PROJECT PROPOSAL

CLASSIFICATION OF TUMORS INTO 2 CATEGORIES (BENIGN OR MALIGNANT) BASED ON FEATURES OF TUMOR CELL NUCLEI

The problem

A tumor can be malignant (cancerous) or benign (non-cancerous). This classification is made by a pathologist who examines the tumor cells obtained from a biopsy under a microscope. One method employed by pathologists is to examine the shape of a cell nuclei and make a judgement based on it.
The problem that is encountered in this regard, is that the shape is defined by many parameters such as the area of a nuclei, it's perimeter and numerous other parameters. Making a judgement on which of these parameters are more important relative to other parameters requires a quantitative analysis of the differences in these parameters between the malignant and benign cases.

The clients and their need for this analysis

The client for this project are cancer pathologists and oncologists. Their decision on tumor classification determines the type of treatment that an oncologist will undertake. The proposed analysis will aid a pathologist in deciding the type of tumor. Also, it will increase the confidence of an oncologist who relies on the report of a pathologist.

Data: What it is and how will it be acquired

The data for this analysis will be a csv file containing 30 features of a cell nuclei of a tumor cell. This will be acquired from the following data base link: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data

Approach outline

In this and the proceeding sections, malignant and benign nuclei will referred to as classes.
1- Read data and store it in a data frame.

2- Perform data wrangling:
- Check data for duplicate values.
- Check data for NaN values.
- Check metadata to determine which of the columns (nuclei features) are informative.
- Check if reshaping (melting or pivoting) the data frame helps in visualization and analysis.
- Normalize the data for training predictive models and visualizing the differences in features between the two classes.

3- Perform visual exploratory data analysis:
- Plot correlation matrix for features to check if there is correlation between different features. This will be important for building predictive models as discussed below.
- Plot box plots overlaid with raw data points to inspect changes in different features between the two classes.
- Calculate percentage changes in features for the two classes relative to each other to gain insight on the features which show the greatest change.
- Plot bar plot for the percentage changes.

4- Perform statistical exploratory data analysis to check which of the features are significantly different between the two classes.
- First check if values of features are normally distributed.
- Based on the test of normalcy, use parametric (values are normally distributed) or non-parametric (values are not normally distributed) statistical test.

5- Based on the analysis performed, there should be enough information to build a coarse visual model showing the difference in nuclei features between the two nuclei.

6- Perform Principal Component Analysis (PCA) if the number of samples compared to number of features is small (which it is as described by the data set's metadata). Moreover, if correlation is seen between different features, this will perform feature extraction and reduce the dimensionality of the data.

7- Build predictive models for the data set:
- Build a k-nearest neighbors model for the data as the first predictive model. Based on the fact that there are two classes with a small number of samples compared to the number of features, this model is a good starting point. Check the accuracy of the model using the f-1 score.

• Build a logistic regression model (logreg) as another predictive model. Check the accuracy of the model using the f-1 score to see if it performs better than the k-NN model.

• Build a logistic regression model (logreg) without performing PCA. This model will provide coefficients for different features for users who want to see the effect of different features on determining the class of a nucleus. Check the accuracy of the model using the f-1 score and compare it to logreg with PCA performed to check whether performing PCA resulted in an improved model.

Deliverables

1- Jupyter notebook with code, markdowns and figures
2- A final analysis report describing the methodology used and the results obtained from the analysis
3- A slide deck summarizing the analysis.