CIND 820 - Proposal (Abstract)

Rama Barham (500903499)

05/21/2024

(a) Problem Context & Chosen Theme:

Breast cancer is one of the most prevalent and life-threatening cancers affecting women worldwide. According to the American Cancer Society, breast cancer is the most commonly diagnosed cancer among women globally, with millions of new cases reported each year. It is also a leading cause of cancer-related deaths among women. Hence, making an early and accurate diagnosis is critical for effective treatment and improved survival rates.

This project aims to analyze the Breast Cancer Wisconsin (Diagnostic) dataset by applying various machine learning (ML) techniques to predict the presence of breast cancer, classifying tumor types, and identifying key features that influence diagnosis. The chosen theme is Predictive Analytics, with a focus on Classification.

(b) Problem being Addressed:

The primary problem addressed in this project is how to accurately classify breast tumors as benign or malignant using machine learning (ML) techniques. The project seeks to answer the following research questions:

- What features are most indicative of whether a tumor is benign or malignant?
- How accurately can tumors be classified based on their computed features?
- What are the key differences in value features between benign and malignant tumors?
- How can visualization techniques help in understanding the distribution and significance of features?

(c) Dataset:

The project will utilize the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository. This dataset includes 569 instances and 32 real-valued attributes. The real-valued attributes were computed from Fine Needle Aspiration (FNA) of breast masses, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

(d) Techniques & Tools:

Relevant Techniques

Classification: Supervised classification, logistic regression.

Feature Selection: Using methods like univariate feature selection to remove irrelevant features. Then apply RFE to identify the most significant features for classification. Use L1 regularization

(Lasso) to further refine feature selection. Then evaluate the selected features using cross-validation to ensure they generalize well.
Model Evaluation: Confusion matrix, accuracy, precision, recall, F1-score.
Data Visualization: Scatter plots, heatmaps, box plots.

*Available Tool*s
Programming Languages: Python will be used for data preparation, modeling, and analysis due to its versatility and extensive libraries for machine learning tasks.
Libraries and Frameworks: Scikit-learn, Pandas, NumPy
Data Processing Tools: Pandas, NumPy
Visualization Tools: Matplotlib, Seaborn, Plotly