Breast Cancer Prediction Project

Develop a machine learning-based classification system that can accurately predict whether a breast tumor is **malignant (cancerous)** or **benign (non-cancerous)** based on cellular characteristics extracted from digitized FNA images.

Success Criteria:

- Achieve >95% accuracy to be clinically useful
- Minimize false negatives (missing cancer cases) as they are most dangerous
- Identify which cellular features are most predictive
- Create an interpretable model that medical professionals can trust

Data Exploration Findings

Dataset Overview

- **Source**: Wisconsin Breast Cancer Dataset (Kaggle/UCI ML Repository)
- **Samples**: 569 patient records
- **Features**: 30 numerical measurements (10 core measurements × 3 statistics)
- **Target**: Binary classification (357 Benign, 212 Malignant)

Findings

- Dataset is remarkably clean.
- Moderate Imbalance (63% Benign, 37% Malignant)
- Many features show right Skewed distributions
- High multicollinearity exists

Model Selection Rationale

No single algorithm is best for all datasets. We tested diverse approaches to find what works best for THIS specific problem.

1) Logistic regression
2) Decision Tree
3) Random Forest
4) SVM

Logistic Regression (Linear Model)

Why chosen:

- Simple, interpretable baseline
- Coefficients show feature importance
- Fast training and prediction
- Works well for linearly separable data

Expected performance: Good if classes are linearly separable

Actual result: 97.4% accuracy ✅ (exceeded expectations!)

Why it worked: Despite multicollinearity, regularization helped. The problem is more linear than expected.

Decision Tree (Tree-based)

Why chosen:

- Handles non-linear relationships
- No scaling required
- Interpretable rules
- Baseline for ensemble methods

Expected performance: Good but may overfit

Actual result: ~95% accuracy (as expected)

Limitation: Single tree less stable than ensembles

Random Forest (Ensemble)

Why chosen:

- Reduces overfitting via averaging
- Handles feature interactions
- Provides feature importance
- Robust to outliers and noise

Expected performance: Excellent (industry standard)

Actual result: 96.5% accuracy ✅

Why it worked: Ensemble averaging captured complex patterns while avoiding overfitting

Support Vector Machine (SVM) (Kernel Method)

Why chosen:

- Excellent for high-dimensional data
- RBF kernel handles non-linearity
- Strong theoretical foundation
- Often best for medical datasets

Expected performance: Very good with proper tuning

Actual result: 98.2% accuracy 🏆 (BEST MODEL!)

Why it worked:

- Found optimal hyperplane with wide margin
- RBF kernel captured non-linear patterns
- Proper scaling was critical
- C and gamma tuning optimized bias-variance tradeoff

Hyperparameter Optimization Strategy

Why we tuned top 3 models:

1. **SVM**: Most sensitive to C and gamma
2. **Random Forest**: n_estimators, max_depth affect performance
3. **Logistic Regression**: C and penalty type matter

Method: GridSearchCV with 5-fold cross-validation

- Systematic search over parameter space
- Cross-validation prevents overfitting to training set
- Balances thoroughness with computational cost

Key Findings

Finding 1: All Models Perform Exceptionally Well

Conclusion: >95% accuracy across all models indicates:

- High-quality, well-curated dataset
- Clear separation between classes
- Features are highly informative
- Problem is well-suited for supervised learning

Finding 2: SVM is the Winner 🏆

Best Model: SVM with RBF kernel (98.2% accuracy)

Why SVM won:

1. Optimal hyperplane with maximum margin
2. RBF kernel captured non-linear relationships
3. Robust to outliers through support vectors
4. Proper scaling + hyperparameter tuning critical

Practical Interpretation:

- Out of 100 predictions, 98 are correct
- Only 2 errors per 100 cases
- Confidence: 98.9% ROC-AUC (nearly perfect discrimination)

Finding 3: Minimal Improvement from Hyperparameter Tuning

False Negatives vs False Positives

Critical Medical Consideration: False negatives (missing cancer) are MORE dangerous than false positives (false alarms)

Our Model's Performance:

```
Confusion Matrix (SVM):
                Predicted
              Benign  Malignant
Actual Benign    70        1      ← 1 False Positive (tolerable)
    Malignant     1       42      ← 1 False Negative (concerning!)
```

Recall = 97.7% means we catch 42 out of 43 malignant cases

Clinical Implication:

- 1 cancer case missed per ~40 tests
- Could adjust threshold to increase sensitivity (catch more cancers) at cost of more false alarms

Key Takeaways

1. **Multiple algorithms work well**, but SVM slightly outperforms
2. **Feature engineering less important** than data quality for this dataset
3. **"Worst" cellular measurements** are most predictive (clinical validation)
4. **Standardization is critical** for distance-based algorithms
5. **Model should assist, not replace** medical professionals
6. **98.2% accuracy** means 1-2 errors per 100 cases → still requires human oversight
7. **Trade-off between accuracy and interpretability** must be considered for medical applications