AI-Statistical Machine Learning Approaches to Liver Disease Prediction

Team ID: PNT2022TMID48272

Faculty Mentor:

D.**Pradhiba**

Team Leader: G.lydia

Team Member: R.priya

Team Member: U.lavanya sri

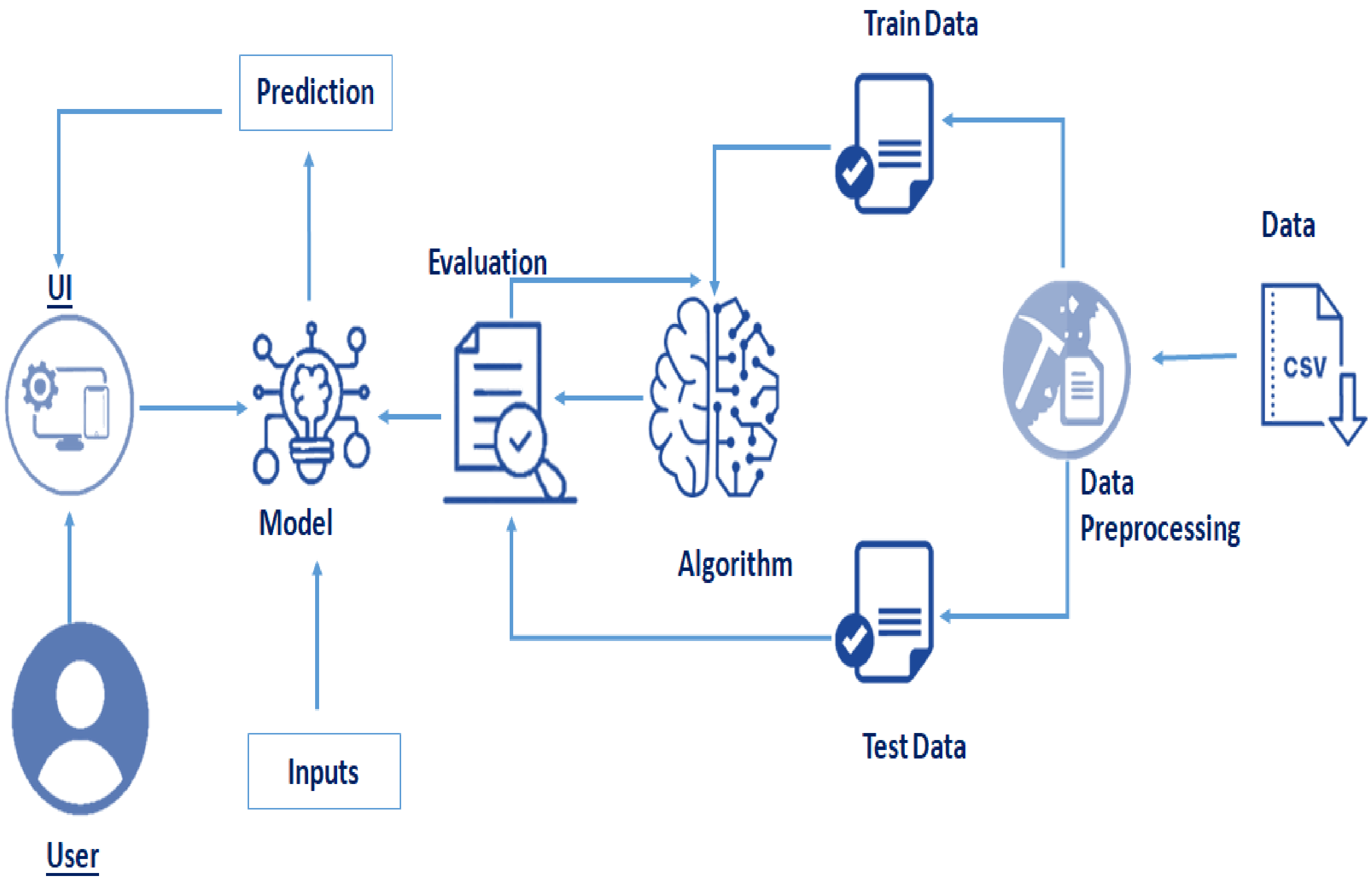
Team Member: G.nagalakshmi

Problem Definition:

Liver diseases avert the normal function of the liver. Mainly due to the large amount of alcohol consumption liver disease arises. Early prediction of liver disease using classification algorithms is an efficacious task that can help the doctors to diagnose the disease within a short duration of time. Discovering the existence of liver disease at an early stage is a complex task for the doctors. The main objective of this project is to analyze the parameters of various classification algorithms and compare their predictive accuracies so as to find out the best classifier for determining the liver disease.This Project examines data from liver patients concentrating on relationships between a key list of liver enzymes, proteins, age and gender using them to try and predict the likeliness of liver disease. Here we are building a model by applying various machine learning algorithms find the best accurate model. And integrate to flask based web application. User can predict the disease by entering parameters in the web application. Medical diagnoses have important implications for improving patient care, research, and policy. For a medical diagnosis, health professionals use different kinds of pathological methods to make decisions on medical reports in terms of the patients' medical conditions. Recently, clinicians have been actively engaged in improving medical diagnoses. The use of artificial intelligence and machine learning in combination with clinical findings has further improved disease detection. In the modern era, with the advantage of computers and technologies, one can collect data and visualize many hidden outcomes such as dealing with missing data in medical research. Statistical machine learning algorithms based on specific problems can assist one to make decisions. Machine learning (ML), data-driven algorithms can be utilized to validate existing methods and help researchers to make potential new decisions. The purpose of this study was to extract significant predictors for liver disease from the medical analysis of 615 humans using ML algorithms. Data visualizations were implemented to reveal significant findings such as missing values. Multiple imputations by chained equations (MICEs) were applied to generate missing data points, and principal component analysis (PCA) was used to reduce the dimensionality. Variable importance ranking using the Gini index was implemented to verify significant predictors obtained from the PCA. Training data ($n_{train} = 399$) for learning and testing data ($n_{test} = 216$) in the ML methods were used for predicting classifications. The study compared binary classifier machine learning algorithms (i.e., artificial neural network, random forest (RF), and support vector machine), which were utilized on a published liver disease data set to classify individuals with liver diseases, which will allow health professionals to make a better diagnosis. The synthetic minority oversampling technique was applied to oversample the minority class to regulate overfitting problems. The RF significantly contributed ($p < 0.001$) to a higher accuracy score of 98.14% compared to the other methods. Thus, this suggests that ML methods predict liver disease by incorporating the risk factors, which may improve the inference-based diagnosis of patients. Keywords: liver disease; demographic variables; prognostic/biochemical variables; statistical learning for variable selection andclassification The liver has many functions such as glucose synthesis and storage, detoxification, production of digestive enzymes, erythrocyte regulation, protein synthesis, and various other features of metabolism. Chronic liver diseases include chronic hepatitis, fibrosis, and cirrhosis. Hepatitis can occur from viral infection (e.g., hepatitis c virus) or auto-immune origin.

Inflammation from hepatitis infection can cause tissue damage and scarring to occur in the liver. Moderate scarring is classified as fibrosis, while severe liver damage/scarring is classified as cirrhosis. Fibrosis and cirrhosis can also occur from alcoholism and non- alcoholic fatty liver disease. When liver disease is diagnosed at an earlier stage, in between infection and fibrosis but before cirrhosis, liver failure can be avoided. Tests, such as a CMP and biopsy, can be conducted to diagnose all forms of liver disease. A CMP with a liver function panel can detect albumin (ALB), alkaline phosphatase (ALP), alanine amino-transferase (ALT), aspartate amino-transferase (AST), gamma glutamyl-transferase (GGT), creatine (CREA), total protein (PROT), and bilirubin (BIL). Diagnosis of a certain liver disease and discovery of its origin are made by interpreting the patterns and ratios of circulating liver-associated molecules measured with the CMP test and compared to values normalized with a patient's age, sex, and BMI. Aminotransferases, AST, and ALT are enzymes that participate in gluconeogenesis by catalyzing the reaction of transferring alpha- amino groups to ketoglutaric acid groups. AST is found in many tissue types and is not as specific to the liver but may denote secondary non-hepatic causes of liver malfunction. ALT is found in high concentrations in the cytosol of liver cells. Liver cell injury cancause the release of both aminotransferases into circulation. When ALT is significantly increased in proportion to ALP, the liver disease is likely from an inflammatory origin(acute or chronic viral hepatitis and autoimmune disease)

Technical Architecture :



How to Preprocess Data in Python:

Here's a step-by-step tutorial on data preprocessing implementation using Python, NumPy and Pandas…In this article, we'll prep a machine learning model to predict who survived the Titanic. To do that, we first have to clean up our data. I'll show you how to apply preprocessing techniques on the Titanic data set.

To get started, you'll need:

- Python
- NumPy
- Pandas
- The titanic data set

For machine learning algorithms to work, it's necessary to convert **raw data** into a **clean data** set, which means we must convert the data set to **numeric data**. We do this by encoding all the **categorical labels** to column vectors with binary values. **Missing values**, or NaNs (not a number) in the data set is an annoying problem. You have to either drop the missing rows or fill them up with a mean or interpolated values. **Note**: Kaggle provides two data sets: training data and results data. Both data sets must have the same dimensions for the model to produce accurate results.

Paper-01

Published in: IEEE International Conference on Systems, Man and Cybernetics

Date of Conference: 06-09 October 2002

Date Added to IEEE *Xplore*: 06 February 2003

Print ISBN:0-7803-7437-1

Print ISSN: 1062-922X

INSPEC Accession Number: 7658757

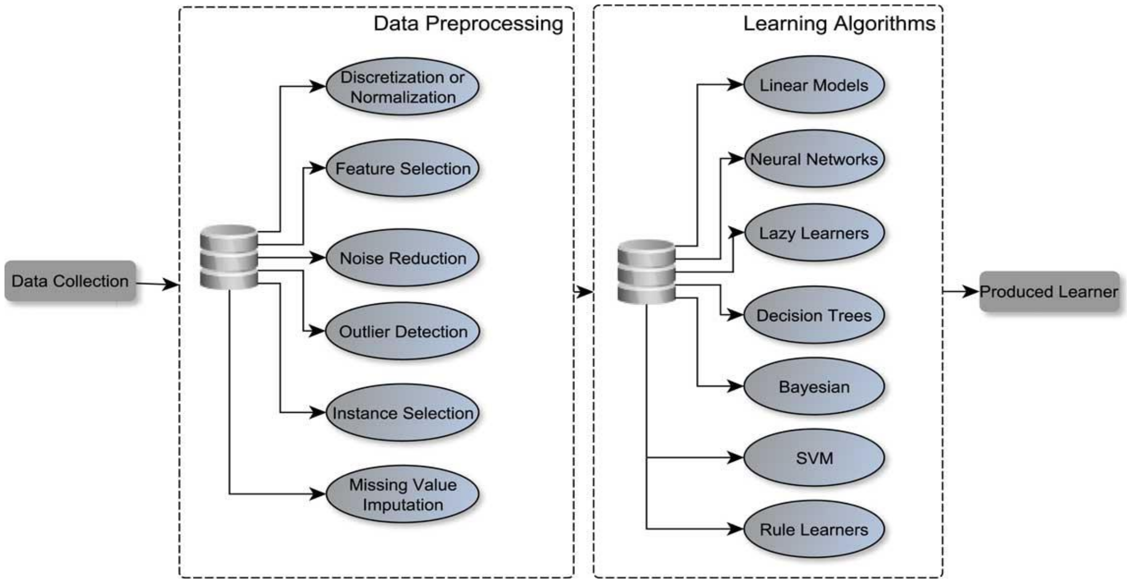
DOI: 10.1109/ICSMC.2002.1176327

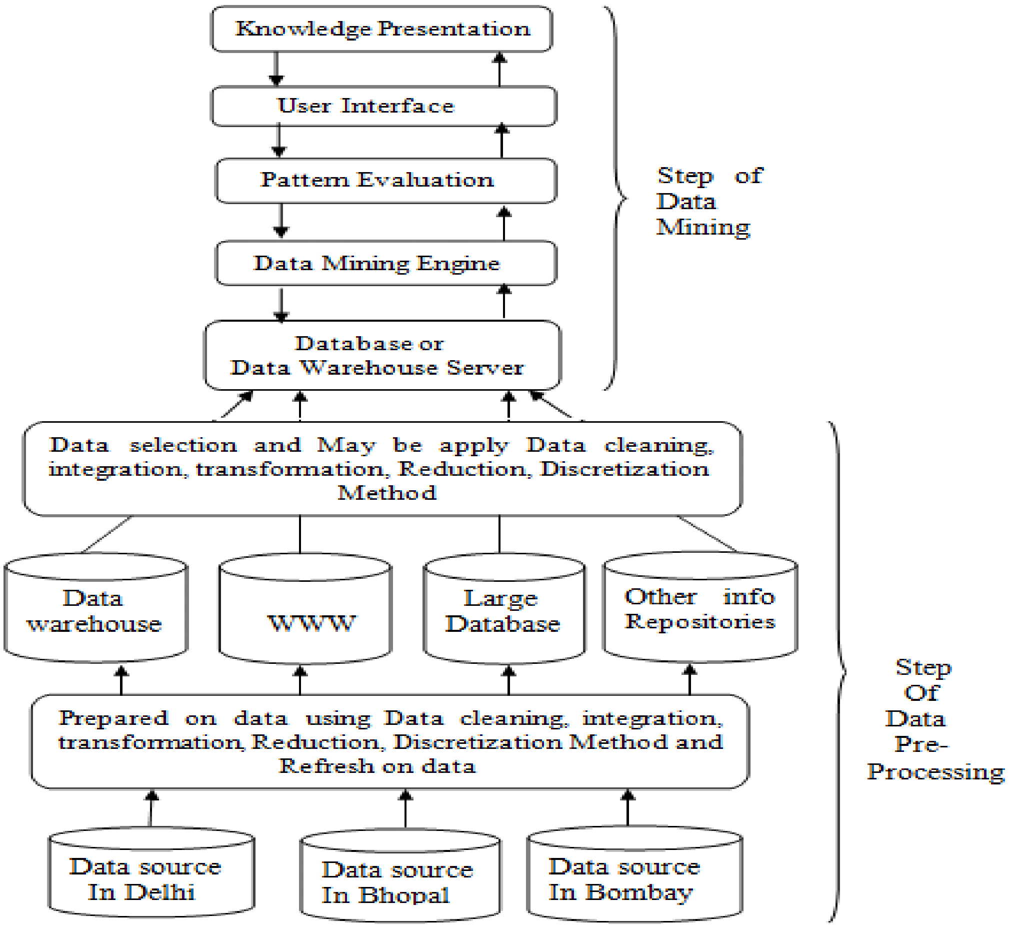
Publisher: IEEE Institute of Electrical and Electronics Engineers

Conference Location: Yasmine Hammamet, Tunisia

Data pre-processing support for data mining:

It is well known that success of every data mining algorithm is strongly dependent on the quality of data processing. In this context it is natural that data pre-processing can be a very complicated task. Sometimes, data pre-processing takes more than half of the total time spent by solving the data mining problem. The paper describes a tool called SumatraTT, the goal of which is to make the process of data pre-processing easier and faster. Basically, SumatraTT (Transformation Tool) is a metadata-driven, platform independent, extensible, and universal data processing tool. These features have been achieved by building the tool as an interpreter of a transformation-oriented scripting language called SumatraScript. SumatraScript a is fully interpreted Java-like language combining together data access, metadata access, and common programming constructions. Furthermore, it supports RAD (Rapid Application Development) technology by providing the library of re-usable transformation templates. The second part of the paper contains a practical application of SumatraTT. It is a task aimed at prediction of water consumption in a regional distribution network.


Data Collection
Data Preprocessing
Discretization or Normalization
Feature Selection
Noise Reduction
Outlier Detection
Instance Selection
Missing Value Imputation
Learning Algorithms
Linear Models
Neural Networks
Lazy Learners
Decision Trees
Bayesian
SVM
Rule Learners
Produced Learner


Knowledge Presentation
User Interface
Pattern Evaluation
Data Mining Engine
Database or Data Warehouse Server
Step of Data Mining
Data selection and May be apply Data cleaning, integration, transformation, Reduction, Discretization Method
Data warehouse
WWW
Large Database
Other info Repositories
Prepared on data using Data cleaning, integration, transformation, Reduction, Discretization Method and Refresh on data
Data source In Delhi
Data source In Bhopal
Data source In Bombay
Step Of Data Pre-Processing

Paper-02

Published in: 2021 IEEE Region 10 Symposium (TENSYMP)

Date of Conference: 23-25 August 2021

Date Added to IEEE *Xplore*: 04 October 2021

ISBN Information:

ISSN Information:

INSPEC Accession Number: 21298123

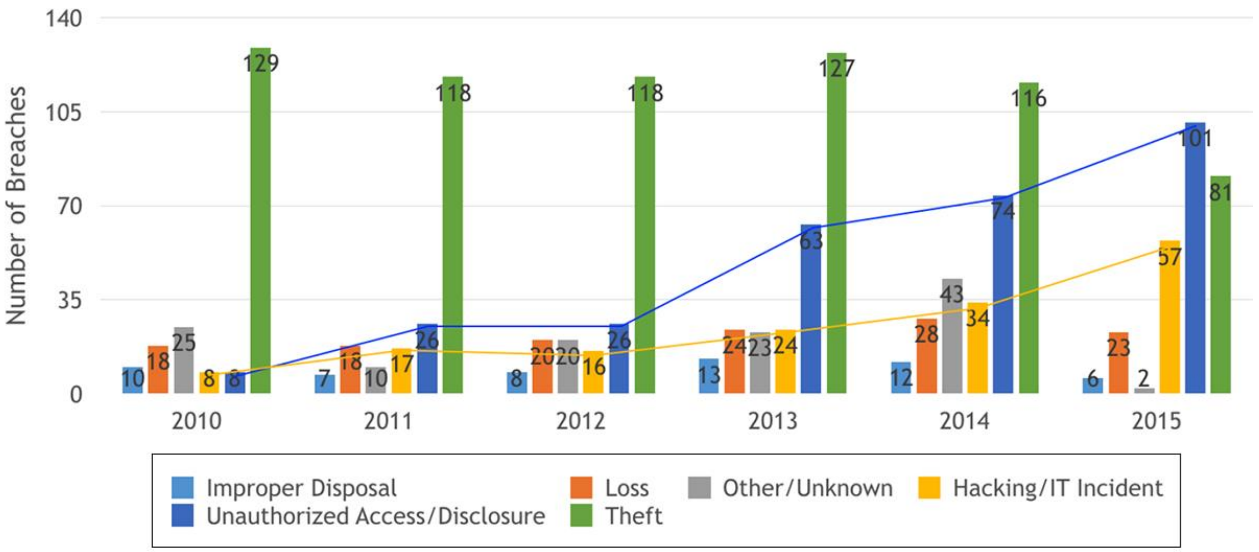
DOI: 10.1109/TENSYMP52854.2021.9550822

Publisher: IEEE

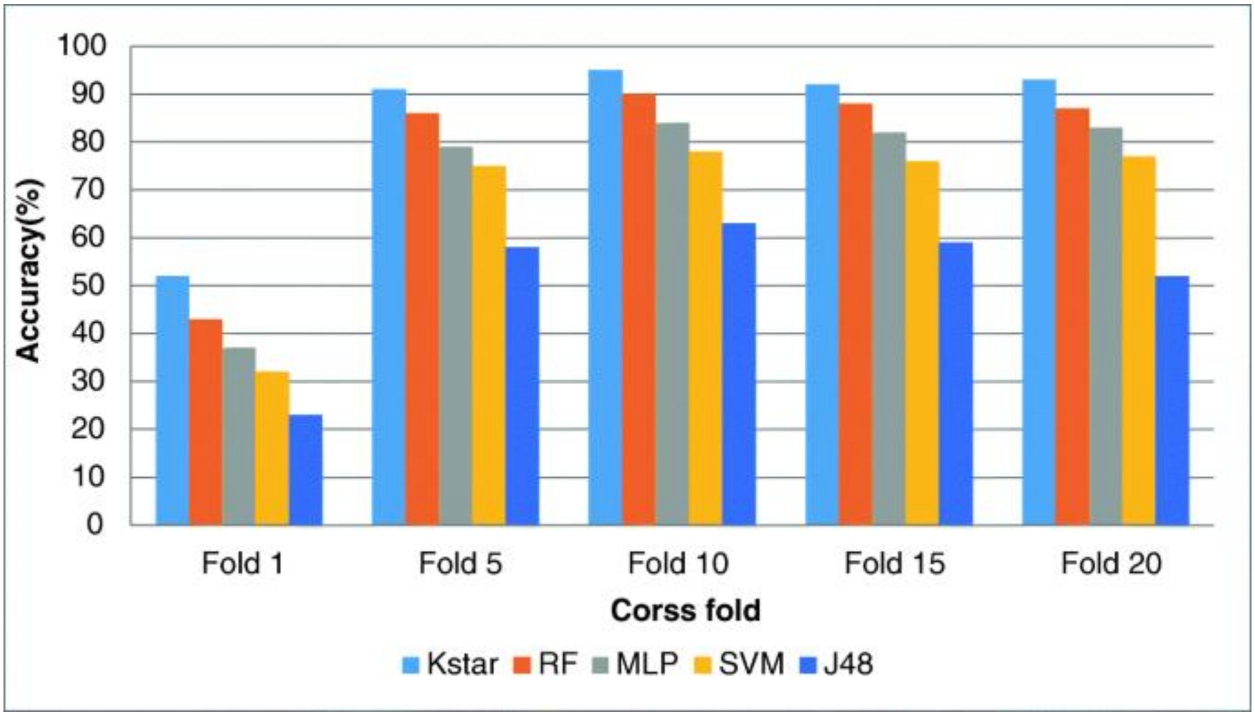
Conference Location: Jeju, Korea, Republic of

Data Visualization and Predictive Analysis for Smart Healthcare:

The healthcare industry is one of the most significant sources of Big Data. It is not feasible to manually interpret and understand the huge amounts of data generated by hospitals accurately. This creates the need for a data analytics and visualization tool. Visualizations are intuitive and help interpret the data easily. It would help the hospital to get insights from the data and to provide better service to the society. The aim of this project was to develop a data analysis and visualization tool for a hospital. This was implemented as a web application. The web application is developed using Django, which is a Python-based free and open-source web framework. For the visualizations embedded in the application, the Python library Altair was used. The application supports the upload of files that are the source for the visualizations and provides interactive visualizations based on the analysis performed. The visualizations can be exported as images using the application. Generating visualizations of choice becomes easier using navigation by menu bars in the application rather than writing complicated queries to the database. The tool ultimately attempts to help the hospital optimize time and resources effectively. Prediction using Long Short Term Memory (LSTM) for pharmacy orders and number of orders per patient will further help the hospital predict trends, patterns and outliers. Analysis tools will help analysing past, current and predict future pharmacy and diagnostics in a hospital which ultimately lead to better quality, efficient smart healthcare.



Source: US DHHS (2019)



Paper-03

Published in: 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)

Date of Conference: 28-30 April 2021

Date Added to IEEE *Xplore*: 04 June 2021

ISBN Information:

Electronic ISBN:

Print on Demand(PoD) ISBN:

INSPEC Accession Number: 208----

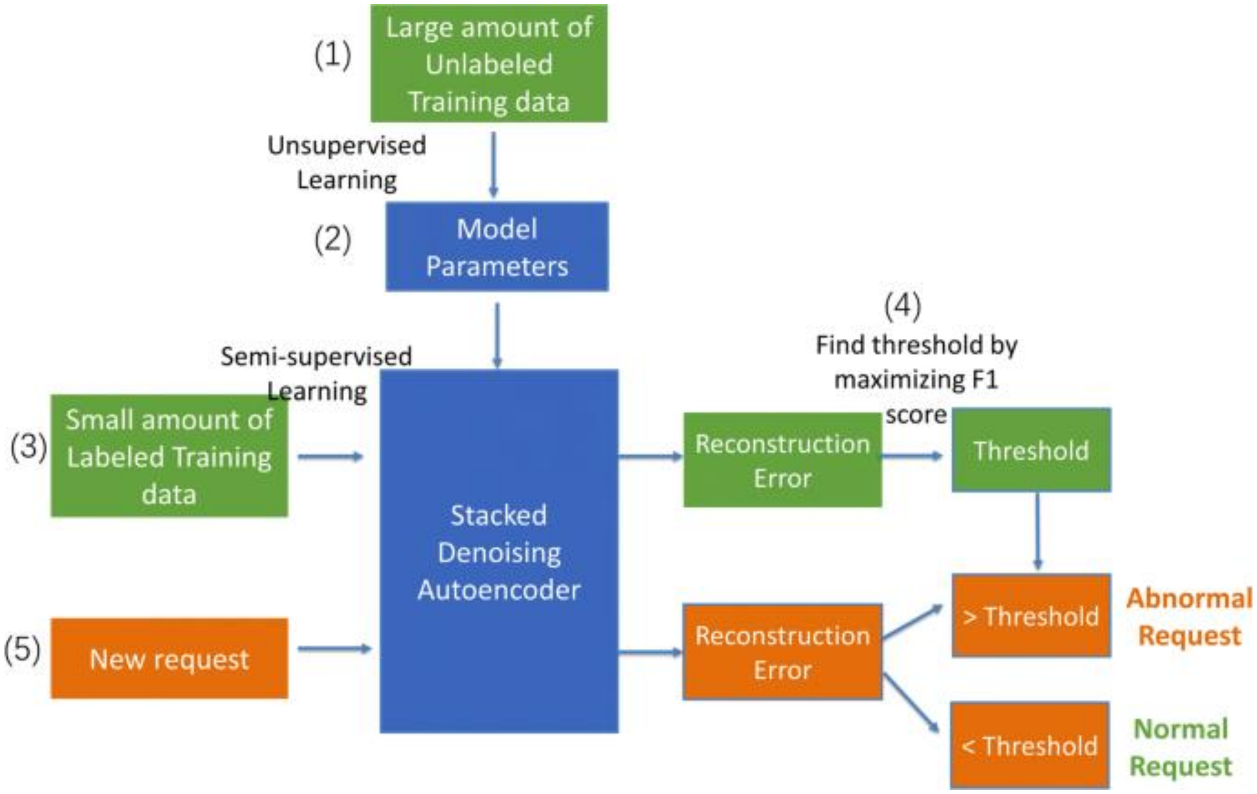
DOI: 10.1109/ICIEM51511.2021.9445368

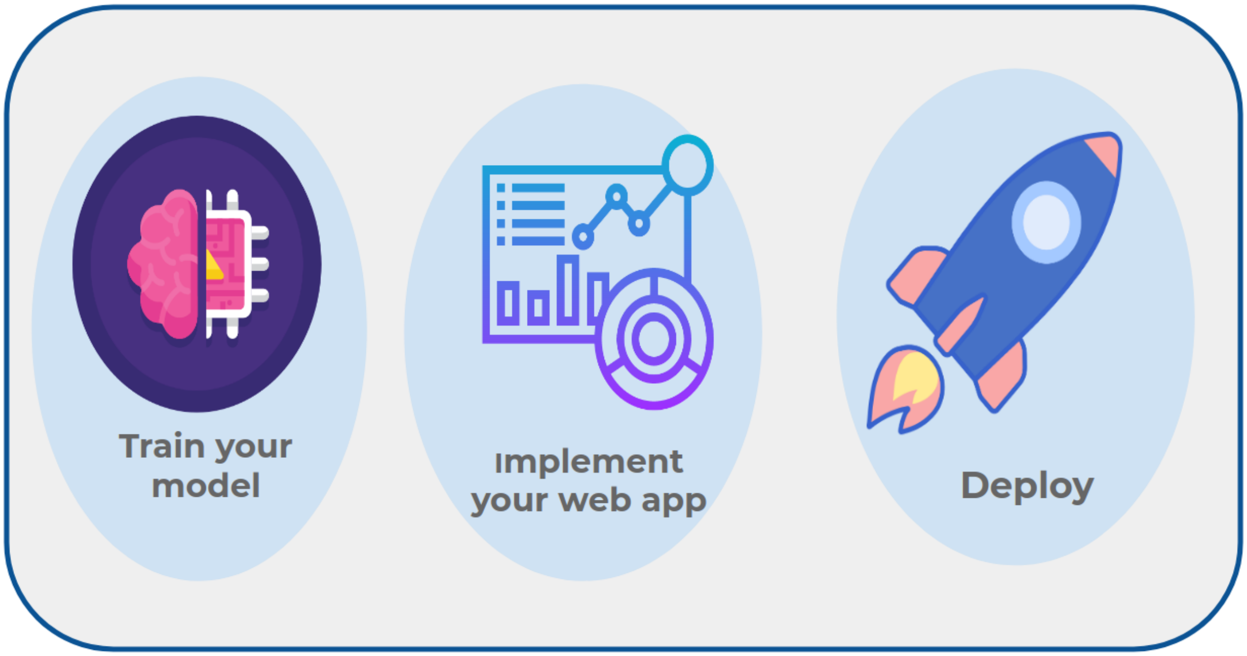
Publisher: IEEE

Conference Location: London, United Kingdom

Web Application Implementation with Machine Learning:

Every Service nowadays has applications. If we want to order food but don't like to talk to someone, we simply go for the web apps or mobile apps and ordered food and it is the more convenient and easy method. so basically, the web application provides a virtual platform where we do lots of tasks from anywhere all around the globe. Here we make a responsive college community web application that helps the college students, faculty, and alumni to interact on one platform. As the objective of this paper is to elaborate the web application working functionality like the creation of fronted part using React Js, backend part by using Django framework, use of the database in web apps to store data in the database, fetch and serve data in the user interface, deployment of a website in the cloud, all components integration and the main part is the use of machine learning where we create a Machine learning NLP Model for text analysis and using Scikit learn library for using python and using flask to deployed in Web apps. and alumni to interact on one platform. As the objective of this paper is to elaborate the web application working functionality like the creation of fronted part using React Js, backend part by using Django framework, use of the database in web apps to store data in the database,


(1)
Large amount of Unlabeled Training data
Unsupervised Learning
(2)
Model Parameters
Semi-supervised Learning
(3)
Small amount of Labeled Training data
Stacked Denoising Autoencoder
(4)
Find threshold by maximizing F1 score
Reconstruction Error
Threshold
(5)
New request
Reconstruction Error
> Threshold
Abnormal Request
< Threshold
Normal Request


Train your model
Implement your web app
Deploy

Paper-04

Authors: Fahad Mostafa

December: 2021

DOI:10.3390/livers1040023

Published in: 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)

ISBN Information:

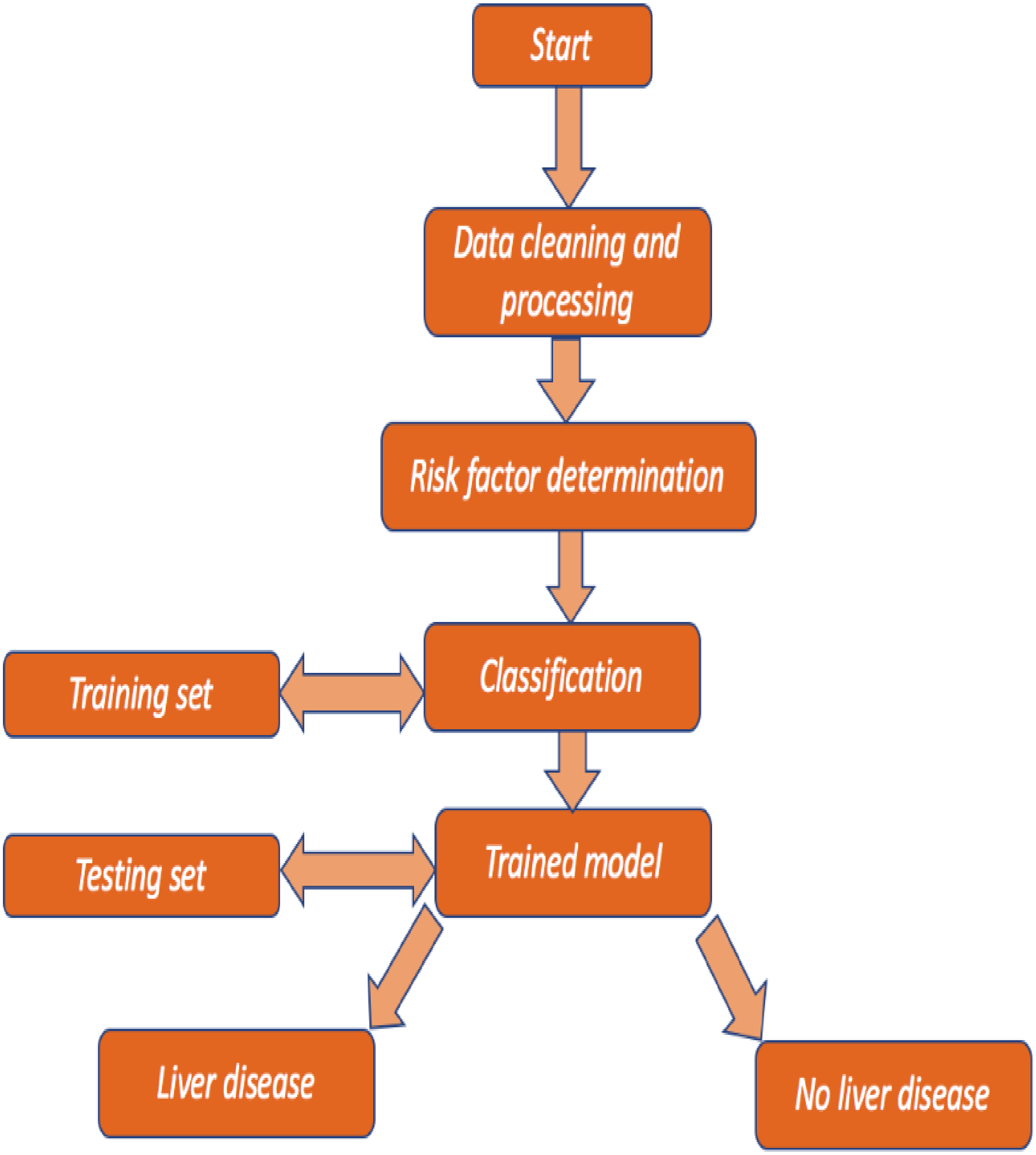
Electronic ISBN:

Statistical Machine Learning Approaches to Liver Disease Prediction:

Medical diagnoses have important implications for improving patient care, research, and policy. For a medical diagnosis, health professionals use different kinds of pathological methods to make decisions on medical reports in terms of the patients' medical conditions. Recently, clinicians have been actively engaged in improving medical diagnoses. The use of artificial intelligence and machine learning in combination with clinical findings has further improved disease detection. In the modern era, with the advantage of computers and technologies, one can collect data and visualize many hidden outcomes such as dealing with missing data in medical research. Statistical machine learning algorithms based on specific problems can assist one to make decisions. Machine learning (ML), data-driven algorithms can be utilized to validate existing methods and help researchers to make potential new decisions. The purpose of this study was to extract significant predictors for liver disease from the medical analysis of 615 humans using ML algorithms. Data visualizations were implemented to reveal significant findings such as missing values. Multiple imputations by chained equations (MICEs) were applied to generate missing data points, and principal component analysis (PCA) was used to reduce the dimensionality. Variable importance ranking using the Gini index was implemented to verify significant predictors obtained from the PCA. Training data ($ntrain = 399$) for learning and testing data ($ntest = 216$) in the ML methods were used for predicting classifications. The study compared binary classifier machine learning algorithms (i.e., artificial neural network, random forest (RF), and support vector machine), which were utilized on a published liver disease data set to classify individuals with liver diseases, which will allow health professionals to make a better diagnosis. The synthetic minority oversampling technique was applied to oversample the minority class to regulate overfitting problems. The RF significantly contributed ($p < 0.001$) to a higher accuracy score of 98.14% compared to the other methods. Thus, this suggests that ML methods predict liver disease by incorporating the risk factors, which may improve the inference-based diagnosis of patients. Keywords: liver disease; demographic variables;

prognostic/biochemical variables; statistical learn- ing for variable selection andclassification The liver has many functions such as glucose synthesis and storage, detoxification, production of digestive enzymes, erythrocyte regulation, protein synthesis, and various other features of metabolism. Chronic liver diseases include chronic hepatitis, fibrosis, and cirrhosis. Hepatitis can occur from viral infection (e.g., hepatitis c virus) or auto-immune origin. Inflammation from hepatitis infection can cause tissue damage and scarring to occur in the liver. Moderate scarring is classified as fibrosis, while severe liver damage/scarring is classified as cirrhosis. Fibrosis and cirrhosis can also occur from alcoholism and non- alcoholic fatty liver disease. When liver disease is diagnosed at an earlier stage, in between infection and fibrosis but before cirrhosis, liver failure can be avoided. Tests, such as a CMP and biopsy, can be conducted to diagnose all forms of liver disease. A CMP with a liver function panel can detect albumin (ALB), alkaline phosphatase (ALP), alanine amino-transferase (ALT), aspartate amino-transferase (AST), gamma glutamyl-transferase (GGT), creatine (CREA), total protein (PROT), and bilirubin (BIL). Diagnosis of a certain liver disease and discovery of its origin are made by interpreting the patterns and ratios of circulating liver-associated molecules measured with the CMP test and compared to values normalized with a patient’s age, sex, and BMI. Aminotransferases, AST, and ALT are enzymes that participate in gluconeogenesis by catalyzing the reaction of transferring alpha- amino groups to ketoglutaric acid groups. AST is found in many tissue types and is not as specific to the liver but may denote secondary non-hepatic causes of liver malfunction. ALT is found in high concentrations in the cytosol of liver cells. Liver cell injury cancause the release of both aminotransferases into circulation. When ALT is significantly increased in proportion to ALP, the liver disease is likely from an inflammatory origin(acute or chronic viral hepatitis and autoimmune disease)

Study Design:



Critical Findings:

The application can be extended to include scanning of barcode on the price tag which decreases the effort of entering the data in the input fields.

A notification system can be enabled in case when the expenses crosses over the income generated by the user to warn him or her about the situation.