Acute Liver Failure Prediction

Sri Padmini Jayanti
Department of Computer Science
Wayne State Universiry
Detroit, MI, USA
gq6158@wayne.edu

TO: Prof. Dongxiao Zhu
Department of Computer Science
Wayne State University
Detroit, MI, USA
dzhu@wayne.edu

***Abstract*—This dataset called Acute Liver Failure (ALF) is a xlsx datasheet consisting of demographic and health information of Indian adults. It is produced by the JPAC Center of Health Diagnosis and Control and has input values collected through various primary methods. The most optimal Machine Learning method for medical data is the Support Vector Machine has been implemented to analyze and predict risk areas from this dataset.**

Keywords—support vector machine, SVC, prediction, accuracy, visualization

I. INTRODUCTION AND BACKGROUND

This paper in the IEEE official research proposal format is to discuss about the machine learning concepts implemented on the Acute Liver Failure datasheet provided by JPAC Center of Health and Diagnosis. This dataset is exclusive real-time data collected using several methods like direct interviews, examinations and blood samples by interacting with several Indian adults of age 20 years and older. This data ranges from 2008-2009 and 2014-2015 surveys and contains fields that help in detecting cirrhosis, a long-term chronic liver disease or failure. Since the data is as huge as details of 8,785 adults, Support Vector Machine (SVM) has been used to divide it into simpler forms and predict the necessary outputs.

II. PREPROCESSING

A. Choosing the method of implementation

Extensive preprocessing is required for large datasets, especially while using Support Vector Machine methods. Firstly, SVM has been chosen because of its ability to divide the dataset using a hyperplane at any point in the training set. In medical datasets especially, the ability to predict identifications of a particular disease or general classification of a set of rules for a disease to occur can be predicted easily, like in our case of the Acute Liver Failure.

B. Importance of Preprocessing

The data in the ALF dataset had to be studied thoroughly to understand which data is necessary for our findings and which headers in the sheet could be eliminated. This is to improve the robustness of the data which can make our work simpler.

C. Preprocessing procedure

Firstly, exploratory data analysis is conducted to study the dataset in detail. Then, missing values are identified by scanning all the data in the sheet which leads to dropping the unnecessary columns in the dataset. So, here we have dropped the 'ALF' column since it was creating many NULL spaces from its values. Then, features that would be useful and easy to understand for the implementation of SVM are identified and displayed. There were many other missing values in the data even after dropping some columns. So, the missing values were taken care of using the "Imputer" method and "iloc" functions from the sklearn package. "REGION" and "GENDER" columns were strings when all others were numerical. So, encoding had to be done for those columns for them to be synchronized. Finally, according to the project prompt, training and testing sets were divided with respect to the given percentage values.

```
Int64Index: 6000 entries, 0 to 5999
Data columns (total 15 columns):
Age                        6000 non-null int64
Gender                     6000 non-null object
Region                     6000 non-null object
Weight                     6000 non-null float64
Height                     6000 non-null float64
Body Mass Index            6000 non-null float64
Obesity                    6000 non-null float64
Waist                      6000 non-null float64
Maximum Blood Pressure     6000 non-null float64
Minimum Blood Pressure     6000 non-null float64
Good Cholesterol           6000 non-null float64
Bad Cholesterol            6000 non-null float64
Total Cholesterol          6000 non-null float64
Dyslipidemia               6000 non-null int64
PVD                        6000 non-null int64
dtypes: float64(10), int64(3), object(2)
```

Fig a. The table of robust dataset after the preprocessing stages where 30 columns have been reduced to 15 important columns.

III. DISCUSSION OF SVM AND RESULTS

As discussed in the introduction, Support Vector Machine methods are the best fit for medical data predictions and results. Other methods like Logistic regression, linear regression and Naïve Bayes have been successful with predicting results from medical data, but for this particular data set that we have, SVM is beneficial due to concept of greatest margin. The hyperplane concept used to classify data into sets which is implemented in SVM acts as the main winner for our dataset.

A. Using sklearn and scikit packages

Implementing SVM algorithm from scratch is not an easy task. So, with the permission from the term project prompt, extensive usage of sklearn and scikit packages has been done. Firstly, the data is trained using the svm library from the scikit-learn package. Since we are performing a classification task here, we have used svc or the support vector class from the scikit-learn's svm library. Gaussian classification is enough for this data so, we have used 'rbf' kernel for classification. Since there is non-linearly separable data in our datasheet, we had to use any of the kernel SVM has to be used. Gaussian kernel SVM can project the non-linearly separable data lower dimensions to separable higher dimensions so that different types of classes are allocated in different dimensions.

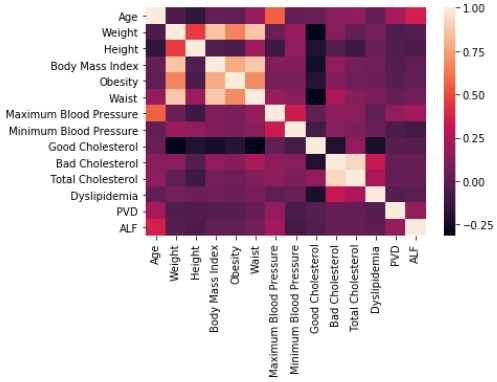

Fig b. The heatmap of the dimensions of the robust and preprocessed dataset showing recall values of several demographics.

B. *Predictions from SVM*

- In order to understand the classification ability of an algorithm, certain parameters of evaluation must be implemented. Some of them are: confusion matrix, precision, recall, classification report and so on.
- Most of these evaluation methods are found in the "metrics" library of the scikit-learn. It contains the "confusion_matrix","classification_report", "accuracy_score" and "precision_recall_fscore_support" to calculate confusion matrix, classification report, accuracy, precision and recall respectively.
- All the results for the evaluation and prediction methods are available in the python file attached. However, the most important result is of the accuracy of the implemented SVM algorithm being nearly 83%.
- The recall is also an important measure to study the data. The higher the recall is, the better the predictions will be.
- For the implemented SVM algorithm, recall value has come up to 0.6 which is a decent strength as a result for a classifier.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

Identify applicable funding agency here. If none, delete this text box.

- Since we are claiming SVM to be the best solution for this dataset, we have implemented Naïve Bayes in another python file to compare the results of both of then to provide proof of the efficiency of the algorithms.
- Accuracy and recall can be considered as the important features to validate the algorithm, especially with medical datasheets, so we will compare SVM and Naïve Bayes with respect to these two evaluations. Naïve Bayes is lower than SVM in both respects with accuracy being 79% and recall being about 0.5
- There is very precise prediction of liver failure with respect to several demographics given in the dataset. Since SVM classifies through a hyperplane, it is much effective for a huge dataset like this and hence Naïve Bayes has been difficult to implement for such a huge dataset.

IV. CONCLUSION/SUMMARY

With a huge dataset like the Acute Liver Failure datasheet we had for this project, strong classifiers need to be implemented in order to have precise predictions. For this purpose, SVM has served to be the best out of all the other researched and implemented methods. Firstly, we preprocess the data thoroughly, plot some heat maps for clarity and evaluate the algorithm based on several evaluation features implemented through scikit-learn and sklearn packages.

REFERENCES

[1] "Top AI algorithms for healthcare", Sciforce. Retreived from: https://medium.com/sciforce/top-ai-algorithms-for-healthcare-aa5007ffa330

[2] K. Rahul, "Acute Liver Failure", Kaggle. Retreived from: https://www.kaggle.com/rahul121/acute-liver-failure

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.