CEN 359 INTRODUCTION TO MACHINE LEARNING

PROJECT DOCUMENTATION

Liver-Disease

Prepared by:

Abdulkarem Khamis

Proposed to:

Elma Avdić, Assist. Prof. Dr.

Zehra Sikira, Teaching Assistant

18/01/2023

Contents

1. Introduction

Liver Disease: It can happen in some conditions where the liver is fully or partially not functioning, it happens due to some reasons:

- Obesity
- Alcohol misuse
- An undiagnosed hepatitis infection
- Polluted water

This disease causes the skin, body fluids, and the white part of the eye to turn into a yellowish, Confusion or other mental difficulties and Severe fatigue, and some other pains…

Normally a person can be diagnosed with liver disease byrunning some tests by doctors such as:

- Laboratory tests.
- Imaging tests.
- Biopsy.

But in our case, we will use the technology (ML) to predict the infected person.

We will apply some ML models to identify some symptoms that are directly related to liver-disease to diagnose a patient.

2. Problem Formulating

Since we have seen how people’s lives are being affected by pains that they suffer so it is advisable to use the technology and its resources such as machine learning to predict and to enhance their lives.

So, we are going to utilizing supervised learning techniques such as Logistic-Regression and Supported Vector Machine (SVM) to classify whether a tested person is classified with liver-disease or not.

To get a better explanation we should read some formulization definitions such as the definition that was mentioned by Tom Mitchell where he said,

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

So, according to the above definition we should define our “Task, Experience, and the performance”

Task: our problem is classifying patients as having a liver disease or not by relying on our data points on dataset. We define our task as the problem that we need to classify a patient as having a liver-disease or does not have by relying on our data points on dataset.

Experience: from dataset we could know what factors are mainly related to the disease, and according to that experience we can apply our algorithms. And lastly taking into consideration how our models are being effective or not.

Performance: define the accuracy of our models.

As for each dataset there should be some important aspects related to it, so as for our data: we have 10 features (9 Numerical, 1 categorical) and 1 label (categorical), mostly all of them are important.

And to better use the dataset we must use data pre-processing to be suitable for our algorithms.

3. Methods

For a better explanation, we will have to apply several steps.

The objective of the steps is to use machine learning techniques to improve the performance of predicting the liver-disease class.

There are 4 steps and some sub-steps:

- Data collection.
- Data pre-processing (data analysis, non-values-handling, data visualization, data mapping, data scaling, data splitting).
- Models.
- Models evaluation/testing.

Data Collection:

Our dataset was imported from Kaggle.

The data was collected by an Indian researcher (for only Indians), so we will utilize our dataset for training, testing, and evaluating models.

4. Pre-processing Data:

Data analysis

In this stage we imported the dataset and viewed the shape of the dataset and the columns (features) to know how to proceed with the data in the pre-processing stage.

Feature	*Datatype*	*Description*
Age	*numerical*	*Respondent age*
Gender	*categorical*	*Respondent gender*
Total-Bilirubin	*numerical*	*is a yellowish pigment that is made during the breakdown of red blood cells, highly than usual indicate having liver-disease*
Direct-bilirubin	*numerical*	*High levels in blood may indicate that liver isn't clearing bilirubin properly*
Alkaline-Phosphatase	*numerical*	*high levels is a sign of a liver problem*
Alkaline-Phosphatase	*numerical*	*high levels is a sign of a liver problem*
Alanine-Aminotransferase	*numerical*	*being released into the bloodstream by damaged cells and the high level indicate a liver damage, the normality is 4 to 36 U/L*
Total-Protiens	*numerical*	*the quantity in Liver, the normal range is 6.0 to 8.3 grams per deciliter (g/dL), low-level indicates to (liver disease) or the protein isn't being digested or absorbed properly, high-level indicates to dehydration.*
Albumin	*numerical*	*checking liver functionality, low level refers to a liver disease, normality is 3.4 to 5.4 g/dL*
Albumin_and_Globulin_Ratio	*numerical*	*Albumin made in liver, Globulin helps fighting infections and move nutrients throughout the body, both of them are proteins found in the blood, the normality ratio is 1 to 2 low level is a sign for a liver problem*
Dataset(output)	*Numerical(binary)*	*No liver-Disease: = 0, patients having liver-Disease: =1*

Table 1: Description of dataset.

The collected dataset originally contained 583 data points for 10 input (numerical and categorical) features and 1 target label (output). The target label includes two classes [2, 1], 2 = healthy, 1 = is patient.

Null handling

The data contained 4 observations having nulls in “Albumin_and_Globulin_Ratio” feature, and since it is only small number, we can simply remove them. The final number of observations is 579.

Data visualization

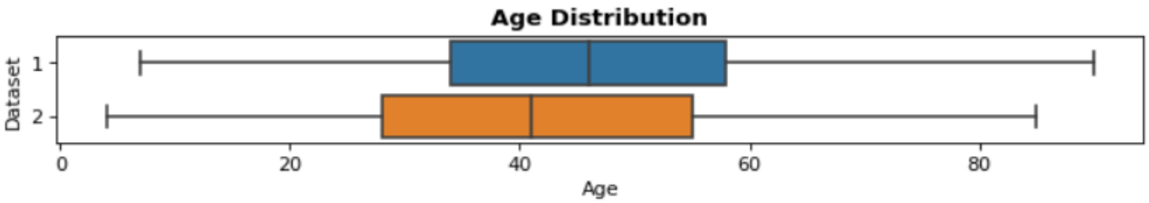
In this stage we viewed the distribution of the numerical data, and the ratio of the categorical data (Gender and Output).

a) Normal distribution: A normal distribution is symmetric and the mean, median and mode are equal where the predictors, or independent variables, follow a normal distribution, when the predictors are normally distributed, it means that the majority of the data points are concentrated around the mean, and fewer data points as you move away from the mean. This allows logistic regression to make accurate predictions by estimating the parameters of the model.

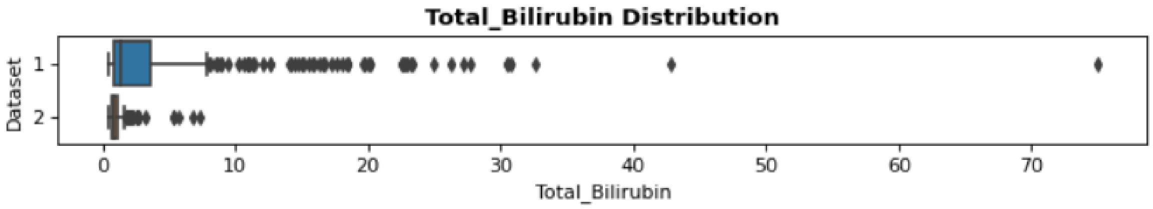
b) Right-Skewed distribution (positive skew): Where we get more data points on the right side of the distribution and fewer data points on the left side, this can lead to a bias in the model towards the majority class and make it difficult for the model to correctly classify the minority class, we will refer to it as "RSD".

c) Left-Skew-distribution (negative Skew): Where we get more data points on the left side of the distribution and fewer data points on the right side, this can lead to a bias in the model towards the minority class and make it difficult for the model to correctly classify the majority class, we will refer to it as "LSD".

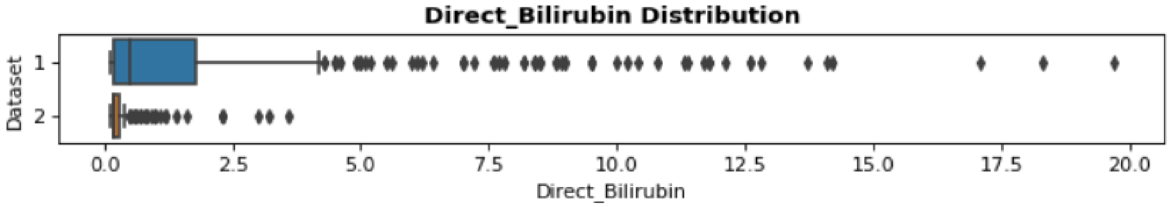
d) Outliers: means that there are some observations in the dataset that are significantly different from most of the data points.
Outliers can cause the model to estimate parameters that are not representative of most of the data, leading to biased results. This can cause the model to make incorrect predictions and have poor generalization performance.



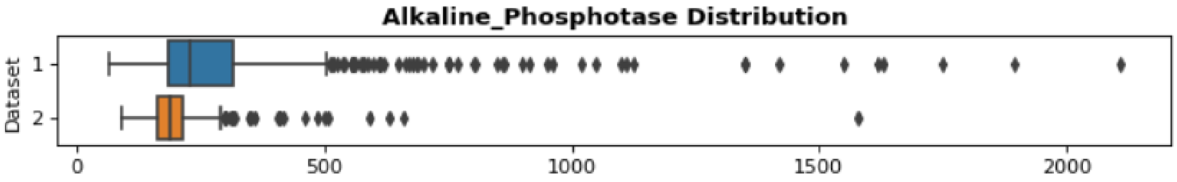
a. **Both are normally distributed**



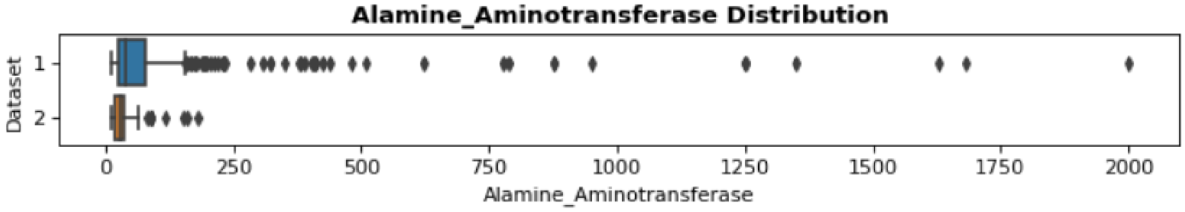
b. 1:RSD and with too many outlier
2: RSD with many outliers



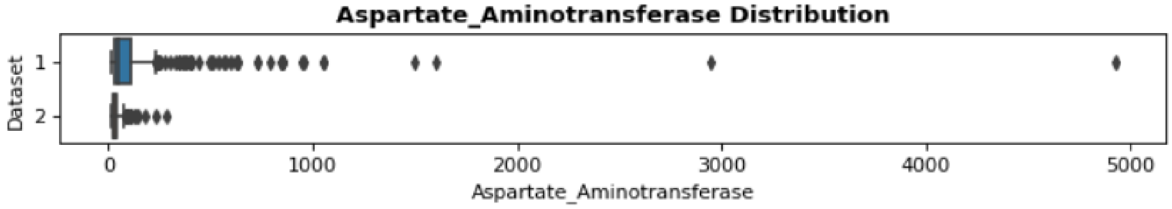
c. 1. RSD with too many outliers
2. LSD with too many outliers



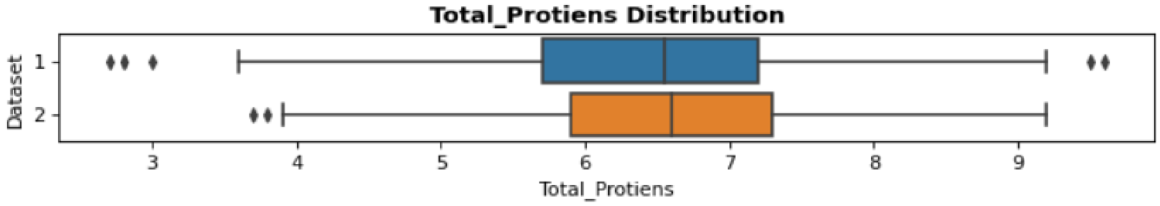
d. 1. RSD with too many outliers
2. Normal distribution with a lot of outliers



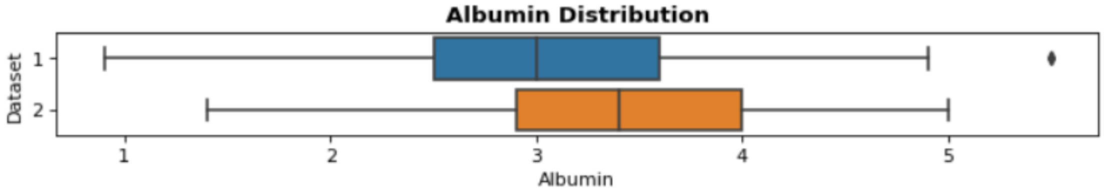
e. 1. RSD with too many outliers
2. LSD with a few outliers



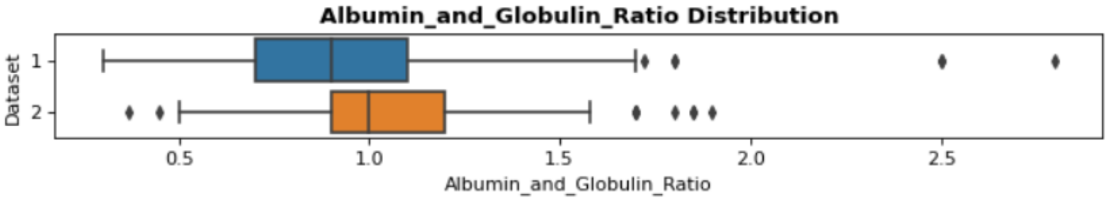
f. 1. RSD with too many outliers
2. LSD with a few outliers



g. 1: LSD with a very few outliers
2: Normal Distribution with 0 outlier



h. 1. RSD with 1 outlier
2. RSD with no outliers



L. 1. Normally distributed with a few outliers

2. RSD with a few of outliers with a few outliers

Figure 1: Numerical distribution of all numerical features

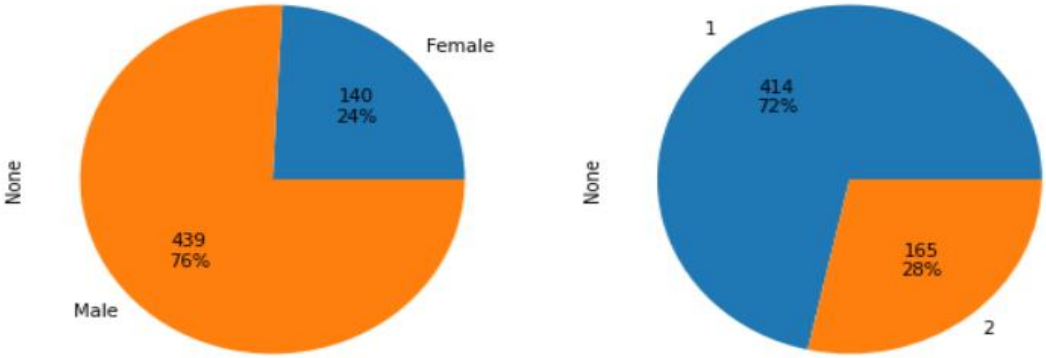


Figure 2:gender and output ratio (categorical features); Gender feature to the left, Output to the right

Pie Chart: it is a type of chart that is used to represent data in a circular format, with each section of the pie representing a different category of data. The size of each section is proportional to the quantity of data it represents. Pie charts are often used to visualize data that is broken down into parts of a whole, such as percentage of total or relative frequencies.

We visualized our categories using Pie Chart since we have a limited number of categories (2 categories), so it would be the best representative.

First chart shows the gender representation of data, 24% Female, 76% Male, (140 out of 579 Female) and (439 out of 579 Male)

Second chat shows the diagnoses (output) representation of the data on total and the healthy people on total, (72% patient) and (28% healthy) in other words 414 out of 579 are patient and 165 out of 579 are healthy.

Data mapping

We have put the [dataset] label into a new list so that we can use it sufficiently when we test our models.

In this stage we mapped the categorical data, Output data was mapped to be: 0: healthy and 1: Patient.

Also, Gender feature was mapped using get-dummies () method which is part of *oneHotEncoder* library which convert categorical data into a binary features.

get-dummies () generate features for each category in the gender's feature, so we had to take one feature out because it contains the same data as the other, so if one feature is 0 then the other will be automatically 1.

We used the classification library to show the classification report for each model.

Data scaling

For numerical features we have used scaling, *standardScaler* () to eliminate biases of all models to sum features on the expense of others (the model without it will give more importance to the features with higher values where we want to eliminate that).

First, we trained *MinMaxScaler* method, but it gives bad accuracy for all models then we used *standardScaler* method.

Data splitting

In the data splitting stage, our heart disease dataset is divided into a 60% training set, 20% validation set, and 20% as the testing set. The training set is utilized to train and make the model learn the hidden features/patterns in the data. The validation set is utilized to validate our model performance during training. Validation set is a set of data which is separate from the training set. The test set is a separate set of data used to test the model after completing the training.

5. Models

We will be using two supervised machine learning algorithms to classify patients:

- ✓ Logistic Regression.
- ✓ SVM (support-vector-machine).

Logistic Regression: is a classification model and it is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes.

SVM: Support vector machines (SVMs) are a type of supervised machine learning algorithm that can be used for classification or regression tasks. The goal of an SVM is to find the hyperplane in a high-dimensional space that maximally separates the different classes. In the case of binary classification, the SVM finds the hyperplane that separates the two classes with the largest margin, which is the distance between the closest data points in each class and the hyperplane.

In the case of multi-class classification, the SVM finds the hyperplane that separates the different classes by constructing a one-vs-rest model for each class

i. Process of Models Training

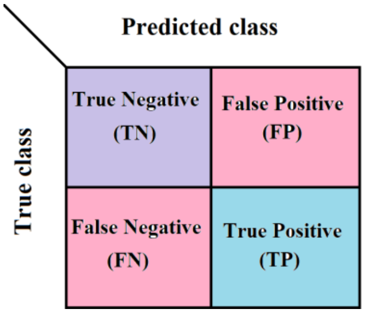
- Firstly, we report each library for each model then we defined the models with their parameters and then we trained them.
- We let the models to predict the output of the test set then we evaluate the model using the evaluation on what we said on the evaluation section
- So, shortly our models work as follows
    - Library importing
    - Defining parameters
    - Training
    - Predicting
    - Testing and evaluation of the prediction.

6. Models Evaluation

- Accuracy: is one of the most common evaluation metrics used in classification tasks. It is calculated as the proportion of correctly classified instances out of the total number of instances in the test set.
- Testing: refers to the process of evaluating the performance of a machine learning model on a set of data that it has not seen before the purpose is to is to estimate the model's ability to generalize to new data and to identify any potential issues or biases in the model.

- Evaluating: is the process of assessing the performance of a machine learning model using metrics and techniques, to determine the quality of the model and its ability to generalize to new data.

- Having the confusion matrix as an evaluation method to evaluate the results we get from models.
- Confusion matrix is utilized for the performance evaluations of the methods used after the classification. For binary classification, the scheme of the confusion matrix is below.



True class	Predicted class	
	True Negative (TN)	False Positive (FP)
	False Negative (FN)	True Positive (TP)

Figure 3 confusion matrix example.

WHERE:

0: N

1: P

True Positives (TP): we correctly predicted that they do not have illness.

True Negatives (TN): we correctly predicted that they do have illness.

False Positives (FP): we incorrectly predicted that they do not have illness.

False Negatives (FN): we incorrectly predicted that they do have illness.

The objective of confusion matrix is mainly used to evaluate the performance of classification models, it allows us to see number of correct and incorrect predictions made by the model and help us to know where the model is making mistakes.

The confusion matrix compares the predicted values from the model with the actual values from the test data set

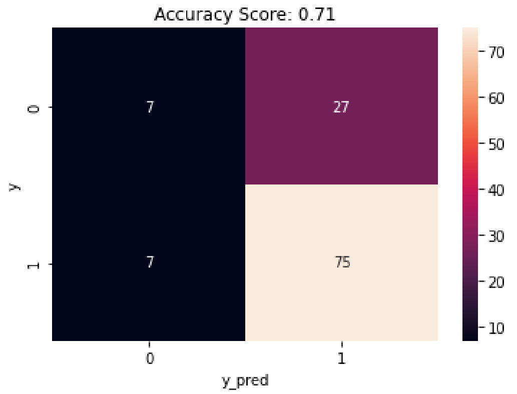
The accuracy calculated using these performance metrics according to the values in the confusion matrix which is done according to equations below.

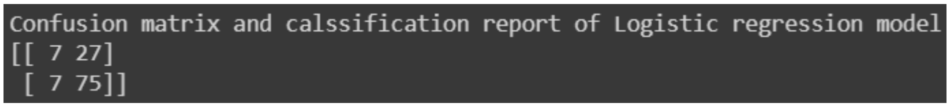
The accuracy: TN+TP/Total.

Total = total number of observations.

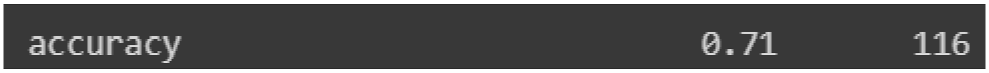
Log-Reg:

Test:



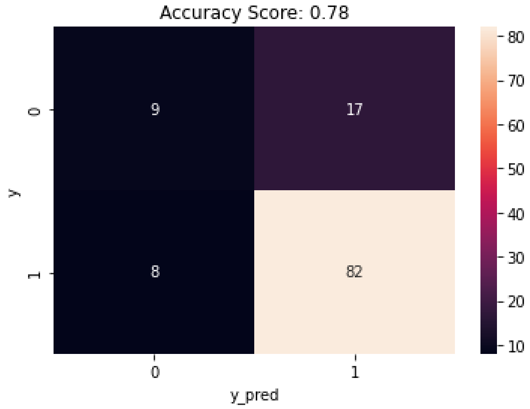


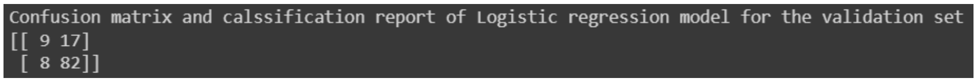
```
Confusion matrix and calssification report of Logistic regression model
[[ 7 27]
 [ 7 75]]
```



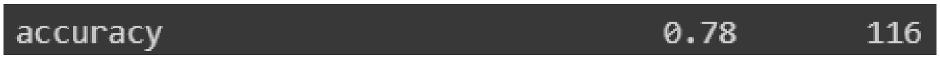
```
 accuracy                           0.71        116
```

Validation:





```
Confusion matrix and calssification report of Logistic regression model for the validation set
[[ 9 17]
 [ 8 82]]
```



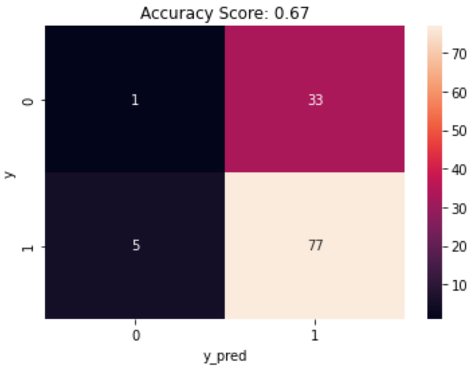
```
accuracy                           0.78        116
```

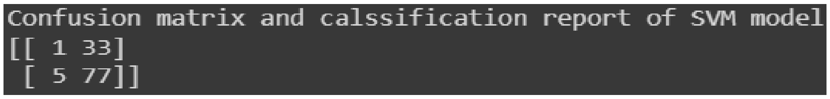
We have got the below results after the pre-processing stage and applying this model to our data.

Log-Reg	Accuracy
Test	0.71
Validation	0.78

SVM:

Test:

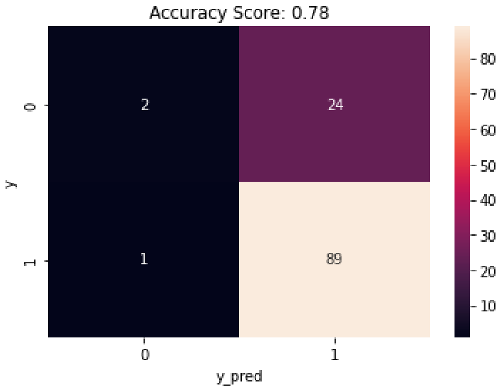


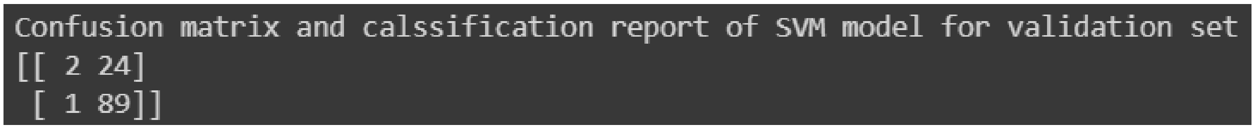


```
Confusion matrix and calssification report of SVM model
[[ 1 33]
 [ 5 77]]
```

```
 accuracy                           0.67        116
```

Validation:





```
Confusion matrix and calssification report of SVM model for validation set
[[ 2 24]
 [ 1 89]]
```

```
 accuracy                           0.78        116
```

We have got the below results after the pre-processing stage and applying this model to our data.

SVM	Accuracy
Test	0.67
Validation	0.78

7. Results

To state what is the best model

Based on results on both the accuracy of each model on both test set and evaluation set we found out that the Log-Reg is the most suitable model for this dataset with accuracy of 71 on test and 78 on evaluation set.

8. Feature selection:

I chose to not apply feature selection (feature importance) due to small number of features. And usually feature selection causes a drop of accuracy and we want to avoid it because of the relatively low accuracy of our models.

Limitation:

Due to the lack of the small number of observations we could not get a valid neural network model (it is a naïve predictor which based on random chance) we knew it by looking to the result of the Neural-Networks (NN) where it showed us that all of them are being diagnosed to be patient.

9. Conclusion

In this project I have utilized two learning machine language techniques, which are Logistic-Regression and SVM (support vector machine).

And we assessed their accuracy in identifying liver-disease.

The Accuracy of a given test set for a classifier is the percentage of test set instances that are classified correctly by using the classifier. The classifier will classify the data set which is being tested.

10. GitHub

https://github.com/Kareem-kh97/Liver-Disease-Machine-Learning

11. IDE:

I have used Colab (Collaboratory) to train the models since it is runs on the cloud and it does not need to be downloaded on my laptop as an Editor, since other IDEs require to set up a local development environment and they may run properly or not.

12. References

https://www.kaggle.com/code/harisyammnv/liver-disease-prediction/data

https://scikit-learn.org/stable/modules/svm.html

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

https://towardsdatascience.com/an-introduction-to-preprocessing-data-for-machine-learning-8325427f07ab

https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing