Prediction of Liver Fibrosis

line 1: Dhruv Sharma
line 2: *Master's in Computer Science*
line 3: *Bishop University*
line 4: Sherbrooke, Canada
line 5: DSHARMA19@ubishops.ca

line 1: Rajiv Israni
line 2: *Master's in Computer Science)*
line 3: *Bishop University*
line 4: City, Country
line 5: RISRANI19@ubishops.ca

line 1: 3rd Aishwarya Naik
line 2: *Master's in Computer Science)*
line 3: *Bishop University*
line 4: City, Country
line 5:ANAIK19@ubishops.ca

line 1: 4th Apoorva Tikle
line 2: *Master's in Computer Science)*
line 3: *Bishop University)*
line 4: City, Country
line 5: ATIKLE19@ubishops.

***Abstract*—** LIVER Fibrosis prediction is one of the challenging and widely explored research field in machine learning. Hepatitis C virus (HCV) infection affects more than 170 million people worldwide. Egypt has the highest prevalence of hepatitis C in the world with prevalence rates reaching 13%-15%. Thus, HCV represents a major public health and economic problem in Egypt. HCV infection is marked by high tendency to persistence and evolution to chronic hepatitis with development of serious consequences such as cirrhosis and liver cancer in some patients. The aim of this project is to use statistical learning(machine learning) to predict liver fibrosis by discovering the hidden predictive patterns from medical databases. This paper covers how a relevant feature extraction method can help to achieve high performance in Liver Fibrosis.

Keywords*—Machine-Learning,Statistics,Training,Feature Extraction,Classifiers,Accuracy.*

I. Introduction

Hepatitis C is a viral infection that causes liver inflammation, sometimes leading to serious liver damage. The hepatitis C virus (HCV) spreads through contaminated blood. Until recently, hepatitis C treatment required weekly injections and oral medications that many HCV-infected people couldn't take because of other health problems or unacceptable side effects. That's changing. Today, chronic HCV is usually curable with oral medications taken every day for two to six months. Still, about half of people with HCV don't know they're infected, mainly because they have no symptoms, which can take decades to appear. For that reason, the Centers for Disease Control and Prevention recommends a one-time screening blood test for everyone at increased risk of the infection. The largest group at risk includes everyone born between 1945 and 1965 — a population five times more likely to be infected than those born in other years. fter decades of hepatitis C infection, cirrhosis may occur. Scarring in your liver makes it difficult for your liver to function. A small number of people with hepatitis C infection may develop liver cancer. Advanced cirrhosis may cause your liver to stop functioning.

II. Ease of Use

Protect yourself from hepatitis C infection by taking the following precautions.Stop using illicit drugs, particularly if you inject them. If you use illicit drugs, seek help.Be cautious about body piercing and tattooing. If you choose to undergo piercing or tattooing, look for a reputable shop. Ask questions beforehand about how the equipment is cleaned. Make sure the employees use sterile needles. If employees won't answer your questions, look for another shop.Practice safer sex. Don't engage in unprotected sex with multiple partners or with any partner whose health status is uncertain. Sexual transmission between monogamous couples may occur, but the risk is low.

A. Cause

Hepatitis C infection is caused by the hepatitis C virus (HCV). The infection spreads when blood contaminated with the virus enters the bloodstream of an uninfected person.Globally, HCV exists in several distinct forms, known as genotypes. Seven distinct HCV genotypes and more than 67 subtypes have been identified. The most common HCV genotype in the Egypt is type 1.Although chronic hepatitis C follows a similar course regardless of the genotype of the infecting virus, treatment recommendations vary depending on viral genotype. Long-term infection with the hepatitis C virus is known as chronic hepatitis C. Chronic hepatitis C is usually a "silent" infection for many years, until the virus damages the liver enough to cause the signs and symptoms of liver disease. The following symptoms are Bleeding easily, Bruising easily,Fatigue,Poor appetite, Yellow discoloration of the skin and eyes,Dark coloured urine,Itchy Skin,Fluid Buildup in your abdomen,swelling in legs,Weight Loss,Confusion,Drowsiness,Slurred Speech,SpiderLike blood vessels on the skin.

III. Related work and benchmarking study

Good benchmarks are very important for quick, quantitative and fair means of analyzing different learning algorithms in the field of deep learning. To understand neural networks and its implementations, we went through large number of papers and literature. This report covers the effective techniques followed to improve performance on datasets EMNIST and MNIST.

A. Classifiers

Random forest classifier:
from sklearn.ensemble import RandomForestClassifier

A random forest is a meta-estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses mean to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.
A diverse set of classifiers is created by introducing randomness in the classifier construction. The prediction of the ensemble is given as the averaged prediction of the individual classifiers.
Random forest classifier can be used for both classifications as well as regression functionalities. It provides higher accuracy. It handles missing values and maintains the accuracy of a large proportion of data.
Random forest algorithm can handle binary features, categorical and numerical features. There is tiny bit of pre-processing which needs to be done. Rescaling and transformation of data is not needed in here.
In this, we can split the process to multiple machines to run. This results in faster computation time.

Random forest is great with high dimensional data since we are working with subsets of data.

B. **Logistics Regression**

from sklearn.linear_model import LogisticRegression

Logistic regression is a linear model for classification rather than regression. Logistic regression is also known as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the predictions of possible outcomes of a single trial are modeled using a logistic function. This implementation can fit binary, One-vs-Rest, or multinomial logistic regression or Elastic-Net regularization. Logistic Regression is a Machine Learning classification algorithm which is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, True etc.) or 0 (no, failure, False etc.).

we can *transform* are **linear regression** curve to a **logistic regression** curve because our **linear regression** curve won't fit our *binary group* models properly and our **logistic regression** curve can only go in range of **0 to 1** and which is the key to understand **classification** using **logistic regression** curve. The sigmoid function also known as the logistic function is going to be the key to using logistic regression to perform classification. The sigmoid function takes in any value as a input and outputs it in between 0 and 1.

C. ***Gaussian Naïve Bayes***

from sklearn.naive_bayes import GaussianNB

GaussianNB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

The parameters sigma (Variance) and mu (mean) are estimated using maximum likelihood. Gaussian (or Normal distribution) is the easiest to work with because you only need to estimate the mean and the standard deviation from your training data.
Gaussian distribution, is probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

Gaussian distributions have some unique properties that are valuable in analytic studies. For instance, any linear combination of a fixed collection of normal deviates is a normal deviate. Many results and methods can be derived analytically in explicit form when the relevant variables are normally distributed.

D. ***ADA Boosting***

from sklearn.ensemble import AdaBoostClassifier

An AdaBoost [1] classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. he core principle of AdaBoost is to fit a sequence of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction. The data modifications at each so-called boosting iteration consist of applying weights, , …, to each of the training samples. Initially, those weights are all set to, so that the first step

simply trains a weak learner on the original data. For each successive iteration, the sample weights are individually modified and the learning algorithm is reapplied to the reweighted data.

E. ROC Curve

from sklearn.metrics import roc_curve

A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate.
ROC curves are frequently used to show in a graphical way the connection between sensitivity and specificity for every possible cut-off for a test or a combination of tests. In addition, the area under ROC curve gives an idea about the benefits of using number of tests.
The curves of different models can be compared directly in general or for different thresholds.
The area under the curve (AUC) can be used as a summary of the model skill.
The function takes both the true outcomes (0,1) from the test set and the predicted probabilities for the 1 class. The function returns the false positive rates for each threshold, true positive rates for each threshold.

IV. Results

A. Random Forest with combining of ADA Boosting to increase the performance of the model.

Random Forest with using ADA boosting accuracy =0.5240384615384616

B. Doing Comparison without Random Boosting

Random Forest without boosting accuracy = 0.49759615384615385

C. Combing Logistic Regression with ADA Boosting

Logistic regression with using ADA boosting accuracy =0.53125

D. Gaussian Naïve Bayes without ADA Boosting

Gaussian Naive Bayes = 0.5408653846153846

E. Average Score

Average score of rand for = 0.46747207903780075
Average score of log reg =0.48913230240549826

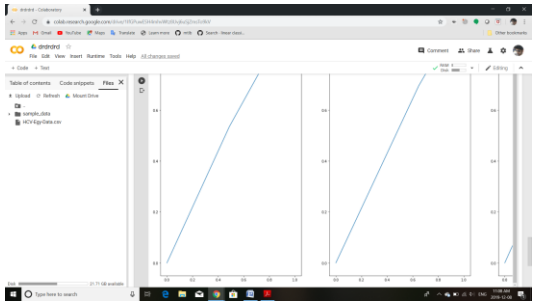
The best result we got is the Gaussian Naïve Bayes with 54% to predict the Liver Fibrosis.

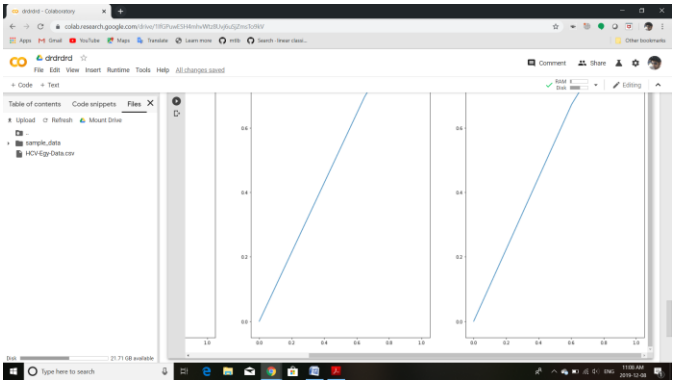
The following top 10 features are:

	Specs	Score
26	RNA EF	1.042778e+06
22	RNA Base	3.752706e+05
24	RNA 12	3.462333e+05
23	RNA 4	3.166259e+05
25	RNA EOT	3.565566e+04
13	Plat	5.611844e+03
11	RBC	4.565762e+02
10	WBC	9.229231e+01
15	ALT 1	5.850758e+00
2	BMI	3.939171e+00

F. Figures and Tables

ROC CURVE: Text(0.5, 1.0, 'area under roc curve= 0.53')





Acknowledgment

In this model we got to know about the prediction of Liver Fibrosis and top 10 features through which we can detect the prediction.The following output we got is 54% with Gaussian NaïveBayes,following further with Logistic Regression,Random Forest,with the use of ADA Boosting.

References

1.Random Forest through scikit learn documentation:

https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

2.Logistic Regression classifier

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

3.Gaussian Naive Bayes:

https://scikitlearn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

4.ADA Boosting

https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html