Midterm Project: Data Mining

October 5, 2023

1 BACKGROUND

Classification in predicting patient outcomes involves using machine learning and data analysis to categorize patients into groups based on their characteristics, aiding in forecasting their future health results. The goal of this project is to predict whether breast cancer patients have recurrence events or not.

2 DATASET

You can find the dataset at this link. The dataset includes 201 instances of no recurrence and 85 instances of recurrence. The instances are described by 9 attributes and most are categorical features. For example, the levels of node-caps are yes and no. The levels of the breast-quad are left-up, left-low, right-up, right-low, and central.

- Class: no-recurrence-events, recurrence-events
- age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
- menopause: lt40, ge40, premeno.
- tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
- inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
- node-caps: yes, no.
- deg-malig: 1, 2, 3.
- breast: left, right.
- breast-quad: left-up, left-low, right-up, right-low, central.
- irradiat: yes, no

3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a preliminary data analysis technique that involves summarizing and visualizing data to understand its key characteristics, uncover patterns, and identify potential insights before more formal modeling or hypothesis testing. Common EDA methods include summary statistics, data visualization, correlation analysis, outlier detection, trend analysis, and more. Perform at least two exploratory data analyses and summarize your findings. This is a very open-ended question.

4 DATA PRE-PROCESSING

There are many steps that are needed for most datasets before a Machine Learning algorithm can be implemented on them. For this dataset:

- First, replace the missing values of each feature with the mode of that feature. For example, if the feature node-caps is missing in an instance, replace the missing value with whichever level of node-caps is more common in your data (yes or no). In this data set, missing values are shown with '?'

- Dealing with categorical features: Some machine learning algorithms can handle numerical and categorical data, while others only handle numeric values. Label encoding and one-hot encoding are two common techniques used to convert categorical data into numeric format in machine learning and data analysis. Here's a brief explanation of each:

  Label Encoding:
  Label encoding is used when you have a categorical feature with ordinal relationships, meaning there is a clear order or ranking among the categories. It assigns a unique integer label to each category, where the labels are typically assigned in ascending order of importance or rank. For example, if you have a "Size" feature with categories ["Small", "Medium", "Large"], label encoding might convert them to [0, 1, 2], respectively.

  One-Hot Encoding:
  One-hot encoding is suitable when you have nominal categorical data, where there is no inherent order or ranking among the categories. It creates binary columns for each category and assigns a 1 to the column corresponding to the category of the data point, and 0 to all other columns. For example, if you have a "Color" feature with categories ["Red", "Green", "Blue"], one-hot encoding would create three binary columns, like [1, 0, 0], [0, 1, 0], [0, 0, 1] for "Red", "Green", and "Blue", respectively.

  The original data:

Size	color
Small	red
large	blue
medium	green

  After conversion:

Size	Red	Green	Blue
0	1	0	0
2	0	0	1
1	0	1	0

  You may find Scikit-Learn documentation for label encoder and one-hot encoder. Please note the scikit-learn implementation of decision tree does not support categorical variables for now. Conduct research and convert the categorical features of this data set appropriately.

5 METHODS AND MODELS

Use at least two different classification techniques for predicting patient outcomes, one of which has to be an ensemble method.

- Model overfitting occurs when a machine learning or statistical model learns the training data too well, capturing noise, random fluctuations, or minor details that don't generalize to new, unseen data. This results in a model that performs exceptionally well on the training data but poorly on new, unseen data. To combat overfitting, techniques such as regularization, cross-validation, and increasing the amount of training data are often employed to help models generalize better. Explain if your models are overfitted or not. If they are overfitted, elaborate on your efforts to address it. Again, this can be an open-ended question.

- K-fold cross-validation is a technique used in machine learning to assess the performance of a predictive model. It involves dividing the dataset into K subsets, training and testing the model K times using different subsets for testing each time and then averaging the performance metrics to provide a more robust evaluation of the model's generalization ability. Use K-fold cross-validation for assessing your model performance.

- *This is for graduate students only.* Confusion matrices are often used to evaluate the performance of classification models. They show the number of true positives, true negatives, false positives, and false negatives, which can be useful for assessing model accuracy. Create one confusion matrix for each model used.

6 Conclusion

Include a summary of objectives, model performance results, key findings, and limitations. Offer recommendations for future work and highlight the real-world impact or applications of your model. Conclude with a brief, impactful statement about the project's significance.

7 References

Use the IEEE citation format: numerical citations in square brackets to refer to sources and provide straightforward formatting for references. See IEEE Citation Guidelines.