# Predictive Modeling Analysis: Breast Cancer Diagnosis

**Overview**

This study explores predictive modeling techniques for breast cancer diagnosis using the Wisconsin Breast Cancer Dataset. Four classification models—Decision Tree, k-Nearest Neighbors (k-NN), Logistic Regression, and Support Vector Machine (SVM)—were implemented and evaluated based on multiple performance metrics. The primary goal was to classify tumors as malignant or benign while optimizing model performance through hyperparameter tuning.

**Modeling Process**

The dataset was preprocessed by handling missing values, encoding categorical variables, and scaling numerical features where necessary. Given the class imbalance (more benign cases than malignant), **recall was chosen as the key evaluation metric** to minimize false negatives, ensuring that malignant cases were correctly identified. **ROC-AUC was used as a secondary metric** to assess overall model discrimination capability. A **nested cross-validation** approach was used for hyperparameter tuning, ensuring unbiased performance estimates.

**Results and Findings**

* **Logistic Regression** emerged as the best-performing model with **98% recall and an ROC-AUC of 0.998**, making it the most reliable choice for identifying malignant cases.
* **SVM** closely followed, achieving **95% recall and an ROC-AUC of 0.997**, showing strong overall performance.
* **k-NN** had **93% recall and an ROC-AUC of 0.982**, indicating moderate sensitivity to malignant cases.
* **Decision Tree**, while interpretable, had the lowest recall (**93%**) and ROC-AUC (**0.949**), making it less favorable in an imbalanced dataset.

**Model Comparison Table**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC-AUC** |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Decision Tree** | 0.94 | 0.91 | 0.93 | 0.92 | 0.949 |  |
| **k-NN** | 0.95 | 0.93 | 0.93 | 0.93 | 0.982 |  |
| **Logistic Regression** | **0.97** | **0.95** | **0.98** | **0.97** | **0.998** |  |
| **SVM** | 0.97 | 0.98 | 0.95 | 0.96 | 0.997 |  |

**Model Evaluation**

A **confusion matrix** analysis confirmed that **Logistic Regression had the fewest false negatives**, making it the safest option for medical diagnosis. The **ROC curve** demonstrated strong model separability, while the **Lift curve** highlighted superior predictive power in high-risk cases.

**Conclusion**

Due to the dataset imbalance, **recall was prioritized** in model selection to avoid missing malignant diagnoses. **Logistic Regression provided the best trade-off between recall and ROC-AUC, making it the most suitable model.** SVM also performed well, while Decision Tree and k-NN had lower recall.