Heart Disease Prediction - ML Model Performance Analysis Report

Objective

To evaluate and compare the performance of different machine learning models (Logistic Regression and SVM) on heart disease prediction, with and without feature reduction techniques like feature selection and PCA.

1. Logistic Regression Model

Feature Reduction Techniques Used:

Method	Parameters	Train Accuracy	Test Accuracy
Feature Selection	k = 8	84.70%	83.60%
PCA	**n_components = 9**	**85.53%**	**85.26%**
Simple (No Reduction)	-	85.53%	80.32%

Analysis:

- Best Test Accuracy: Achieved using PCA (85.26%), closely followed by Feature Selection (83.60%).
- Observation: Feature reduction via PCA maintained model generalization better than no reduction (which overfits slightly).

2. SVM Model

A. With PCA-Based Feature Extraction

Kernel	n_components	Train Accuracy	Test Accuracy
Linear	7	83.47%	83.60%
Polynomial	5	84.71%	75.40%
Gaussian (RBF)	5	83.47%	77.04%

B. With Feature Selection

Kernel	k	Train Accuracy	Test Accuracy

Linear	10	86.36%	81.96%
Polynomial	7	91.32%	80.32%
Gaussian (RBF)	8	91.32%	83.60%

C. Without Feature Reduction (All Features)

Kernel	Train Accuracy	Test Accuracy
Linear	85.95%	81.96%
Polynomial	97.10%	75.40%
Gaussian (RBF)	97.10%	78.68%

SVM Model Analysis [(poly(deg = 3, coef0 = 1) , RBF(gamma = 0.2)]

- With PCA:
  • Best Test Accuracy: Linear kernel (83.60%) with n_components=7.
  • Polynomial and Gaussian show signs of overfitting due to lower test accuracy despite reasonable training scores.

- With Feature Selection:
  • Best Test Accuracy: Gaussian (83.60%) with k=8.
  • Polynomial kernel gives very high train accuracy but test accuracy drops—suggesting overfitting.

- Without Feature Reduction:
  • Overfitting observed in Polynomial and Gaussian kernels (97.10% train but <76% test).
  • Linear kernel shows balance: 85.95% train, 81.96% test.

3. Neural Network Model

Method	Parameters	Train Accuracy	Test Accuracy
Feature Selection	k = 5	82.64%	80.32%
PCA	n_components = 7	88.1%	81.7%
Simple (No Reduction)	-	81.81%	83.6%

Neural Network Model Analysis (2 hidden layers (activation = ReLU))

1. **Simple model achieves highest test accuracy (83.6%)**, indicating the best generalization despite having no feature reduction.

2. **PCA shows highest train accuracy (88.1%)**, but slightly lower test accuracy (81.7%), suggesting possible overfitting due to retained complexity.

3. **Feature Selection (k=5) performs consistently** with balanced train (82.64%) and test (80.32%) accuracy, making it a reliable alternative with reduced dimensionality.

Overall Conclusion

Model	Technique	Best Kernel	Test Accuracy
Logistic Regression	**PCA (n=9)**	**N/A**	**85.26%**
Neural Network	All Features	N/A	83.6%
SVM	PCA (n=7)	Linear	83.60%
SVM	All Features	Linear	81.96%

Diabetes Prediction Model Analysis

1. Logistic Regression Model

Method	Train Accuracy	Test Accuracy
Feature Selection	Train: 78.66%	Test: 76.62%
Feature Extraction	Train: 78.33%	Test: 75.97%
Normal (No Feature Reduction)	Train: 78.82%	Test: 75.97%

Logistic Regression Model Analysis

1. **Feature Selection gave the best generalization**: With a **test accuracy of 76.62%**, it slightly outperformed other methods, indicating that selecting relevant features helped reduce overfitting.

2. **Normal model had highest train accuracy but lower test accuracy**: The **train accuracy of 78.82%** suggests mild overfitting, as the test performance didn't improve.

3. **Feature Extraction didn't significantly help**: It resulted in a slight drop in both train and test accuracy, possibly due to loss of important information during transformation.

2. SVM Model

Feature Strategy	Kernel	Train Accuracy	Test Accuracy
Feature Extraction (n_comp=3)	Linear	Train: 73.15%	Test: 76.62%
Feature Extraction (n_comp=3)	**Poly**	**Train: 76.7%**	**Test: 78.0%**
Feature Extraction (n_comp=3)	Gaussian	Train: 76.05%	Test: 76.62%
Feature Selection (k=6)	Linear	Train: 78.50%	Test: 77.27%

Feature Selection (k=2)	Poly	Train: 76.22%	Test: 76.62%
Feature Selection (k=3)	Gaussian	Train: 78.66%	Test: 75.32%
Simple	Linear	Train: 78.66%	Test: 77.27%
Simple	Poly	Train: 84.52%	Test: 72.72%
Simple	Gaussian	Train: 85.66%	Test: 72.07%

Logistic Regression Model Analysis

1. **Feature Selection with Linear Kernel offers the best balance**: Achieved a solid **train accuracy (78.50%)** and **test accuracy (77.27%)**, indicating good generalization without overfitting.

2. **Simple model with Poly and Gaussian kernels overfits**: Very high train accuracy (**84.52%, 85.66%**) but poor test accuracy (**72.72%, 72.07%**) suggests overfitting due to model complexity.

3. **Feature Extraction with Poly kernel performs best overall**: With **train: 76.7%** and **test: 78.0%**, this setup achieves the **highest test accuracy**, showing the effectiveness of dimensionality reduction + nonlinear decision boundary.

3. Neural Network

Method	Train Accuracy	Test Accuracy
Simple	**Train: 81.56%**	**Test: 76.40%**
Feature Extraction	Train: 79.23%	Test: 72.07%
Feature Selection	Train: 76.71%	Test: 72.72%

Neural Network Model Analysis

1. **Simple model performs best overall**: With the highest **test accuracy (76.40%)**, it shows the model could learn well from raw features without needing reduction.

2. **Feature Selection and Extraction hurt performance**: Both methods led to lower **train and test accuracy**, possibly due to loss of important information during dimensionality reduction.