**Comprehensive Summary and Analysis**

**Project Title: Predicting Diabetes Onset using Support Vector Machines (SVM)**

**1. Introduction**

Diabetes is a **global health challenge** affecting **537 million adults worldwide**, according to the **International Diabetes Federation (IDF, 2023)**. It is a **leading cause of blindness, kidney failure, heart attacks, and amputations**. The **American Diabetes Association (ADA)** recommends early screening to prevent **complications** and reduce healthcare costs.

This project leverages the **Pima Indians Diabetes Database**, a well-established dataset in medical AI research, to classify **patients as diabetic or non-diabetic** using **Support Vector Machines (SVM)**. The **objective** is to evaluate whether **SVM-based classification models** can be a viable tool for **clinical decision support systems (CDSS)** in hospitals.

**2. Data Overview & Exploratory Analysis**

**Dataset:** [Pima Indians Diabetes Database](https://www.openml.org/d/37)

**Source:** National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)

**Patient Group:** Pima Indian women (high-risk group for diabetes)

**Total Instances:** 768 patients

**Features (8):** Diagnostic factors such as **glucose levels, blood pressure, insulin levels, BMI, and age**.

**Target Variable:** Diabetes status (**1 = Diabetic, 0 = Non-Diabetic**).

**Key Findings from Exploratory Data Analysis (EDA):**

* **Outliers** detected in **insulin and skin thickness measurements**; this indicates potential data inconsistencies common in medical datasets.
* **Class imbalance**; more non-diabetic cases than diabetic cases, requiring **resampling techniques**.
* **Missing values replaced** with column **means preventing data loss**.

A graph of a distribution of age

AI-generated content may be incorrect.

A graph of a distribution of insu

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.A group of dots with different colors

AI-generated content may be incorrect.

| **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **preg** | 768.0 | 0.226180 | 0.198210 | 0.000000 | 0.058824 | 0.176471 | 0.352941 | 1.0 |
| **plas** | 768.0 | 0.611465 | 0.152945 | 0.221106 | 0.501256 | 0.587940 | 0.704774 | 1.0 |
| **pres** | 768.0 | 0.592253 | 0.099311 | 0.196721 | 0.524590 | 0.590164 | 0.655738 | 1.0 |
| **skin** | 768.0 | 0.268752 | 0.097285 | 0.070707 | 0.207439 | 0.232323 | 0.323232 | 1.0 |
| **insu** | 768.0 | 0.140260 | 0.110024 | 0.016548 | 0.094326 | 0.094326 | 0.150414 | 1.0 |
| **mass** | 768.0 | 0.483619 | 0.102465 | 0.271237 | 0.409836 | 0.476900 | 0.545455 | 1.0 |
| **pedi** | 768.0 | 0.194990 | 0.136913 | 0.032231 | 0.100723 | 0.153926 | 0.258781 | 1.0 |
| **age** | 768.0 | 0.410381 | 0.145188 | 0.259259 | 0.296296 | 0.358025 | 0.506173 | 1.0 |

**3. Data Preprocessing**

* **Missing values handled:** Replaced zeros in medical features (**plas, pres, skin, insu, mass**) with **mean values**.
* **Data splitting:** **70% for training, 30% for testing**.
* **Encoding:** Categorical target variable encoded for machine learning.

**4. Model Selection & Training**

Several **SVM configurations** were tested:

|  |  |  |
| --- | --- | --- |
| Kernel | Decision Function Shape | Accuracy (Test) |
| Linear | ovo | **74%** |
| RBF | ovo | **74%** |
| Polynomial | ovo | **75% (Best Performance)** |

**Why Polynomial Kernel Performed Best?**

* Captures **non-linear relationships** in medical data.
* Balances **accuracy and generalizability**.

**5. Model Evaluation**

**Performance Metrics on Test Set:**

**Overall Accuracy:** **74%**

**Precision (Diabetic Cases):** **0.64**

**Recall (Diabetic Cases):** **0.59**; indicates missed diabetic cases.

**F1-Score (Diabetic Cases):** **0.61**

**Key Takeaways:**

The model is **better at predicting non-diabetic cases** than diabetic cases.

**False negatives** (missed diabetes cases) are a concern in medical settings.

**Class imbalance negatively impacts recall for diabetic cases.**

**6. Medical Significance & Real-World Implications**

**A. How Can This Be Used in Healthcare Institutions?**

* **Hospitals like Mayo Clinic and Johns Hopkins** use predictive models for early **diabetes screening**.
* This model can be **integrated into hospital EHRs** for automated **risk assessment**.
* **Remote patient monitoring programs** (such as those at **Cleveland Clinic**) could use an improved version of this model to **track high-risk patients**.

**B. Ethical Considerations in AI for Healthcare**

* **Bias and Fairness:** The model **performs poorly for diabetic cases**, requiring **improvements for equity**.
* **Regulatory Compliance:** Any real-world deployment would need to meet **FDA and HIPAA** standards.
* **Interpretability:** Clinicians need **explainable AI models**; SVM could be paired with **SHAP or LIME** to improve transparency.

**7. Limitations & Future Work**

**Limitations:**

* **Class imbalance issue** reduces performance for diabetic cases.
* **Feature engineering was minimal**; additional features (e.g., **HbA1c levels**) could improve accuracy.

**Future Improvements:**

* **Apply Oversampling/Undersampling**. Improve model recall for diabetic cases.
* **Explore Hybrid Models**. Combine SVM with **Random Forest or Neural Networks**.
* **Incorporate Real-World Data from EHRs**. Validate the model in clinical settings.

**8. Conclusion**

This project demonstrates that **Support Vector Machines (SVM)** can serve as a **predictive tool for diabetes classification**. While the model performs **well overall (75% accuracy)**, it has limitations in identifying diabetic patients due to **class imbalance**. Future work should focus on **enhancing recall, integrating real-world hospital data, and improving interpretability** for practical adoption in healthcare settings.

**References**

ScikitLearn. Support Vector Machines. <https://scikit-learn.org/stable/auto_examples/svm/index.html>.

ScikitLearn. Precision\_Recall\_Fscore\_support. <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html>.

ScikitLearn. Train\_Test\_Split. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>.

ScikitLearn. Classifier Comparison. <https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py>.

# Ghoneim, S. (2019, April 2). *Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on?.* Towards Data Science. <https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124>.

Glassner, A. (2021). Chapter 11: Classifiers, Support Vector Machines. *Deep Learning: A Visual Approach*. San Francisco: No Starch Press. <https://learning.oreilly.com/library/view/deep-learning/9781098129019/c11.xhtml?ar%C2%A0&sso_link=yes&sso_link_from=umgc>.

Fawcett, T. (2024). *Learning from Imbalanced Classes.* KDnuggets. <https://www.kdnuggets.com/2016/08/learning-from-imbalanced-classes.html>.

Higuera C., Gardiner KJ., Cios KJ. (2015, June 25). *Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome.* PubMed. 10(6):e0129126. doi: 10.1371/journal.pone.0129126. <https://pubmed.ncbi.nlm.nih.gov/26111164/>.

*Harvard CS109A. Lecture 6: Linear Regression. GitHub.*  <https://harvard-iacs.github.io/2020-CS109A/lectures/lecture06/notebook-3/>.

NIH.(2024, January). *Diabetes Statistics.* <https://www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics>.

<https://www.openml.org/search?type=data&sort=runs&id=37&status=active>

Giri, S. (2024). *AI-driven Predictive Models for Early Det3ection of Diabetes:* *A Review Study*. International Journal of Computer Science and Mobile Computing, 13(9), 24–33. <https://d1wqtxts1xzle7.cloudfront.net/118489533/V13I9202408-libre.pdf?1727526056=&response-content-disposition=inline%3B+filename%3DAI_Driven_Predictive_Models_for_Early_De.pdf&Expires=1741631784&Signature=A6ZYCDvci-MwMcwzDi00y-p7ltj2ISDZXJGskTDsfklQLWvuVZvOdtb07jUbBsgNIkzrTeTkYd3fwiFkQMlws97orhfgtdcqMXzKdIyft0RP3H~q~5fdlRLH~DY6EmA5xrRtH53un6WDZkcRgabx~MdsLc38ZtdyW3xRahUx881e7g1JZRdIHlsRubEvMLT5uFO2xYW7oeGgM55QsMXjW0NhvT8pWlq8mtjWLYrWnWTn4cy8sQkWP3azD5E38t20lDjIU3UdtyQkOiLS-V~vEiG6pC5YMHiMvzsMJ~lZ3sklTd8kYY6sAO12QRTlzSf8yzn14jHpWrNk1tGFScWGKA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA>.

Adewale, T. (2024). *Improving Patient Outcomes with Cloud-driven Artificial Intelligence in Healthcare*. ResearchGate. <https://www.researchgate.net/profile/Tunmise-Adewale-2/publication/387426412_Improving_Patient_Outcomes_with_Cloud-Driven_Artificial_Intelligence_in_Healthcare/links/676d4a4c117f340ec3d9184d/Improving-Patient-Outcomes-with-Cloud-Driven-Artificial-Intelligence-in-Healthcare.pdf>.