**Data Science**

# Project Title: Hepatitis C Prediction

# Data Set Name: Hepatitis C Prediction Dataset

# Data set Link: <https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset> The data was obtained from UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/HCV+data>

# Data Set Description: The data set contains laboratory values of blood donors and Hepatitis C patients and demographic values like age. All attributes except Category and Sex are numerical. The target attribute for classification is Category. The data contains 615 observations and 14 attributes of blood donors and Hepatitis C patients laboratory and demographic values.

# Description of Work Done: The data set was collected from Kaggle. Multiple pre-processing steps were applied on the data for imputation, oversampling and handling data-imbalance. The pre-processed data was stored in a new data-sheet by the name

# CP\_imputed and this pre-processed data was used for the further application so in order to avoid the same process all over again. Later, we implemented RF algorithm which showed the predicted Category variable and the actual Category variable in the same table. Then KNN was employed which gave 99.38% accuracy which showed it was an overfitting model. The irrelevant columns were checked once again and the Serial Number column was removed from the dataset. To avoid the repetition of same steps for multiple algorithms and making the algorithm easy to use as well as to ensure reusability, we used TrainControl method which comprises of k-fold cross validation method to avoid over-fitting and to obtain a refined model. This model was still giving high accuracy of 97%. Then the testdata was oversampled and then the actual final model was ready which gave accuracy of 87.50%, 88.59% and 85.33% respectively. Feature selection method was then employed which usrd rf function for finding the 5 most relevant independent variables. The further computations was done using this new dataset. Logistic regression and SVM were further applied to train the existing model and gave good accuracies of 89.40% and 86.41% respectively. The imputation, feature selection and the main code were later combined to get a single complete code.

# Literature Survey:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sr. no | Title | Journal name and Year | Dataset Name | Data Pre-processing Techniques | Class Imbalance Problems | Algorithms and Accuracy | Evaluation parameters | Feature Selection |
| 1 | Diagnosing the Stage of Hepatitis C Using Machine Learning | Hindawi | HCV-Egy-Data from UCI machine learning repository | mean imputation  Normalization: Yes | Yes | Artificial Neural Network | Precision Miss rate Mean square error | Stage of Fibrosis |
| 2 | Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques | bioengineering/MDPI | NHANES and UCI datasets | Median imputation Data standardization | No | SVM, KNN,  Logistic Regression, DT, XGBoost, ANN | confusion matrix, accuracy, precision, F1 score recall, Area Under Curve, Receiver Operating Characteristics | Embedded method technique |
| 3 | Prediction and Staging ofHepatic Fibrosis in Children with Hepatitis C Virus: A Machine Learning approach | HIR (Healthcare Information Research) | Egyptian children with only HCV | Data cleaning utilizes RF Normalization: NO | No | Random Forest | AST APRI FIB-4 | cutoff, AUC+-SE Sensitivity Specificity |
| 4 | Hybrid model for precise hepatitis‑C classification using improved random forest and SVM method | Scientific reports | UCI HCV dataset | Verifying attributes SMOTE | YES | ANN SVM KNN IRF | Precision Recall F-measure Accuracy | Ranker Algorithm is used to score and rank the dataset features |
| 5 | Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity. | International Journal of Environment Research and public health | Study of adult persons from the north-eastern and southeastern regions of Romania | data cleaning |  | SVM , Naive Bayes, Random Forest, KNN | Accuracy | NA |
| 6 | Hepatitis C Virus prediction based on machinelearning framework: a real-world case study in Egypt | Knowledge and Information Systems | real-world dataset of hepatitis C prevalence among HCWs in Egypt | Normalisation:No |  | NB, RF, Logistic Regression | Accuracy | SFS feature selection method |
| 7 | Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance | Hindawi Computational and Mathematical Methods in Medicine | The data cannot be made freely available. | Data cleaning |  | .Requests for access to these data should be made to [Xiangyong Li, lxy2005123@126.com] | Precision, Sensitivity, F-score, AUC | NA |
| 8 | Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection | MDPI (2023) | The Hospital of Jordan University | Data cleaning |  | (KNNs), Decision tree, Random forest , Neural Network , SMOTE | Accuracy, Recall, Precision, F-measure, R-squared | Sequential feature selection |
| 9 | The scientific progress and prospects of hepatitis C research from 2013 to 2022 | Am J Transl Res (2022) | Web of Science Core Collection (WoSCC) | Visualization method, Time-slicing technique |  | Bibliometric visualization was conducted through VOSviewer and CiteSpace. | Cross validation |  |
| 10 | Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt | Springer (2023) | HCV dataset for Egyptian healthcare workers | Data mining with decision tree, Ensemble learning |  | NaïveBayes, Random forests, Logistic regression, KNN | Accuracy, precision, recall, and F1 scores | SFS |
| 11 | Classification Accuracy of Hepatitis C VirusInfection Outcome: Data Mining Approach | (IJCSEITR) (2013) | HCV dataset forEgyptian healthcareworkers(HCW) | Visualization method,Time-slicing technique |  | Random forests, NaïveBayes,Neural Network | AST | NA |
| 12 | RF-PSSM: A Combination of Rotation Forest Algorithm and Position-Specific Scoring Matrix for Improved Prediction of Protein-Protein Interactions Between Hepatitis C Virus and Human | BIG DATA MINING AND ANALYTICS(2023) | 477 HCV-human interactions dataset |  | NO | PSSM 2DPCA Rotation Forset Algorithm | Q,K | NA |
| 13 | Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm | IEEE Access (2022) | Liver disease datasets,IR dataset | SMOTE | YES | Artificial bee colony Synthetic minority oversampling technique Random forest Logistic regression | ACCURACY SCORE PRECISION SCORE RECALL SCORE F1 SCORE MCC SCORE | LR with feature selection |
| 14 | An Ensemble Learning Approach for Enhanced Classification of Patients With Hepatitis and CirrhosisAn Ensemble Learning Approach for Enhanced Classification of Patients With Hepatitis and Cirrhosis | IEEE Access (2021) | DISCOVERY COHORT VALIDATION COHORT | SMOTE | NO | Random Forests,Decision tree | ACCURACY, PRECISION,SENSITIVIY | UNIVARIATE STATISTICAL TESTS,,MACHINE LEARNING FEATURE SELECTION |
| 15 | Clinical Decision Support System for Liver Fibrosis Prediction in Hepatitis Patients: A Case Comparison of Two Soft Computing Techniques | IEEE Access (2018) | UCI dataset | ROSE | NO | Hill Climbing | Precision Miss rate Mean square error | Sequential feature selection |

# Data Preprocessing (if any): The data contained missing values in various columns and also had some character data in it. Mean imputation method has been deployed for imputing the missing values. The Category i.e the target variable was in characters which was then converted into factor by using as.factor method in R. The data is imbalanced for which oversampling is done using SMOTE function. The data was then split into Training and Testing data in 70:30 ratio. The training and testing data were again over-sampled to avoid any kind of imbalances in the data.

# Feature Selection (if any): There are no feature selection methods employed at first. Feature selection method was applied for finding the 5 most significant features which came out to be as ‘Sex’, ‘AST’, ‘ALT’, ‘GLT’, ‘Age’. Then the work was done using these 5 features and the accuracies were calculated for KNN, RF,

# Logistic Regression, SVM and NB. The set.seed function is an important function

# which ensures reproducibility. After applying set.seed method, AST, ALT, ALP, GGT, BIL were found to be the best 5 factors affecting the category variable.

# Algorithms Implemented: *Explain algorithms with default arguments values or set hyperparameters value.*

# 1] Naïve Bayes:

# Naive Bayes is a simple probabilistic classification algorithm based on Bayes' theorem. It assumes that the features used for classification are independent, even though this assumption is often not true in real-world data. It calculates the probability of an instance belonging to a specific class by multiplying the prior probability of the class and the conditional probabilities of each feature given that class. It then assigns the instance to the class with the highest probability. Naive Bayes is commonly used in text classification, spam detection, and other tasks with discrete data.

# Accuracy: 85.33%

# Precision: 83.50%

# Specificity: 81.67%

# Sensitivity: 88.83%

# 2] KNN:

# The k-Nearest Neighbors (KNN) algorithm is a simple machine learning classification and regression method. It works by finding the k data points in the training dataset that are closest to a new, unlabeled data point and assigns the label or value of the majority of those nearest neighbors to the new data point. KNN is based on the idea that similar data points tend to have similar outcomes. It's a non-parametric, instance-based learning algorithm, and the choice of the value k determines the number of neighbors considered for classification or regression.

# Accuracy: 87.5%

# Precision: 86.59%

# Specificity: 85.56%

# Sensitivity: 89.36%

# 3] Random Forest:

Random Forest is an ensemble machine learning algorithm that combine multiple decision trees to make more accurate predictions. It works by building a collection of decision trees during training and then averaging their predictions (for regression) or taking a majority vote (for classification) to make final predictions. The "random" part comes from using random subsets of the data and random subsets of features for each tree, which helps reduce overfitting and improve the model's

generalization capabilities. Random Forest is known for its versatility, robustness, and high performance across various machine learning tasks.

Accuracy: 88.59%

Precision: 82.88%

Specificity: 78.89%

Sensitivity: 97.87%

1. **SVM:**

SVM is a supervised machine learning algorithm used for classification and regression tasks. It finds a hyperplane that maximizes the margin between different classes in the data. SVM works by identifying support vectors, data points closest to the decision boundary. It can handle linear and nonlinear data separation using different kernel functions.

Accuracy: 86.41%

Precision: 82.55%

Specificity: 79.44%

Sensitivity: 93.09%

1. **Logistic Regression**:

Logistic Regression is a statistical method used for binary classification tasks. It models the relationship between a set of input features and the probability of an outcome belonging to one of two classes. By using a logistic function, it maps input data to a probability score between 0 and 1, making it a popular choice for tasks like spam detection and medical

diagnosis. It estimates the coefficients that best fit the data and can be used to make predictions by thresholding the probability score.

Accuracy: 89.40%

Precision: 86.70%

Specificity: 85.00%

Sensitivity: 93.62%

# CODE:

# #Imputation----

# library(imputeTS)

# d<-read.csv("CP.csv")

# print(d)

# m<-data\_imputed<-na\_mean(d)

# View(m)

# m$Category=as.numeric(m$Category)

# mn<-mean(d$CHOL,na.rm = TRUE)

# print(mn)

# install.packages("smotefamily")

# library(smotefamily)

# m$Sex=as.numeric(m$Sex)

# oversampled\_data <- SMOTE(Category ~ ., m, dup\_size = 1, K = 5)

# SMOTE(m, m$Category, K = 5, dup\_size = 1)

# #Code for feature selection:----

# 

# data\_rfe<-read.csv("CP\_imputed.csv")

# #subsets<-c(1:5,10,15,20)

# X<- data\_rfe[, !colnames(data\_rfe) %in% "Category"]

# Y<-data\_rfe$Category

# #set.seed(335)

# model<-train(X,Y,method="rf")

# ctrl <- rfeControl(functions = rfFuncs, method = "cv", number = 10)

# result <- rfe(X, Y, sizes = c(1:10), rfeControl = ctrl)

# print(result)

# selected\_features <- X[, result$optVariables]

# subset<-new\_dataset[,c("Category","Sex","AST","ALT","GGT","Age")]

# print(subset)

# #----

# library(caret)

# library(ROSE) # For SMOTE

# library(e1071) # For Naive Bayes

# # Convert the target variable to a factor with two levels

# subset$Category <- as.factor(subset$Category)

# levels(subset$Category) <- make.names(levels(subset$Category))

# # Split the data into 70% training and 30% testing

# set.seed(123)

# trainIndex <- createDataPartition(subset$Category, p = 0.7,

# list = FALSE,

# times = 1)

# data\_train <- subset[trainIndex,]

# data\_test <- subset[-trainIndex,]

# # Set up 10-fold cross-validation control with Accuracy as the evaluation metric

# ctrl <- trainControl(

# method = "cv", # Cross-validation

# number = 10, # Number of folds (10-fold)

# classProbs = TRUE, # To get class probabilities

# summaryFunction = twoClassSummary, # Evaluation metrics

# preProcOptions = list(thresh = 0.8), # Threshold for removing near-zero variance predictors

# selectionFunction = "best", # Select the best model based on Accuracy

# savePredictions = "final" # Save final predictions for the test set

# )

# # Apply SMOTE to balance the class distribution in the training data

# data\_train\_balanced <- ovun.sample(Category ~ ., data = data\_train, method = "both", N = 2 \* nrow(data\_train), seed = 42)$data

# # Apply SMOTE to balance the class distribution in the testing data

# data\_test\_balanced <- ovun.sample(Category ~ ., data = data\_test, method = "both", N =

# 2 \* nrow(data\_test), seed = 42)$data

# # Train the Naive Bayes model with cross-validation using balanced training data

# model <- train(

# Category ~ .,

# data = data\_train\_balanced, # Use the balanced training dataset

# method = "nb",

# trControl = ctrl

# )

# # Print the cross-validation results

# print(model)

# # Predict on the test data

# predictions <- predict(model, newdata = data\_test\_balanced)

# # Access confusion matrix and ROC values from the results

# confusion\_matrix <- confusionMatrix(predictions, data\_test\_balanced$Category)

# roc\_values <- model$results$ROC

# print("Confusion Matrix:")

# print(confusion\_matrix)

# print("ROC Values:")

# print(roc\_values)

# 

# Evaluation Parameters used:

# Accuracy

# Precision

# Sensitivity

# Precision

# Results and Discussions: *Quantitative findings and Plots.*

# 

# 

# 

# 

# 12.Conclusions:

# 1. When applied to the hepatitis dataset, Random Forest (RF) exhibited the highest predictive accuracy, making it a robust choice for classifying hepatitis cases.

# 2. Naive Bayes (NB) performed moderately well on this dataset, providing a simpler and faster alternative for classification tasks.

# 3. k-Nearest Neighbors (KNN) showed mixed results, highly dependent on the choice of k and the dataset's characteristics, making it less straightforward for hepatitis classification.

# 4. Support Vector Machine(SVM) :SVM's accuracy and other metrics will depend on the dataset, kernel choice, and hyperparameter tuning.

# 5. Logistic Regression(LR): Logistic regression's performance will vary based on the

# data's linearity and feature importance.Logistic regression's performance will vary based on the data's linearity and feature importance.

# 6. The choice between RF, NB, KNN, SVM and LR should consider the specific dataset and the importance of predictive accuracy, simplicity, and computational efficiency.

# 7. Further experimentation and performance evaluation are crucial to fine-tune the algorithm selection for the hepatitis dataset and optimize the classification results.

# References:

# [1] Butt, M. B., Alfayad, M., Saqib, S., Khan, M. A., Ahmad, M., Khan, M. A., &amp; Elmitwally, N. S. (2021, December 10). Diagnosing the stage of hepatitis C using machine learning. Journal of healthcare engineering. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8748759/>

# [2] Alizargar, A., Chang, Y.-L., &amp; Tan, T.-H. (2023, April 17). Performance comparison of machine learning approaches on hepatitis C prediction employing data mining techniques. MDPI. <https://www.mdpi.com/2306-5354/10/4/481>

# [3] N;, B. N. S. (n.d.). Prediction and staging of hepatic fibrosis in children with hepatitis C virus: A machine learning approach. Healthcare informatics research. <https://pubmed.ncbi.nlm.nih.gov/31406609/>

# [4] Lilhore, U. K., Manoharan, P., Sandhu, J. K., Simaiya, S., Dalal, S., Baqasah, A. M., Alsafyani, M., Alroobaea, R., Keshta, I., &amp; Raahemifar, K. (2023, August 1). Hybrid model for precise hepatitis-C classification using improved random forest and SVM method. Nature News. <https://www.nature.com/articles/s41598-023-36605-3>

# [5] Harabor, V., Mogos, R., Nechita, A., Adam, A.-M., Adam, G., Melinte-Popescu, A.-S., Melinte-Popescu, M., Stuparu-Cretu, M., Vasilache, I.-A., Mihalceanu, E., Carauleanu, A., Bivoleanu, A., &amp; Harabor, A. (2023, January 29). Machine learning approaches for the prediction of hepatitis B and C seropositivity. International journal of environmental research and public health. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9915359/>

# [6] Farghaly, H. M., Shams, M. Y., &amp; El-Hafeez, T. A. (2023, March 2). Hepatitis C virus prediction based on machine learning framework: A real-world case study in Egypt - knowledge and information systems. SpringerLink. <https://link.springer.com/article/10.1007/s10115-023-01851-4>

# [7] Tian, X., Chong, Y., Huang, Y., Guo, P., Li, M., Zhang, W., Du, Z., Li, X., &amp; Hao, Y. (2019, June 11). Using machine learning algorithms to predict hepatitis B surface

# antigen seroclearance. Computational and Mathematical Methods in Medicine. <https://www.hindawi.com/journals/cmmm/2019/6915850/>

# [8] Ali, A. M., Hassan, M. R., Aburub, F., Alauthman, M., Aldweesh, A., Al-Qerem, A., Jebreen, I., &amp; Nabot, A. (2023, March 16). Explainable machine learning approach for hepatitis C diagnosis using SFS feature selection. MDPI. <https://www.mdpi.com/2075-1702/11/3/391>

# [9] Ali, A. M., Hassan, M. R., Aburub, F., Alauthman, M., Aldweesh, A., Al-Qerem, A., Jebreen, I., &amp; Nabot, A. (2023, March 16). Explainable machine learning approach for hepatitis C diagnosis using SFS feature selection. MDPI. <https://www.mdpi.com/2075-1702/11/3/391>

# [10] Tang, X., Gan, P., Huang, S., Pan, X., Wang, R., Wang, P., Xia, H., Zeng, X., Ren, W., Shi, L., Zhou, X., &amp; Linghu, E. (2022, November 15). The scientific progress and prospects of Hepatitis C Research from 2013 to 2022. American journal of translational research. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9730086/>

# [11] Farghaly, H. M., Shams, M. Y., &amp; El-Hafeez, T. A. (2023a, March 2). Hepatitis C virus prediction based on machine learning framework: A real-world case study in Egypt - knowledge and information systems. SpringerLink. <https://link.springer.com/article/10.1007/s10115-023-01851-4>

# [12] Frias, M., Moyano, J. M., Rivero-Juarez, A., Luna, J. M., Camacho, Á., Fardoun, H. M., Machuca, I., Al-Twijri, M., Rivero, A., &amp; Ventura, S. (2021, February 24). Classification accuracy of hepatitis C virus infection outcome: Data Mining Approach. Journal of medical Internet research. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7946589/>

# [13] RF-PSSM: A combination of rotation forest algorithm and position ... (n.d.-i). <https://ieeexplore.ieee.org/document/9962955/>

# [14] Hepatitis C virus detection model by using Random Forest, logistic ... (n.d.-f). <https://ieeexplore.ieee.org/abstract/document/9868346/>

# [15] An ensemble learning approach for enhanced classification of patients ... (n.d.-c). <https://ieeexplore.ieee.org/document/9347427>

# [16] Clinical decision support system for liver fibrosis prediction in ... (n.d.-d). <https://www.researchgate.net/profile/Shaker-El-Sappagh/publication/327850558_Clinical_Decision_Support_System_for_Liver_Fibrosis_Prediction_in_Hepatitis_Patients_A_Case_Comparison_of_Two_Soft_Computing_Techniques/links/5bc6e947458515f7d9bfd8a1/Clinical-Decision-Support-System-for-Liver-Fibrosis-Prediction-in-Hepatitis-Patients-A-Case-Comparison-of-Two-Soft-Computing-Techniques.pdf>