Executive Summary – Tumor Diagnostics with Machine Learning

Author: Åsa Hellstrand Date: April 26, 2025

Goal

The aim of this project was to identify the most effective machine learning model for classifying tumors as benign or malignant, using the Breast Cancer Wisconsin dataset (UCI ML Repository).

Methods

- Data: 683 tumor samples with 10 cellular features (e.g., clump thickness, uniformity of cell size, bare nuclei).
- Models tested:
  - Logistic Regression
  - K-Nearest Neighbors
  - Support Vector Machine (SVM)
  - Random Forest
- Metrics: Accuracy, Precision, Recall, F1-score, and AUC.
- Implementation: Python (scikit-learn) in Jupyter Notebook.

Key Findings

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Logistic Regression	0.956	0.940	0.940	0.940	0.99
K-Nearest Neighbors	0.956	0.923	0.960	0.941	0.99
Support Vector Machine	0.956	0.907	0.980	0.942	0.99
Random Forest	0.971	0.942	0.980	0.961	1.00

- Recall is the most critical metric in tumor diagnostics (detecting all malignant tumors).
- Random Forest achieved the best overall performance, balancing high recall with strong precision and F1-score.

- Small variations in train/test splits sometimes favored SVM, showing that multiple models perform well on this dataset.

Implications

- Ensemble methods like Random Forest are highly effective for structured medical data.
- Robust validation (e.g., cross-validation) is important, since results can shift with different data partitions.
- This project illustrates how machine learning can support reliable, automated tumor diagnostics.

Deliverables

- Code & Notebook → model training, evaluation, and visualizations.
- Report → scientific-style analysis of results.
- Repository → public on GitHub with dataset reference.

Summary: Random Forest is the most suitable model for this dataset with the chosen data split, but both Random Forest and SVM demonstrate strong potential for tumor diagnostics tasks.