PROJECT REPORT

Data Preprocessing Project

California State University, Sacramento

CSC 177

Data Preprocessing Project

Prof. Jagan Chidella

Author:

1. Phuc Dinh
2. Moshley Marcelo
3. Brian Shao
4. Steven Ho
5. Alex Souv
6. Navteg Khalsa

To start this assignment, we went through the tutorials posted on Canvas. In this step, we learned different modules and methods used for data preprocessing. After we finished the tutorials, we searched online for datasets we could use for the projects. It took time as we had to find datasets that we were interested in and unclean. We can't find a dataset that can apply all the techniques we see in the tutorials. So, we went with four different datasets:

1. Dataset 1 (London Air)
2. Dataset 2 (UNSW-NB15)
3. Dataset 3 (diabetes)
4. Dataset 4 (Mobile Phone Price)

Through these four datasets, we applied dropping fields, removing rows with missing values, dropping fields, removing duplicated rows, removing duplicated rows, concatenating, calculating fields, Feature Normalization, shuffling, sorting, saving dataframe applied dropping fields, removing rows with missing values, dropping fields, removing duplicated rows, removing duplicated rows, concatenating, calculated fields, Feature Normalization, shuffling, sorting,saving dataframe.

During the process of cleaning data, we faced multiple issues. For example, in the mobile phone price dataset, the values in specific columns were formatted inconsistently. We have to remove inconsistencies before any other techniques; also, after removing inconsistencies, we can remove more duplicated rows. This indicates that some rows have repeated information in different formats, so we can’t remove them until we fix the inconsistencies. In dataset 2, we realized that the ‘id’ column is not necessary, so we remove it, then remove duplicates. We can’t do this before removing the ‘id’ column as every rows has different id. Through this process, we

learned that the order of techniques used is really important, we have to do it in a proper order rather than just apply it randomly

We already learned that data quality is essential and poor data quality is an unfolding disaster. Here, we only deal with small datasets and already see why data preprocessing takes so much time. We must eliminate missing data, unnecessary features, duplicate data, outliers, or inconsistencies. After cleaning the data, our dataset is smaller, more manageable for us to read, and easier for the model to process. Data preprocessing ensures that we and the model are working with accurate and reliable information, which can lead to more accurate predictions.

Data Split

In this part, we decided to work with our diabetes dataset as the dataset is pretty much clean and we want to predict diabetes based on other attributes. Here we use all of the columns except the Outcome as x. And the Outcome column for Y. We split the data, 75% for training and 25% for testing. Here are the mean and std of each dataset:

- Glucose:
    - Training set mean: 120.52604166666667
    - Testing set mean: 122.0
    - Training set std: 31.30199630736171
    - Testing set std: 33.96625982181027
- BloodPressure
    - Training set mean: 68.94965277777777
    - Testing set mean: 69.57291666666667
    - Training set std: 19.113006722286883
    - Testing set std: 20.110555584040394

- SkinThickness
    - Training set mean: 20.73263888888889
    - Testing set mean: 19.947916666666668
    - Training set std: 15.694852459319465
    - Testing set std: 16.72905100325358
- Insulin
    - Training set mean: 81.27256944444444
    - Testing set mean: 75.38020833333333
    - Training set std: 115.65762221613798
    - Testing set std: 114.17978027672025
- BMI
    - Training set mean: 32.01597222222222
    - Testing set mean: 31.922395833333336
    - Training set std: 8.125176445501436
    - Testing set std: 7.131822974665879
- DiabetesPedigreeFunction
    - Training set mean: 0.4719513888888889
    - Testing set mean: 0.47165104166666666
    - Training set std: 0.3358161055745912
    - Testing set std: 0.3183402606796537

Compare the two sets: the training data and the test data and analyze it, developing an intuition and meaning of your results.

Splitting data is an essential step in machine learning. The training set will only be used to train, and the test set (unseen data) will be used to test how well a model is generalized. This process will prevent overfitting when the model performs well on the training set but poorly on the test set. In our case, the mean and std of columns from the two datasets are pretty close together.

Developing and documenting human insights with human interpretation on preprocessed data and possible effect on predictions

We already learned that data quality is essential and poor data quality is an unfolding disaster. Here, we only deal with small datasets and already see why data preprocessing takes so much time. We must eliminate missing data, unnecessary features, duplicate data, outliers, or inconsistencies. After cleaning the data, our dataset is smaller, more manageable for us to read, and easier for the model to process. Data preprocessing ensures that we and the model are working with accurate and reliable information, which can lead to more accurate predictions

Summary

Throughout this project, we learned that data preprocessing is an essential process and it require patient as it will take a lot of time. We also learned that we have to understand the dataset, what we are looking for to apply proper techniques rather than just apply them blindly.