Tumor Cancer Prediction

Project description:

This project focus on predicting whether a tumor is a Benign or malignant using Tumor Cancer Dataset which contain various of attributes that might correlate to having benign or malignant tumor.

Different mathematical models are trained using the dataset, and the prediction of each model is compared, and the result is the common prediction, this method will help us improve the prediction accuracy, and lower the error.

The main idea of the project is using **ensemble learning** to predict the tumor, which require combining the result of

Data Cleaning:

Using pandas library, we create data frame and use build in functions to clean the data set, the operations are:

1.drop Index column so it doesn't affect the model.

2.Removing duplicated rows.

3.Remove rows containing Nan values.

```
df = df.drop(columns='Index', axis=0)
df = df.drop_duplicates()
df = df.dropna(how='any', axis=0)
```

4.use correlation matrix to remove columns with high correlation, because highly correlated features will not improve the model but might even lower its accuracy, also too many features might overfit the model because it requires a more complex model.

```
cor_matrix = df.corr().abs()
upper_tri = cor_matrix.where(np.triu(np.ones(cor_matrix.shape), k=1).astype(bool))
to_drop = [column for column in upper_tri.columns if any(upper_tri[column] > 0.95)]
df1 = df
for i in to_drop:
    df1 = df1.drop(i, axis=1)
```

Models test score:

1.SVM Linear:

predict score is: 0.9635036496350365%

2.Logistic regression:

predict score is: 0.9562043795620438%

3.Bayes:

predict score is: 0.948905109489051%

4.Decision tree:

predict score is: 0.927007299270073%

5.SVM polynomial:

predict score is: 0.9197080291970803%

6.KNN:

predict score is: 0.9051094890510949%

7.SVM RBF:

predict score is: 0.9051094890510949%

8.SVM sigmoid:

predict score is: 0.5547445255474452%

Cumulative score:

predict score is: 0.948905109489051%

Conclusion:

1.picking the model for cumulative prediction is important because it may affect the overall performance.

2.models require larger dataset, so they don’t give the same score and prediction.

3. it require more time and space.