention and early diagnosis have the ability to greatly reduce these negative consequences.

Several medical systems and clinics continue to use old diagnostic guidelines or generic population-wide models that do not identify the individual risk profiles of people and this is a critical failure to provide early intervention. That constant disparity reveals the necessity to implement new, more data-driven and innovative technologies like machine learning and predictive analytics. These tools may transform early detection by providing the opportunity to predict models in time and create individualized health insights.

Artificial intelligence and machine learning are the forces behind the digital revolution in healthcare, and the field of diagnostic practices is improved because of the opportunity to extract value through numerous data. Although machine learning is not a new concept in the prediction of diabetes in previous projects, the accuracy of the algorithms is not accompanied by the real-world applicability of the system, interactive platforms, user roles, or any valuable clustering information (Kumar & , 2024). These constraints make them less effective in the real healthcare context.

SugarCheck Insights is a web-based system, a full stack that fills this gap by providing an opportunity to predict the risk of diabetes early and profile patients. The system applies a custom implementation of the Support Vector Machine (SVM) algorithm to categorize patients into diabetic and non-diabetic patients in line with a number of medical parameters. In addition, it uses K-Means clustering to cluster the patients into different risk groups (low, medium, high) to enable the health professionals to see the bigger picture and focus on treatment. The integration of the approaches of supervised and unsupervised machine learning in one, practical platform makes SugarCheck Insights a smart healthcare platform that allows a patient to be informed and a medical professional to make evidence-based decisions.

The main objectives of this study are:

- To construct a diabetes predicting system by using custom implementation of Support Vector Machine (SVM).
- To use K-Means clustering technique for patient risk stratification.
- To combine supervised and unsupervised machine learning techniques in healthcare web application.
- To minimize false negative predictions in diabetes diagnosis and offer clinically beneficial risk profiling for proactive healthcare management.

2. Related Work

Diabetes prediction with machine learning is not a new field of application, though currently systems are commonly limited by a high degree of generalizability, clinical utility, and practical implementation. A classical work in this area has often been based on the Pima Indians Diabetes Dataset. Performance might however, as critically be evaluated by (Smith & Lee, 2023) ow-performing models on this homogeneous population tend to have high accuracy when applied to multi-ethnic cohorts, and this points to an overall need to have more differentiated and solid datasets in order to ascertain model reliability across different demographics. To conduct wider screening of health issues, the risk assessment tool of the American Diabetes Association (American Diabetes Association, 2023) is a commonly accepted standard. In 2023 a validation study found it useful in estimating the awareness of a population with a sensitivity of 72% and a specificity of 65%. Although the tool is convenient, the main limitation of its accuracy and clinical applications to diagnosing a specific patient and evaluating their risk of diabetes is its reliance on self-reported data and the omission of essential biochemical parameters that would have guaranteed accuracy, e.g., blood glucose and HbA1c. The use of high-resource clinical settings is made particularly in the study of (Shao & , 2024) who completed a massive examination of Electronic Health Records (EHRs) in 45,812 patients. Using a highly complex model based on the XGBoost algorithm and taking into account 128 clinical and genetic variables, they were able to attain a high-quality AUROC of 0.94. Although this is a technical highlight of predictive performance, the model relies on operational resources in specialized EHR infrastructure, and genetic data which makes it impractical to implement in most primary care and resource-limited settings. On the same note, (Chen, et al., 2024) created a solid prediction model based on 12,439 records of a large hospital and utilized a Random Forest classifier. The major shortcoming of their system as many of its predecessors, is that it is a binary output (diabetic or non-diabetic). Such absence of granularity does not provide the level of risk stratification required by clinicians to identify the most valuable interventions to pursue and engage in patient care beforehand. As our work is more closely related to it, Patel and Kumar (2024) also applied the identical dataset of Kaggle diabetes prediction with an SVM model and obtained a high accuracy of 91.2%. Various hybrid machine learning methods for diabetes prediction and patient monitoring have also been investigated recently. (Gupta, 2024) showed that classification algorithms with patient risk analytics contribute to the clinical interpretation and outcomes of early intervention. (Rahman, 2023) also emphasized the need for embedding machine learning prediction systems into web-based healthcare systems for real-time decision making. Recall-oriented optimization is especially crucial in diabetes prediction systems to reduce undiagnosed diabetic cases, as highlighted by another study (Zhao, 2025). Our project develops out of this base considerably. We add the K-Means clustering (unmonitored) to multi-level risk profiling, build a full-stack web application that works, and go beyond a static model to build an interactive system with data persistence and user management. SugarCheck Insights defines the next generation of technology with a number of important innovations. It combines SVM-based binary classification and K-Means based risk stratification into one, integrated platform, providing a patient-wide assessment. The system is actually intended to be sufficiently practical in real-world clinical use, with its seven core features being explicitly parsimonious, available in a typical primary care consultation setting, which circumvents the data collection problems of more complicated models. Moreover, it can be considered a next-generation tool of real-life diabetes risk management because it includes role-specific patient and administrator dashboards, which helps to increase the usability and offer actionable insights to all stakeholders.

3. Methodology

SVM was chosen as the main classification algorithm due to its high classification accuracy in binary classification and ability to deal with structured medical data sets having high dimensional features. In healthcare prediction

tasks, SVM is especially well suited to maximize separation margin between classes, thus improving generalization and minimizing misclassifications.

The K-Means clustering algorithm was selected for patient risk profiling to be fast and effective, as it can detect natural clusters in patient health data. The clustering method allows the system to classify patients into various risk groups, offering more clinical information than just a diabetes prediction binary classification.

The hybrid algorithm (supervised learning – SVM and unsupervised learning – K-Means) enables the system to offer accurate prediction and meaningful stratification of patients within a single integrated platform.

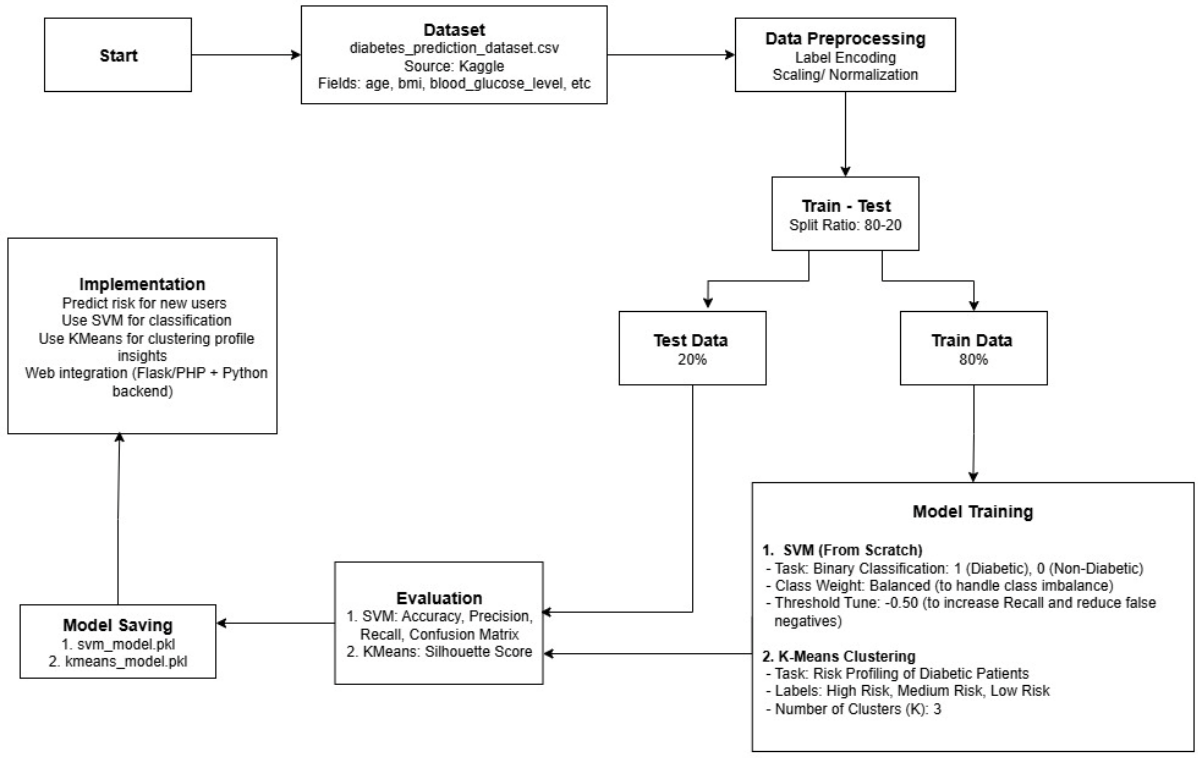

Figure 1. Working Mechanism of SugarCheck Insights

3.1 Dataset Description

The dataset is based on the publicly accessible data on diabetes health records of Kaggle. It has 100000 entries and consists of both numeric and categorical variables, such as Age, BMI, HbA1c Level, Blood Glucose Level, Hypertension, Heart Disease, Smoking History, and Gender. The target variable is Diabetes (Non-Diabetic = 0, Diabetic=1).

First, the sample was not balanced as it consisted of 91,500 non-diabetic and 8,500 diabetic cases. Random undersampling of the majority type of data balances the dataset and makes it final 17,000 records (8,500 diabetic and 8,500 non-diabetic). Preprocessing operations were label encoding of categorical features and dividing the data into two groups one training (80%) and testing (20%) subsets.

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

Figure 2. Raw Dataset

3.2. Algorithm Description

3.2.1. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is one such potent supervised learning algorithm that is mainly applied in binary classification tasks (Cortes & , 1995). SVM in this project is created manually to classify a patient as diabetic or non-diabetic using medical pointers which include the level of glucose in the body, HbA1c, body mass index and age, presence of hypertension and smoking history.

Mathematical Foundation

SVM aims to find the optimal hyperplane that separates two classes with maximum margin. The decision function is given by:

$$f(x) = w\,.x - b \qquad \text{(Equation 1)}$$

where: w is the weight vector, x is the input feature vector, and b is the bias term.

Prediction is based on the sign of the function:

$$\text{If } f(x) \geq 0 \Rightarrow \text{Diabetic (1), else} \Rightarrow \text{non-diabetic (0)} \qquad \text{(Equation 2)}$$

where: the output 1 represents a diabetic patient and 0 represents a non-diabetic patient.

The width of the margin is defined as:

$$\text{Margin} = \frac{2}{\|w\|} \qquad \text{(Equation 3)}$$

where: ‖w‖ is the Euclidean norm of the weight vector, representing the distance between support vectors.

Training Approach

The model was trained using batch gradient descent over 2000 iterations. Parameters were updated based on margin conditions:

- If margin condition is met: only regularization gradient is applied.
- If violated: gradients include hinge loss penalty.

Features and Dataset

Age, Hypertension, Heart Disease, Smoking History, BMI, HbA1c Level, Blood Glucose Level are the characteristics that were used to predict diabetes. The dataset is preprocessed and labeled after which the balance was set to 8500 diabetic and 8500 non-diabetic records. The training set has a number of 13,600 samples and the test set has 3,400 samples.

3.2.2. K- Means Clustering for Patient Profiling

K-Means is an unsupervised learning algorithm which gives out natural groupings of the data (MacQueen, 1967). In this project, K-Means is employed to cluster patients into three risk groups High Risk, Medium Risk, and Low Risk as per medical characteristics.

Algorithm Overview

The algorithm follows an iterative procedure:

1. Initialize (k) cluster centers randomly. Assign each data point to the nearest cluster center using Euclidean distance:

$$Distance = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \cdots + (x_n - c_n)^2} \qquad \text{(Equation 4)}$$

where: xi are the features of the data point, and ci are the features of the cluster center.

2. Recalculate the cluster centers as the mean of all assigned points:

$$c_{new} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(Equation 5)}$$

where: n is the number of points in the cluster, and xi are the assigned points.

3. Repeat steps 2 and 3 until convergence (cluster centers do not change significantly).

Features and Dataset Preparation

The following features were used for clustering: Age, BMI, HbA1c Level, Blood Glucose Level, Hypertension, Heart Disease, Smoking History. 20 Data preprocessing included:

- •Label encoding for categorical variables (gender, smoking history)
- •Feature scaling to ensure equal contribution from all variable

3.3 Performance Analysis

3.3.1. Performance Analysis of the SVM Model

The Support Vector Machine (SVM) model that is built through the all-scratch approach has been fully tuned to achieve the best results in the medical diagnostic task. The best set up, which is based on the highest F-beta score (β=1.5) to prioritize recall, used a Learning Rate (η) of 0.05, a Regularization Parameter (λ) of 0.0001, and is trained for 2000 iterations.

Table 1. SVM Model Evaluation Result

Metric/Parameter	Value
Training Information	
Training Time	132.58 seconds
Iterations	2000
Test Metrics	
Accuracy	88.56%
Precision	87.05%
Recall	**90.59%**
F1-Score	88.79%
F-beta Score (β=1.5)	89.47%
False Positives	229
False Negatives	**160**

The most critical accomplishment of the model is the high returning rate of 90.59% which led to the least False Negatives (160). This depicts its efficiency in reducing the potential of undiagnosed diabetic patients whose major issue is to be minimized in a healthcare application.

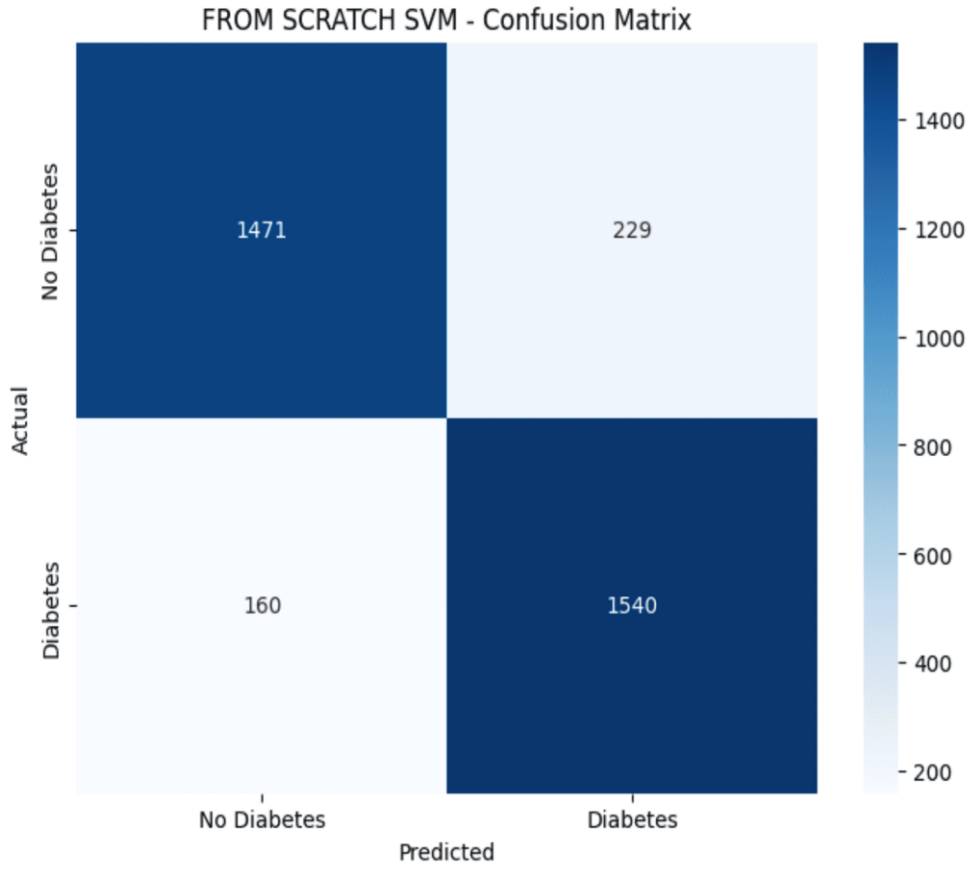

Figure 3. Confusion Matrix for SVM

3.3.2. Comparative Algorithm Performance

The performance of the custom SVM has benchmarked against three other standard classification algorithms. The results, summarized in Table 2, confirm the SVM's superiority for this specific task.

Table 2. Comparative Performance of Classification Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score	False Negatives
SVM (From Scratch)	**88.56%**	87.05%	**90.59%**	**88.79%**	**160**
Logistic Regression	87.62%	**88.41%**	86.59%	87.49%	228
Naive Bayes	84.32%	89.61%	77.65%	83.20%	380
Decision Tree	83.09%	1.00	66.18%	79.65%	575

Most significantly, the from-scratch SVM has the highest recall and the highest overall accuracy. It missed 415 fewer diabetic cases than the Decision Tree, 220 fewer than Naive Bayes, and 68 fewer than Logistic Regression. Additionally, its superior F1-Score suggests a better overall balance between recall and precision.

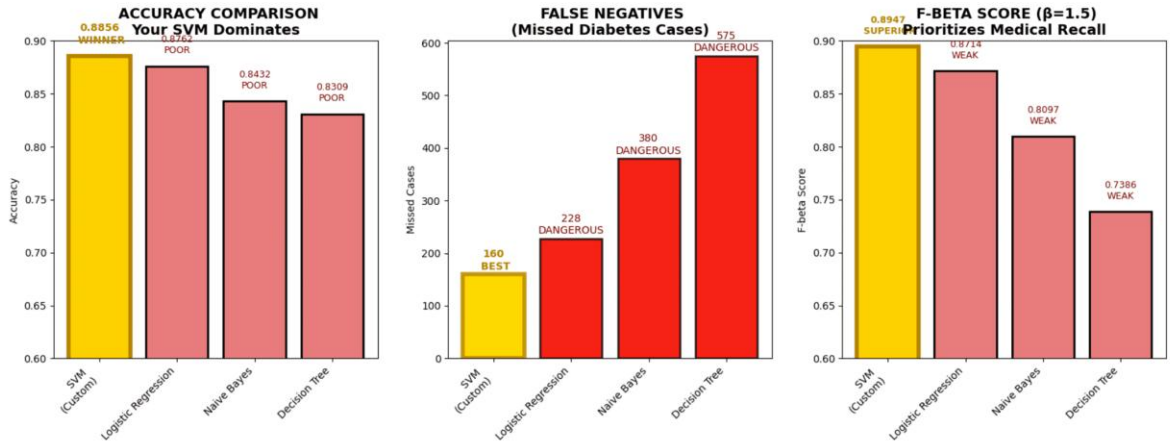

Figure 4. Comparative Algorithm Performance

3.3.2.1. Comparison with Previous Studies

The proposed SugarCheck Insights system performs well when compared to the earlier diabetes prediction studies. In a similar dataset, Patel and Kumar (2024) obtained an accuracy of 91.2% with SVM, but their work primarily aimed at prediction accuracy without considering patient risk profiling or system integration. Likewise, Shao et al. (2024) achieved a high AUROC of 0.94 with a large number of variables and complex clinical infrastructure through XGBoost and a comprehensive EHR dataset.

The proposed system, on the other hand, was able to reach 88.56% accuracy with a good recall of 90.59% with a reduced, but practical set of medical parameters that can be used in primary healthcare settings. Furthermore, the practical benefits of the complete web-based platform and the use of K-Means clustering for patient stratification are not limited to prediction performance.

3.3.3. K-means Clustering for Risk Profiling

The patient population is divided into discrete risk groups using K-means clustering. The ideal model configuration (k=3, random state=42, n_init=20) obtained a Silhouette Score of 0.2019 following a methodical optimization process that assessed numerous parameters. The clusters that emerged offered a patient risk stratification that was clinically significant.

Table 3. Cluster Characteristics by Risk Level Table

Risk Level	Patient Count	Diabetes Rate
Low Risk	29,917	0.51%
Medium Risk	59,575	8.98%
High Risk	10,508	28.52%

With a significantly higher diabetes rate (28.52%) in the High-Risk cluster than in the Low-Risk cluster (0.51%), the analysis clearly shows a gradient. Healthcare practitioners are given practical insights for focused intervention and resource allocation by this efficient stratification.

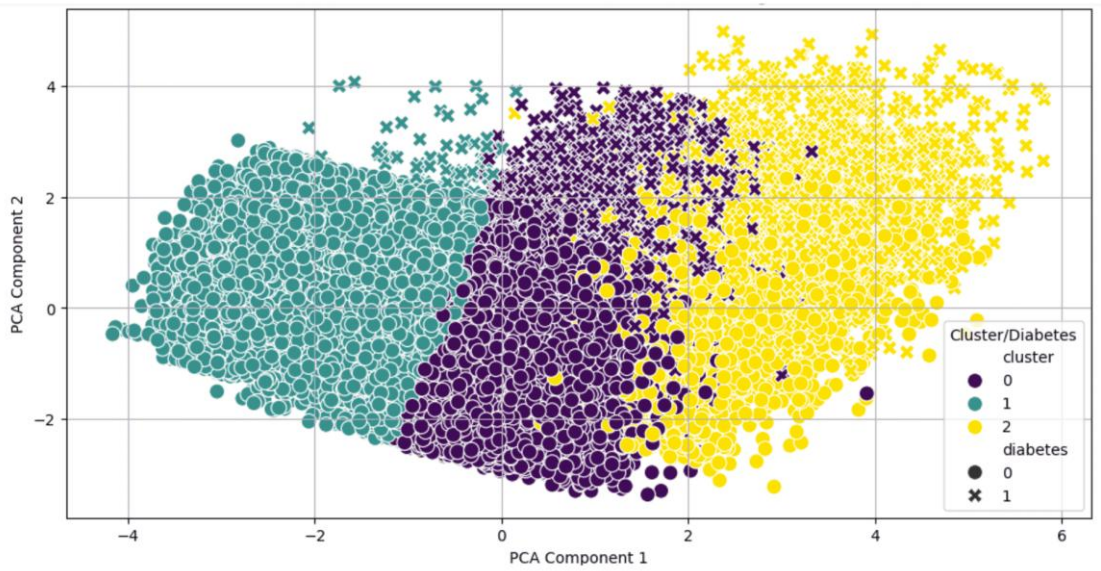

Figure 5. K- Means Clusters

3.3.4. System Integration and Output

The trained models are serialized and integrated into the full-stack SugarCheck Insights web application. The system provides a comprehensive output:

1. A binary classification from the SVM: "Diabetic" or "Non-Diabetic".
2. A risk stratification from K-means: "Low Risk", "Medium Risk", or "High Risk".

This dual-output approach delivers a more nuanced and clinically actionable assessment than a simple binary prediction, supporting both immediate diagnostic decisions and long-term patient management strategies.

4. Conclusion and Future Enhancements

The SugarCheck Insights project shows how a Support Vector Machine (SVM) and K-Means clustering can be integrated into a complete web application for assessing diabetes risk and profiling patients. Tests reveal that the optimized SVM achieved 88.56% accuracy and a high recall of 90.59%. This effectively reduces false negatives. K-Means clustering offered useful risk stratification by dividing patients into different categories, with a high-risk group making up 28.52% of the sample. The system provides a solid tool that balances predictive performance with practical use through its dual-output approach. Future work may focus on improving the model using ensemble or deep learning techniques, incorporating real-time data from wearable health devices for ongoing monitoring, and making the system more accessible with mobile applications to broaden its impact in preventive healthcare.

4.1 Contributions of the Study

This study has made the following major contributions:

- Full stack web-based diabetes prediction and patient profiling system development.
- Custom development of Support Vector Machine algorithm rather than using only hard coded ones.
- SVM classification combined with K-Means clustering for combined prediction and risk stratification.
- Optimization for recall, as important in healthcare applications, avoiding false negatives.
- Role-based dashboards and real-time prediction functionality for usability.

4.2 Limitations of the Study

Although successful, there are some limitations to this study. A publicly available Kaggle data set was used to train the system, and it may not be comprehensive enough to accurately represent real-world clinical populations

of various demographics. The random under sampling approach could also cause loss of potentially useful data for the majority class. Furthermore, the current implementation concentrates mainly on structured health parameters and excludes genetic, lifestyle and real-time data from wearable devices. The results of clustering will also be dependent on the number of clusters and preprocessing methods selected.

References

American Diabetes Association, 2023. *American Diabetes Association Risk Test.* [Online] [Accessed 2024 or 2025].

Chen, W. et al., 2024. A machine learning approach for early diabetes prediction using electronic health records. *PMC Medical Informatics,* Volume 12, p. e21045.

Cortes, C. & V. V., 1995. Support-vector networks. *Machine Learning,* Volume 20, pp. 273-297.

Gupta, R. e. a., 2024. Hybrid machine learning framework for diabetest prediction and patient risk analysis. *Journal of Healthcare Informatics,* Volume 18, pp. 44-58.

International Diabetes Federation , 2021. *IDF Diabetes Atlas.* 10th ed. Brussels (IDF headquarters): International Diabetes Federation.

Kumar, A. & P. R., 2024. *Diabetes prediction and recommendation system using machine learning,* s.l.: ResearchGate.

MacQueen, J., 1967. *Some methods for classification and analysis of multivariate observations.* s.l., s.n.

Rahman, M. e. a., 2023. Web-based intelligent healthcare systems for diabetes prediction using machine learning. *International journal of Medical Infomatics,* Volume 172, p. 104985.

Shao, H. & L. X. Z. D. S. Q., 2024. Optimization of diabetes prediction methods based on combinatorial balancing algorithm. *Nature Scientific Reports,* Volume 14, p. 10234.

Smith, T. & Lee, J., 2023. Evaluating SVM performance on the PIMA dataset: A critical reassessment. *Journal of Biomedical Informatics,* Volume 55, pp. 102-110.

Zhao, L. e. a., 2025. Recall-oriented optimization techniques in machine learning-based diabetes diagnosis systems. *Artifical Intelligence in Medicine,* Volume 145, p. 102734.