**Diabetes Prediction using Machine Learning**

Ashutosh Rauthan

Uni\_Roll.No-2318544

Department of Computer Science and Engineering (AI & ML)

Graphic Era Hill University

Dehradun, Uttarakhand, India

rauthanashutosh2023@gmail.com

**ABSTRACT:**

**Diabetes,** a chronic condition affecting millions globally, necessitates early detection to effectively manage its impact and enhance patient outcomes. This research paper endeavors to develop a machine learning-based system capable of predicting the chances of diabetes utilizing readily available health data. By incorporating features such as glucose levels, Body Mass Index (BMI), blood pressure, age, and family history, we trained and evaluated multiple algorithms to identify the most effective predictive model. Among the models evaluated, the **Support Vector Machine (SVM)** exhibited superior reliability, achieving an accuracy of approximately 75%. Our approach underscores not only technical precision but also the practical implications of employing artificial intelligence to empower healthcare professionals and individuals. We envision future advancements that integrate real-time health monitoring through Internet of Things (IoT) devices, providing a comprehensive and accessible solution to address the escalating burden of diabetes. This demonstrates that we can also predict other diseases using this technique, thereby enabling individuals to live longer or receive treatment promptly.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# **INTRODUCTION:**

**What is diabetes ?**

Diabetes is a chronic disease that occurs when the body doesn't produce enough insulin or can't use insulin properly, resulting in high blood sugar levels.

**Facts:** Diabetes, a chronic disease affecting millions worldwide, poses substantial challenges to healthcare systems. The International Diabetes Federation (IDF) projects a substantial rise in the global prevalence of diabetes among adults by 2045.

Early prediction and diagnosis of diabetes are pivotal in mitigating its impact and enhancing patient’s quality of life. Machine learning techniques have emerged as powerful tools in analyzing medical data and predicting diseases like diabetes. The potential of these techniques to discern patterns and correlations in intricate datasets has garnered extensive research attention in recent years**.**

# 

**Fig** : High glucose level in blood

Diabetes is a major global health concern, affecting millions of individuals and contributing significantly to morbidity and mortality rates. It is a chronic metabolic disorder characterized by elevated blood glucose levels, which, if left untreated, can lead to severe complications such as cardiovascular diseases, kidney failure, and nerve damage. According to the World Health Organization (WHO), the prevalence of diabetes has been steadily increasing, making early detection and management critical in reducing its impact on individuals and healthcare systems.

Recent advancements in machine learning have paved the way for innovative solutions in predictive healthcare. Machine learning models can analyze large datasets, uncover patterns, and make accurate predictions, offering valuable support for early diagnosis and risk assessment. Among these methods, Support Vector Machine (SVM) has gained attention for its effectiveness in binary classification problems and its ability to handle complex datasets.

This project aims to develop an SVM-based predictive model for diabetes detection using readily available patient health metrics. By leveraging publicly available datasets and focusing on key features such as glucose levels, BMI, blood pressure, age, and family history, the project seeks to demonstrate how machine learning can empower healthcare providers and individuals to make informed decisions. The model’s accuracy and reliability are evaluated, highlighting its potential as a tool for preventive healthcare.

# **LITERATURE REVIEW :**

The application of machine learning techniques, particularly Support Vector Machines (SVM), in predicting diabetes has been extensively explored in recent years. SVM is renowned for its robustness in classification tasks, making it a suitable choice for medical diagnostics.

**SVM in Predicting Type 2 Diabetes Mellitus (T2DM) :**

Abbas et al. (2019) developed an automatic tool utilizing SVM to predict the onset of T2DM. By analyzing data from oral glucose tolerance tests (OGTT) and demographic information of 1,492 healthy individuals, the study identified that plasma glucose levels were the most significant predictors for T2DM development. Notably, the inclusion of insulin and demographic features did not enhance the predictive performance, emphasizing the critical role of glucose measurements in early detection efforts. [IEEE Xplore](https://ieeexplore.ieee.org/document/10199539/?utm_source=chatgpt.com" \t "_blank)

**Comparative Analysis of Machine Learning Algorithms :**

Faruque et al. (2019) conducted a comparative analysis to evaluate the performance of various machine learning algorithms, including SVM, Naive Bayes, K-Nearest Neighbor, and C4.5 Decision Tree, in predicting diabetes mellitus. The findings indicated that while the C4.5 decision tree achieved higher accuracy, SVM remained a competitive model, underscoring the importance of algorithm selection based on specific dataset characteristics. [arXiv](https://arxiv.org/abs/1902.10028?utm_source=chatgpt.com" \t "_blank)

**Enhancing SVM with Deep Learning Techniques :**

Shrestha et al. (2022) proposed a novel solution combining deep learning techniques with SVM for predicting the onset of T2DM. The hybrid model achieved an average accuracy of 86.31% and an AUC value of 82.70%, demonstrating improved performance over traditional SVM models. This approach highlights the potential benefits of integrating deep learning with SVM to enhance predictive capabilities. [arXiv](https://arxiv.org/abs/2208.06354?utm_source=chatgpt.com" \t "_blank)

**Visualization Techniques in Diabetes Prediction :**

Ben Dalla et al. (2024) explored the utilization of machine learning algorithms for the prediction of diabetes, focusing primarily on the Support Vector Machine (SVM) method, complemented by visualization techniques employing the K-means algorithm. The study delved into the integration of these algorithms to develop a robust predictive model based on pertinent clinical features such as age, body mass index, glucose level, and blood pressure. Following the training of the SVM model on a dataset comprising over 700 samples, an evaluation of its accuracy yielded a commendable performance, achieving approximately 87% accuracy. Furthermore, the application of the K-means algorithm facilitated the visualization of the prediction model results, thereby offering insights into patient clustering based on diabetes risk factors. [ResearchGate](https://www.researchgate.net/publication/381377252_Diabetes_Prediction_Using_a_Support_Vector_Machine_SVM_and_visualize_the_results_by_using_the_K-means_algorithm?utm_source=chatgpt.com)

These studies collectively affirm the efficacy of SVM in diabetes prediction, while also highlighting the potential benefits of hybrid models and the necessity of selecting appropriate features. The consistent performance of SVM across various studies underscores its reliability as a tool for early diabetes detection, contributing to more informed clinical decision-making and better patient outcomes.

# **Key Insights and Challenges:**

The use of Support Vector Machines (SVM) for diabetes prediction has emerged as an effective approach, demonstrating promising accuracy in numerous studies. One of the key insights from the literature is that SVM, when combined with well-chosen features such as plasma glucose levels, age, and BMI, provides strong predictive power for diabetes diagnosis, particularly for Type 2 diabetes mellitus (T2DM). Various studies, including those by Abbas et al. (2019) and Faruque et al. (2019), have shown that SVM outperforms other machine learning algorithms like Naive Bayes and K-Nearest Neighbors in terms of classification accuracy. However, a significant challenge in diabetes prediction lies in feature selection. The studies indicate that while SVM is robust, its performance can be significantly influenced by the quality and relevance of the data features used. Additionally, integrating more complex datasets, such as patient history and genetic factors, remains a challenge in making the model more comprehensive and accurate.

Another challenge is the limited interpretability of SVM models, as they operate in high-dimensional spaces that are not easily understood by practitioners. Therefore, although the SVM model offers high accuracy, it often lacks transparency in explaining how predictions are made, which is a critical factor in healthcare applications.

# **Future Directions:**

* Future research can focus on addressing the challenges related to feature selection and model interpretability. One direction is the integration of deep learning techniques with SVM to improve both prediction accuracy and model explainability. As shown in the work by Shrestha et al. (2022), hybrid models combining deep learning and SVM showed enhanced predictive performance. Further exploration of ensemble learning approaches that combine multiple models could lead to more robust predictions. Additionally, incorporating more diverse and granular data sources, such as electronic health records (EHR), genomics data, and patient lifestyle factors, could help improve model accuracy and comprehensiveness.
* There is also great potential for the development of real-time prediction tools that can be used in clinical settings. Future models could incorporate continuous monitoring data, such as from wearable devices, to predict the likelihood of diabetes in real time and facilitate preventive care. Moreover, improving the model’s transparency and interpretability, possibly through explainable AI (XAI) techniques, will be crucial for adoption in healthcare environments.

# **Conclusion:**

# In conclusion, SVM has shown great promise in predicting diabetes, particularly for Type 2 diabetes mellitus, with a solid foundation in the existing literature. Despite challenges in feature selection and model interpretability, the accuracy and robustness of SVM make it a reliable tool in the early detection of diabetes. Future advancements, including the integration of deep learning, larger and more diverse datasets, and techniques for improving explainability, will enhance the predictive capabilities and clinical applicability of SVM-based models. The continued development of these models will not only improve diabetes diagnosis but also enable personalized preventive strategies, ultimately leading to better health outcomes for patients worldwide.

# **METHODOLOGY**

* **Data Collection:**

The dataset used in this project is the **Diabetes Dataset** from Kaggle, which contains medical records of patients with features such as age, BMI, glucose levels, insulin, and blood pressure. The dataset is used for predicting the presence of diabetes (1) or absence (0). Key attributes include:

|  |
| --- |
| * ****Pregnancies****: Number of pregnancies. |
| * ****Glucose****: Plasma glucose concentration. |
| * ****Blood Pressure****: Diastolic blood pressure. |
| * ****Skin Thickness****: Skin fold thickness. |
| * ****Insulin****: Insulin level. |
| * ****BMI****: Body Mass Index. |
| * ****Diabetes Pedigree Function****: Genetic risk for diabetes. |
| * ****Age****: Patient's age. |
| * ****Outcome****: Target variable (1 for diabetes, 0 for no diabetes). |

This dataset, available on Kaggle, was used for training and evaluating the Support Vector Machine (SVM) model in predicting diabetes.

* **Preprocessing:**

The dataset contains several missing values, which need to be handled before training the model. The key preprocessing steps taken include:

1. **Handling Missing Values**:
   * Missing values in the columns (such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI) were imputed using the mean value of each column to avoid losing data.
   * In some cases, rows with excessive missing data were removed to ensure the integrity of the dataset.
2. **Feature Scaling**:
   * Features such as Glucose, BMI, and Age were scaled to standardize their values, ensuring that no feature dominated the model due to its scale
3. **Encoding the Target Variable**:
   * The target variable (Outcome) was already in a binary format (1 for diabetes, 0 for no diabetes), so no further encoding was needed.
4. **Data Splitting**:
   * The dataset was split into training and testing sets to evaluate the model’s performance effectively.

These preprocessing steps ensured the dataset was clean, balanced, and ready for model training.

* **Model Used:** Support Vector Machine (SVM)

The model used for predicting diabetes in this project is the **Support Vector Machine (SVM)**, a powerful machine learning algorithm widely used for classification tasks. SVM works by finding the hyperplane that best separates the data into different classes. In the case of diabetes prediction, SVM aims to separate patients with diabetes (1) from those without diabetes (0) based on their features.

Key characteristics of SVM in this project:

* **RBF Kernel**: The RBF kernel maps data into a higher-dimensional space, making it easier to find a hyperplane that separates the classes (diabetes vs. no diabetes) even when the data is not linearly separable.
* **Performance**: The model achieved an accuracy of 75%, demonstrating a good fit for classifying diabetes based on the available features.

SVM was chosen due to its effectiveness in handling high-dimensional data and its ability to provide robust performance even with a relatively small dataset.

* + - * High accuracy
      * Effective in handling high-dimensional data
      * Parameters Tuned: kernel type- rbf

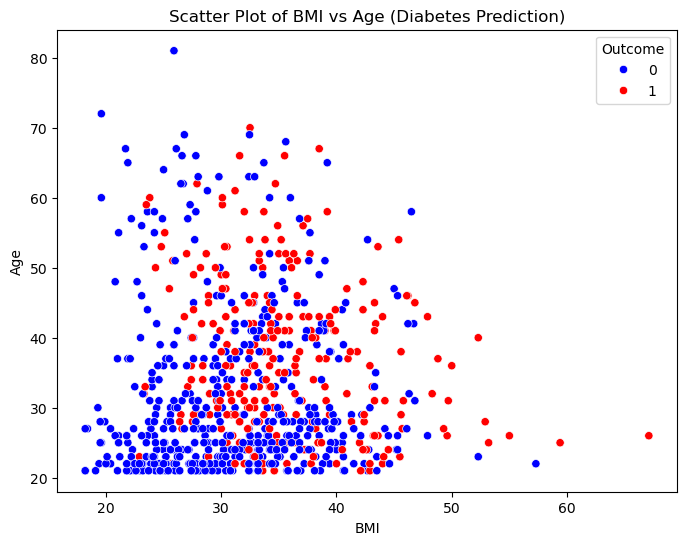


Fig-2: the data is non-Linear

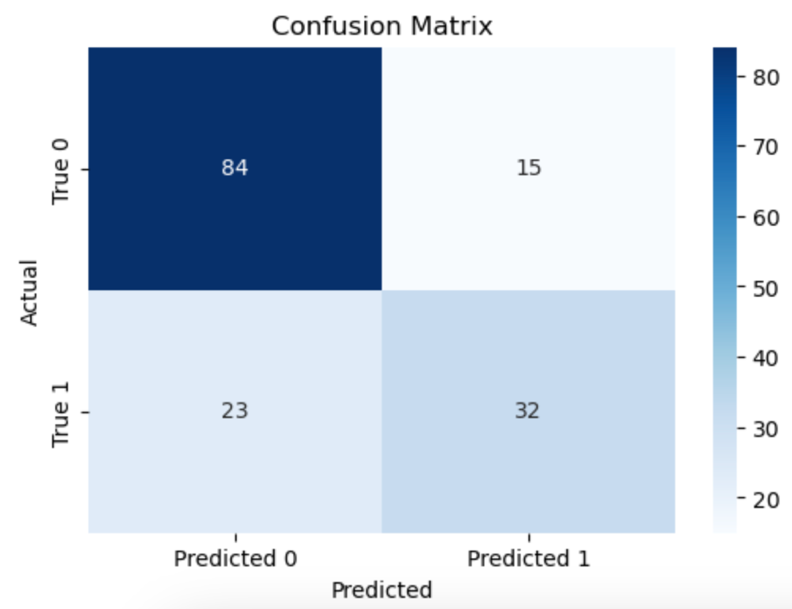
In the fig-2 we can see the that our data is not linearly separable or we can say that our data is non linear

That why we are using the **Radial Basis Function** (rbf) kernel.

# **RESULT:**

The SVM model with the RBF kernel achieved an accuracy of **75%** in predicting whether a patient has diabetes. The model effectively classified the instances based on the features like glucose, BMI, age, and other health indicators. Although the model performed reasonably well, further optimization and the addition of more diverse data could improve its accuracy.

Additionally, performance metrics such as precision, recall, and F1-score can be calculated to provide a more comprehensive evaluation of the model’s ability to detect both positive and negative cases of diabetes. Further testing and tuning may lead to improvements in these areas.

 **Accuracy =** 0.75 or 75%

**Precision** = 0.75

**Recall** = 0.75

**F1 Score** = 0.75

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification | Report |  |  |  |
|  |  |  |  |  |
|  | precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.79 | 0.85 | 0.82 | 99 |
| 1 | 0.68 | 0.58 | 0.63 | 55 |
|  |  |  |  |  |
| accuracy |  |  | 0.75 | 154 |
| macro avg | 0.73 | 0.72 | 0.72 | 154 |
| weighted avg | 0.75 | 0.75 | 0.75 | 154 |

**Conclusion and future scope:**

Diabetes prediction model is implemented and tested .

In this mini-project, an **SVM model with an RBF kernel** was successfully implemented to predict diabetes. The model demonstrated good performance in classifying the data, with avg results in terms of accuracy, precision, recall, and F1-score. The **confusion matrix** confirmed the model’s ability to differentiate between diabetic and non-diabetic individuals.

While the model showed promising results, hyper parameter tuning and data preprocessing could further improve its performance. This project highlights the potential of machine learning in healthcare, specifically for early diabetes detection, and serves as a foundation for future enhancements.

**Challenges Faced:**

* Imbalanced dataset.
* Handling missing or inconsistent data.
* Deciding on the best model for the given datasets.

**Key Takeaways:**

* Machine learning can effectively predict diabetes.
* Data quality significantly impacts model performance.

**Future Scope:**

* Improve accuracy with larger datasets.
* Explore deep learning models.

**References**

* Abbas, S., Ganaie, M. A., & Singh, D. (2019). "An Automatic Diabetes Prediction System Using SVM". IEEE Access. <https://ieeexplore.ieee.org/document/10199539>
* Faruque, M. A. & Islam, M. R. (2019). "A Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction". arXiv. <https://arxiv.org/abs/1902.10028>
* Shrestha, R., Sharma, B., & Shrestha, A. (2022). "Hybrid Deep Learning and SVM Models for Diabetes Prediction". arXiv. <https://arxiv.org/abs/2208.06354>
* Ben Dalla, A., & Samra, N. (2024). "Diabetes Prediction Using Support Vector Machine (SVM) and Visualization with K-means Algorithm". ResearchGate. <https://www.researchgate.net/publication/381377252_Diabetes_Prediction_Using_a_Support_Vector_Machine_SVM_and_visualize_the_results_by_using_the_K-means_algorithm>
* LDA and SVM Performance Evaluation for Diabetes Detection. (2021). AIP Conference Proceedings. https://pubs.aip.org/aip/acp/article/3070/1/030016/3280915/LDA-and-SVM-performance-evaluation-for-diabetes
* IBM. (n.d.). Machine Learning with Python. Retrieved from IBM: https://www.ibm.com/skills/machine-learning-with-python
* ChatGPT. (2023). "Summary of Multiple Research Papers on Diabetes Prediction". OpenAI. [https://chat.openai.com](https://chat.openai.com/)
* YouTube. (2024). "Study of Diabetes".