Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 6

Due: March 29, 2023(100 points)

Name: Jie Zhang

Github Name: JieUpup

Email: zhang.jie4@northeastern.edu

1. What problem are you going to be tackling on your project?

   Contribute to improving breast cancer diagnosis

   Ultimately lead to better treatment outcomes for breast cancer patients.

   Improve the model performance and accuracy.

2. Why is that an interesting/useful application of data mining?

   First, this topic is a meaningful exploration which relates to real life problems.

   Second, the model can be used for medical areas.

   Lastly, it is a good way to practice our data mining techniques in this project.

3. What models/techniques (clustering/classification/etc.) are you envisioning to apply?

Data Preprocessing: We will preprocess the dataset to handle missing values, outliers, and other anomalies. We will also perform feature selection and extraction to identify the most relevant features for diagnosis. The dataset will choose the .cvs type which we are already familiar with.

Model Selection: We will evaluate the performance of various machine learning algorithms, such as K-means, logistic regression, decision trees, and support

vector machines, to identify the best model that can accurately diagnose breast cancer.

Model Tuning: We will perform hyperparameter tuning on the selected machine learning algorithm to improve its performance.

Model Evaluation: We will evaluate the performance of the developed model using various performance metrics, such as accuracy, precision, recall, and F1 score, and compare it with existing state-of-the-art methods.

4. Where are you going to get the data?

https://www.kaggle.com/search?q=breast+cancer+datasetSize%3Alarge+datasetFileTypes%3Acsv

I will choose one dataset in this open source website: Kaggle. The size will be no smaller than 5GB and the type is .cvs.