Breast Cancer Prediction Using Machine Learning

Final Report

Introduction:

Breast cancer is one of the most common cancers in women worldwide. Early diagnosis significantly improves the chances of successful treatment and survival. In this project, we developed and compared several machine learning classification models to predict whether a tumor is malignant or benign based on various features in a publicly available dataset.

Objective:

The goal of this project is to:

- Use the Breast Cancer Wisconsin dataset to predict the nature of tumors.
- Build a reliable, accurate classification model.
- Compare the performance of different machine learning algorithms.
- Visualize and interpret the results to gain insights into model behavior and dataset characteristics.

Methodology:

Tools & Libraries Used

- Python (programming language)
- Pandas, NumPy – for data manipulation
- Scikit-learn – for machine learning models and preprocessing
- Matplotlib, Seaborn – for visualizations

Steps Followed

1. Data Loading: Loaded breast_cancer_data.csv into a Pandas DataFrame.
2. Data Preprocessing:
    - Dropped null values (none found).
    - Separated features (X) and target labels (y).

    - Scaled features using StandardScaler.

3. Model Building: Trained and evaluated the following classifiers:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Random Forest
    - K-Nearest Neighbors (KNN)
    - Naive Bayes
4. Evaluation Metrics:
    - Accuracy
    - Classification Report (Precision, Recall, F1-score)
    - Confusion Matrix
5. Visualization:
    - Correlation Heatmap
    - Accuracy Comparison Bar Chart
    - Confusion Matrix Heatmap

Code and Implementation Details:

All steps were implemented in Python. Below is a brief summary (refer to the attached notebook/script for full code):

```
# Load and preprocess data
df = pd.read_csv("breast_cancer_data.csv")
X = df.drop("target", axis=1)
y = df["target"]


# Split and scale
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y)
scaler = StandardScaler()
```

```
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Model training
models = {
    "Logistic Regression": LogisticRegression(max_iter=10000),
    "SVM": SVC(),
    "Random Forest": RandomForestClassifier(n_estimators=100),
    "KNN": KNeighborsClassifier(),
    "Naive Bayes": GaussianNB()
}
```

Each model was trained and evaluated with consistent preprocessing for fair comparison.

Results and Observations:

Model Accuracy

Model	**Accuracy (%)**
Logistic Regression	97.37
SVM	97.37
Random Forest	96.49
KNN	95.61
Naive Bayes	93.86

Visuals Included:

- Correlation Heatmap showed that radius_mean and perimeter_mean were strongly correlated with the target.
- Confusion Matrix revealed that misclassifications were minimal in top-performing models.

- Accuracy Bar Plot made model comparison intuitive.

Key Insights:

- Logistic Regression and SVM gave the best results.
- Feature scaling significantly improved model performance (especially SVM and KNN).
- The dataset was balanced, making accuracy a good performance metric.

Conclusion

- This project successfully demonstrated the power of machine learning in medical diagnostics. We evaluated five different classifiers, finding that **Logistic Regression and SVM** performed best with **over 97% accuracy**.
- The structured methodology, clear evaluation, and visual insights contribute to building reliable diagnostic tools, and the approach can be extended to more complex healthcare data in the future.

Attachments and Files:

File Name	**Description**
breast_cancer_data.csv	**Dataset used for training**
classification_project.ipynb	**Jupyter notebook with full code and outputs**
model_comparison_plot.png	**Accuracy comparison visualization**
confusion_matrix_svm.png	**Confusion matrix for SVM**