Machine Learning Project: Cervical Cancer Prediction

Business Objective

The business objective of this project is to develop a classification model which can utilize a patient's past health history to predict whether they have a high risk for cervical cancer.

Dataset

The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The data contains patient information, including health history and responses to personal history survey questions. There are 35 features in the dataset, and the target variable 'Biopsy' was used to classify the patient as high risk vs. not high risk. The target variable is a binary indicator with values of 0 and 1, where 1 indicates high risk. Dataset: https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors.

Modeling

The dataset was evaluated using k means clustering, linear kernel support vector machine (SVM), RBF kernel support vector machine, and logistic regression models. All models underwent hyperparameter tuning using GridSearchCV from scikit-learn. LDA and PCA dimensionality reduction were tested to determine whether either technique would improve model accuracy. The RBF model was selected for further fine tuning of hyperparameters, along with testing the performance of a polynomial kernel.

Results

The RBF kernel SVM model with no additional dimensionality reduction showed the highest accuracy with test data. For most models tested, LDA and PCA dimensionality reduction showed lower accuracy.

	No Additional Dimensionality Reduction		**LDA (1 component)**		**PCA (7 components)**	
Model	Training Accuracy	Test Accuracy	Training Accuracy	Test Accuracy	Training Accuracy	Test Accuracy
K Means	0.9359	0.936	0.965	0.9419	0.9359	0.936
SVM LinearSVC	0.965	0.936	0.965	0.9419	0.691	0.75
SVM RBF	0.9636	0.9535	0.9636	0.9419	0.7959	0.8314
Logistic Regression	0.9636	0.936	0.9621	0.9419	0.691	0.7442
SVM RBF**	0.9636	0.9535	0.9636	0.9419	0.7959	0.8314
SVM Polynomial**	0.9767	0.9477	0.9679	0.9477	0.8848	0.8953

**: Model results after further fine tuning of hyperparameters.

Conclusion

After testing various models, the RBF kernel SVM with no additional dimensionality reduction appeared to perform best for classifying patients as having high risk for cervical cancer. This model resulted in a training accuracy of 0.9636 and a test accuracy of 0.9535.

To improve this model in the future, it may be worth exploring deep learning modeling methods and additional dimensionality reduction techniques. Utilizing this dataset to calculate a cancer risk score rather than classification may also be more beneficial for patient care; however, this must be run in an unsupervised manner as this data is not currently available.