Diabetes prediction classifier comparison and evaluation

Ashis Bayen,

Four year B.Sc (Physics Major), Bangabasi college

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project focuses on applying machine learning techniques to a small cancer dataset consisting of demographic and clinical features. The goal was to predict metastasis status of patients using classification algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The project involved data preprocessing (handling categorical variables, scaling, and splitting into training and testing sets), model development, and comparative performance evaluation using metrics like accuracy, precision, recall, F1-score and ROC-AUC. The results demonstrate how predictive models can support clinical decision making by identifying patterns within cancer data. Although the dataset is small and synthetic, the workflow replicates real-world medical analytics. This work also helped develop practical skills in Python programming, pandas, scikit-learn, and visualization libraries. The findings highlight the importance of data quality, feature encoding, and model validation.

---

2. Introduction

Cancer remains one of the leading health challenges globally, and data-driven approaches are increasingly being used to improve diagnosis and treatment planning. This project aims to illustrate how machine learning methods can classify cancer-related outcomes, specifically predicting metastasis status, from patient data. The technology stack used includes Python, pandas, NumPy, matplotlib, seaborn, and scikit-learn. We performed a background survey of classification techniques and model evaluation metrics. The procedure involved collecting a small cancer dataset, encoding categorical variables, splitting the data, training multiple classifiers, and comparing their performance. The purpose of doing this project was to gain hands-on experience with data preprocessing, machine learning model development, and evaluation during the internship period.

Topics covered during first two weeks of internship:

- Introduction to Python programming and Jupyter Notebooks

- Data cleaning and preprocessing in pandas
- Exploratory data analysis and visualization (matplotlib, seaborn)
- Basics of classification algorithms (Logistic Regression, KNN, SVM)
- Model evaluation metrics: accuracy, precision, recall, F1-score, ROC-AUC
- Introduction to GitHub for version control and code sharing

---

3. Project Objective

Objectives of the project:

- To explore a cancer dataset and understand the relationships among demographic and clinical features.
- To preprocess and encode categorical data for machine learning applications.
- To develop and compare multiple classification models (Logistic Regression, KNN, SVM) for predicting metastasis status.
- To evaluate model performance using standard metrics and identify the most effective classifier.
- To gain practical experience in Python and machine learning as part of the internship program.

---

4. Methodology

Step-by-step process:

1. **Data Collection**: Used the provided cancer_dataset.csv (30 records).
2. **Data Cleaning & Preprocessing**:
    - Identified categorical variables (Gender, Cancer_Type, Stage, Treatment_Type).
    - Used One-Hot Encoding for categorical variables.
    - Standardized the features to improve model performance.

3. **Train/Test Split**: Split the dataset into 75% training and 25% testing sets.
4. **Model Development**: Trained three classification algorithms—Logistic Regression, KNN, and SVM—using scikit-learn pipelines.
5. **Model Evaluation**: Calculated accuracy, precision, recall, F1-score, and plotted ROC curves and confusion matrices.
6. **Tools Used**: Python 3.x, Jupyter Notebook, pandas, NumPy, scikit-learn, matplotlib, seaborn.
7. **Version Control**: All code stored in a GitHub repository (insert link here).

(Optional: you can include a simple flowchart of these steps in your report.)

---

5. Data Analysis and Results

Descriptive Analysis:

- Dataset of 30 records with features like Gender, Cancer Type, Stage, Metastasis, and Treatment Type.
- Metastasis column chosen as the target variable (Yes/No).

Model Performance Summary (example):

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.25	0.33	0.20	0.25	–
KNN	0.38	0.50	0.20	0.29	–
SVM	0.38	0.50	0.40	0.44	–

(Include your actual numbers, histograms, and ROC/confusion matrix screenshots here.)

---

6. Conclusion

This project demonstrated the end-to-end process of using machine learning to predict cancer metastasis from a small dataset. Despite the small sample size, we successfully performed data cleaning, encoding, model training, and

evaluation. Among the three models tested, SVM achieved the best recall and F1-score for identifying metastasis cases. This hands-on experience strengthened our understanding of Python, data preprocessing, and model evaluation. For future work, the model could be improved by increasing the dataset size, using cross-validation, and exploring advanced algorithms (e.g., Random Forest, Gradient Boosting).

---

7. Appendices

- Scikit-Learn Documentation: https://scikit-learn.org/stable/
- Python pandas: https://pandas.pydata.org/
- Cancer Dataset (provided during internship)
- Chat GPT