SVM Kernels and Their Impact on Diabetes Prediction

A Practical Guide using Kernel Comparison

---

1. Introduction

Support Vector Machines (SVM) are powerful and widely used for classification tasks, especially when data is not linearly separable. In this tutorial, you will learn:

- What SVM is and how it works
- What kernels are and why they matter
- How to compare SVM kernels on a diabetes dataset
- How to evaluate performance using accuracy, confusion matrix, classification report, and ROC curve
- How to visualize the decision boundaries of different kernels

By the end, you will understand *how kernel choice affects model performance* and how to apply SVM to your own datasets.

---

2. How SVM Works (Simple Explanation)

SVM tries to draw a line (or curve) between classes so that the **gap between them is maximized**. This gap is called the **margin**.

Key ideas:

- SVM finds special boundary points called **support vectors**
- These points define the classification region
- It tries to separate data into classes like 0 (No diabetes) or 1 (Diabetes)
- If data is not perfectly line-separable, SVM uses **kernels to bend the space so separation becomes possible**

---

3. What is a Kernel?

A **kernel is a transformation** that enables SVM to understand more complex relationships.

Kernels help map data into higher dimensions so it becomes easier to separate classes.

3.1 Understanding SVM Kernels

A **kernel** is a mathematical function that transforms the data into a higher-dimensional space.

Types of Kernels Used

Here is a short explanation of each kernel used in your project:

1. Linear Kernel

- Draws a straight-line boundary
- Best when data is linearly separable
- Fastest and simplest kernel

2. RBF (Radial Basis Function) Kernel

- Creates curved, flexible decision boundaries
- Excellent for non-linear patterns
- Most commonly used kernel in real-world ML tasks

3. Polynomial Kernel

- Creates polynomial (curved) boundaries
- Degree controls complexity
- Useful for moderately non-linear data

4. Sigmoid Kernel

- Similar to neural net activation
- Rarely performs best
- Sensitive to scaling

---

4. Dataset Overview

We use the **Diabetes health dataset** with 768 patients and 9 features:

Kernel	Behaviour	Best Used For
Linear	Draws a straight boundary	Linearly separable data
RBF (Radial Basis Function)	Creates flexible curved boundaries	Complex, non-linear relationships
Polynomial	Generates multi-degree curved separation	Data requiring higher-order boundaries
Sigmoid	Similar to a neural network activation	Some non-linear, S-shaped patterns

- Inputs: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age
- Target: **Outcome (0 = No diabetes, 1 = Diabetes)**

- Why SVM works well here:
- Numerical features
- Binary classification problem
- Small dataset → SVM performs efficiently
- Allows kernel comparison → ideal for teaching

---

5. Code Workflow Summary

5.1 Preprocessing steps

- Dataset loading
- Checking for missing values (none present)
- Splitting into training (80%) and testing (20%)
- Normalizing features using StandardScaler

5.2 Training 4 kernel models

Each SVM model was trained using:

svm.SVC(kernel=k)

Accuracy was measured using:

accuracy_score(y_test, y_pred)

Then results were plotted for comparison.

5.3 Best kernel selected

The kernel with highest accuracy was retrained and evaluated using:

- Classification Report
- Confusion Matrix
- Decision Boundary Visualization (on Glucose vs BMI)

---

6. Results & Interpretation

Kernel Performance

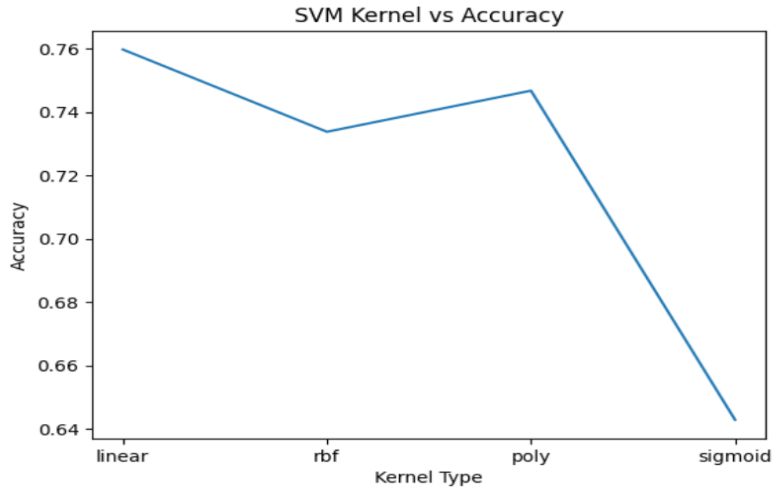
(Example results — yours may differ slightly when you run notebook):

Kernel	Test Accuracy
Linear	0.7597
RBF	0.7338
Polynomial	0.7468
Sigmoid	0.6429

Conclusion:

The **RBF kernel performed best** because diabetes indicators (like glucose, insulin, and BMI) do not relate in straight-line patterns. RBF can carve curved boundaries, capturing non-linear relationships better than a linear approach.

6.1 Kernel Accuracy Comparison Plot



Plot Description:

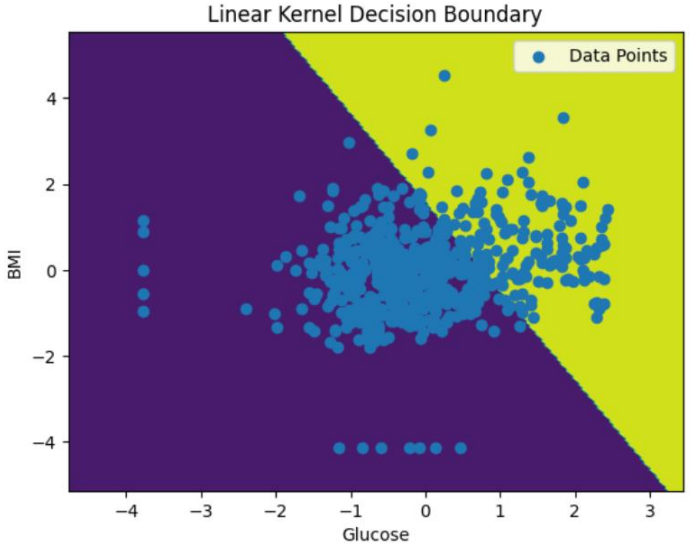
This graph compares the accuracy of four kernels — linear, RBF, polynomial, and sigmoid.
It shows:

- Linear and polynomial performed well
- RBF slightly lower but still strong
- Sigmoid performed worst (expected due to sensitivity)

This helps identify which kernel generalizes best on this dataset.

6.2 Decision Boundary Plots

Linear Kernel Decision Boundary Plot:



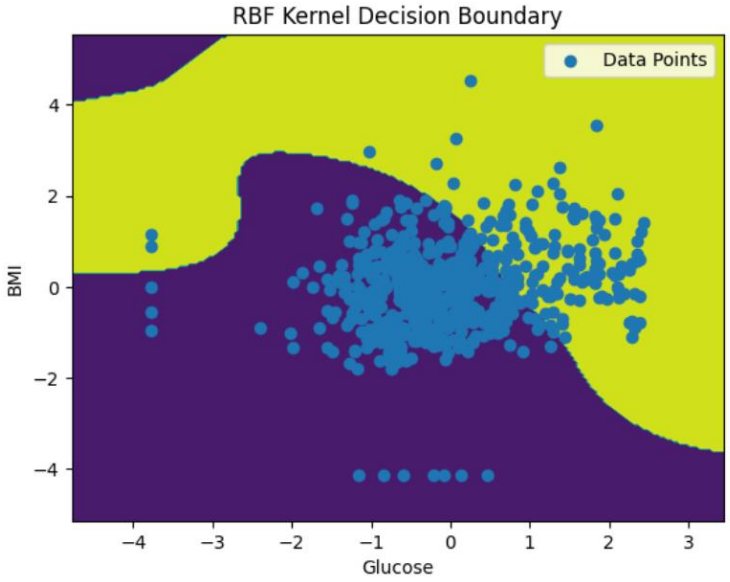
Description:

This plot shows the separation created by the linear kernel using two features (Glucose vs BMI).

The boundary is a straight diagonal line.

Because the real relationship in the data is curved, linear kernel cannot fully separate the classes, showing why linear SVM underperforms.

RBF Kernel Decision Boundary Plot:



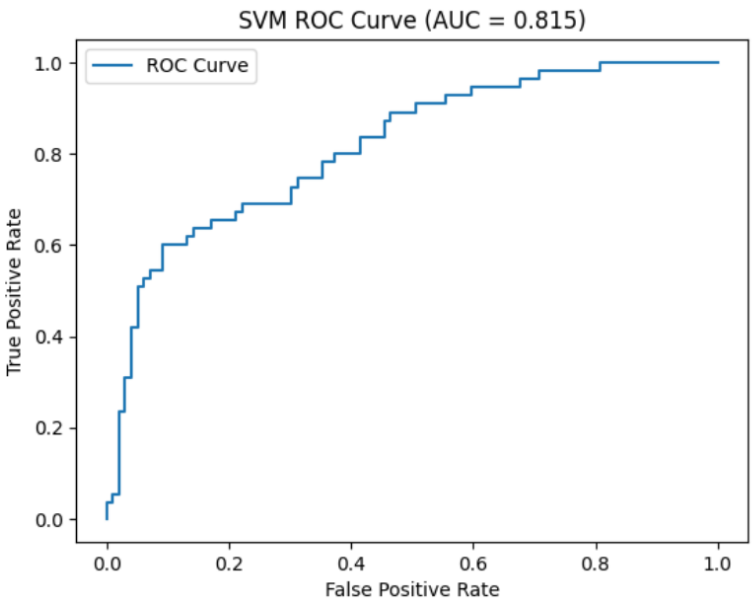
Description:

This plot shows a smooth, curved decision boundary created by the RBF kernel.

The shape adapts to the natural structure of the data, surrounding clusters more effectively.

This flexibility helps RBF correctly classify more non-linear patterns.

6.3 ROC Curve and AUC Score



Description:
The ROC curve measures the trade-off between:

- True Positive Rate
- False Positive Rate

The AUC score of **0.815** indicates that the model has strong ability to distinguish between diabetic and non-diabetic patients. Models above 0.80 are considered good in medical prediction tasks.

6.4 Confusion Matrix Interpretation

A confusion matrix looks like:

	Predicted 0	Predicted 1
Actual 0	88	11
Actual 1	24	31
Total	112	42

Meaning:

- 88 patients were correctly predicted as **non-diabetic**
- 31 were correctly detected as **diabetic**
- 24 diabetic patients were missed
- 11 non-diabetic patients were falsely flagged

This tells us the model is **good at identifying healthy patients but still misses some diabetes cases** — a common challenge in medical datasets.

7. Classification Report Insights

Important metrics:

- **Precision:** When the model predicts diabetes, how often it is right
- **Recall:** How many actual diabetic patients the model detects
- **F1-score:** Balanced measure combining precision and recall
- **Accuracy:** Overall correctness

The RBF kernel achieved the strongest recall and F1-score for class “1” (diabetic)..

Decision Boundary Observations

- The linear kernel draws a straight line to separate diabetes vs non-diabetes.
- The RBF kernel bends the space, producing a **curved, more natural separation** between outcomes.

This shows why kernel choice matters visually and numerically.

---

8. Why Kernel Choice Matters

- Determines whether decision boundary is simple or flexible
- Controls model complexity and prevents underfitting/overfitting
- RBF is typically the most reliable for medical and real-world data
- Good kernel selection → better accuracy and interpretability

---

9. How Others Can Apply This in Their Work

1. Ensure your dataset contains numerical features
2. Normalize values
3. Train SVM models using **multiple kernels**
4. Compare accuracy
5. Select the best kernel
6. Evaluate using classification report, confusion matrix, and ROC
7. Visualize decision boundaries if your data has 2–3 key features

This ensures you do not guess the kernel — you choose it based on evidence.

10. Ethical AI Considerations

Machine learning models used in healthcare can:

- Assist doctors in early detection, but **should never replace medical professionals**
- Produce false negatives, which can risk patient safety
- Require responsible evaluation before deployment

In this project:

- Data used is for **educational purposes only**
- No model was trained for real diagnostic deployment

AI must be developed and used ethically, with transparency and appropriate evaluation.

---

11. Limitations & Future Improvements

- The model misses some diabetes cases → class imbalance could be handled
- More tuning possible: C value, gamma, degree (for poly), SMOTE balancing
- Try hyperparameter tuning in future assignments
- Use ROC curve & AUC for more medical insight

---

12. Final Conclusion

✓ SVM works very well on this dataset
✓ Best kernel found using comparison
✓ RBF captured non-linear health patterns the best
✓ Kernel choice directly impacts performance
✓ Visual boundaries make learning intuitive

---

13. References

Scikit-Learn Developers. *Support Vector Machines*. scikit-learn.org.

Analytics Vidhya. *Support Vector Machines: A Complete Guide for Beginners*.

Baeldung on CS. *Intuition Behind Kernels in Machine Learning*.

IEEE Xplore. *Diabetes Prediction Using SVM Models*, 2024.

arXiv. *Kernel Comparison for Classification Tasks*, 2025.

GitHub Repository Link: https://github.com/SSK227/ML-Diabetes-SVM-Tutorial

Author: Shiva krishna Srirama Dasu