Juxtaposition on Classifiers Modeling Hepatitis Diagnosis Data

ICCVBIC 2018

A Presentation by

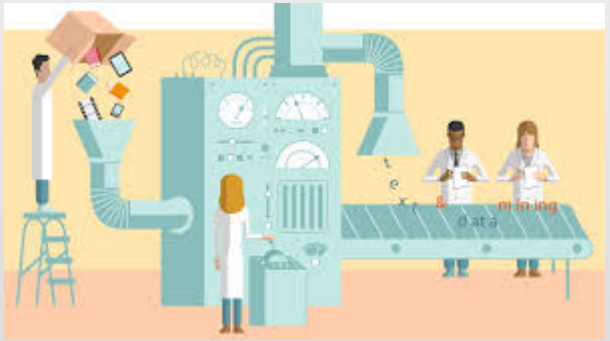
Preetham Ganesh	cb.en.u4cse15435@cb.students.amrita.edu Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
Harsha Vardhini Vasu	cb.en.u4cse15417@cb.students.amrita.edu Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
Keerthanna Govindarajan Santhakumar	cb.en.u4cse15420@cb.students.amrita.edu Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
Raakheshsubhash Arumuga Rajan	cb.en.u4cse15437@cb.students.amrita.edu Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
Bindu K R	j_bindu@cb.amrita.edu Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

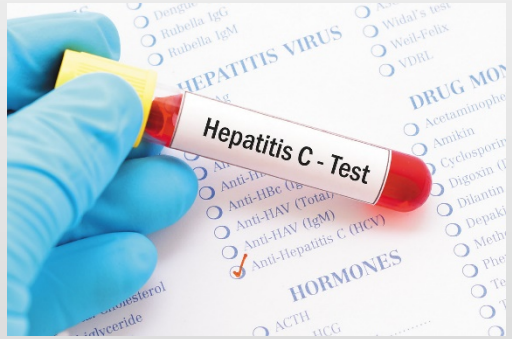
Content

- Introduction
- Motivation
- Related Works
- Methodology
  - Dataset Description
  - Process Flow
  - Classifier Models
  - Performance Measures
- Results and Discussion
- Conclusion

Introduction

- Machine Learning plays a crucial role in predicting unforeseeable parameters in different domains which has always been difficult for human prediction
- Data Mining plays a crucial role in mining the necessary features for prediction as the medical datasets has loads of information
- Hepatitis C, an acute or chronic disease that causes infection in the liver has approximately affected 130 – 170 million people in the world [1]





Motivation

- Hepatitis C is found worldwide
- Globally estimated 71 million people have chronic hepatitis C infection
- A serious number of those who chronically infected develop cirrhosis or liver cancer
- Antiviral medicines can cure more than 95% of people with the infection but access to diagnosis and treatment is low [2]
- Currently there is no vaccine for Hepatitis C

Motivation

- Hepatitis C can transmitted sexually and can be passed from an infected mother to her baby
- Estimated obtained from modeling suggest that worldwide, in 2015 there were 1.75 million new HCV infections [2]
- Globally 23.7 new HCV infections per 100,000 people [2]

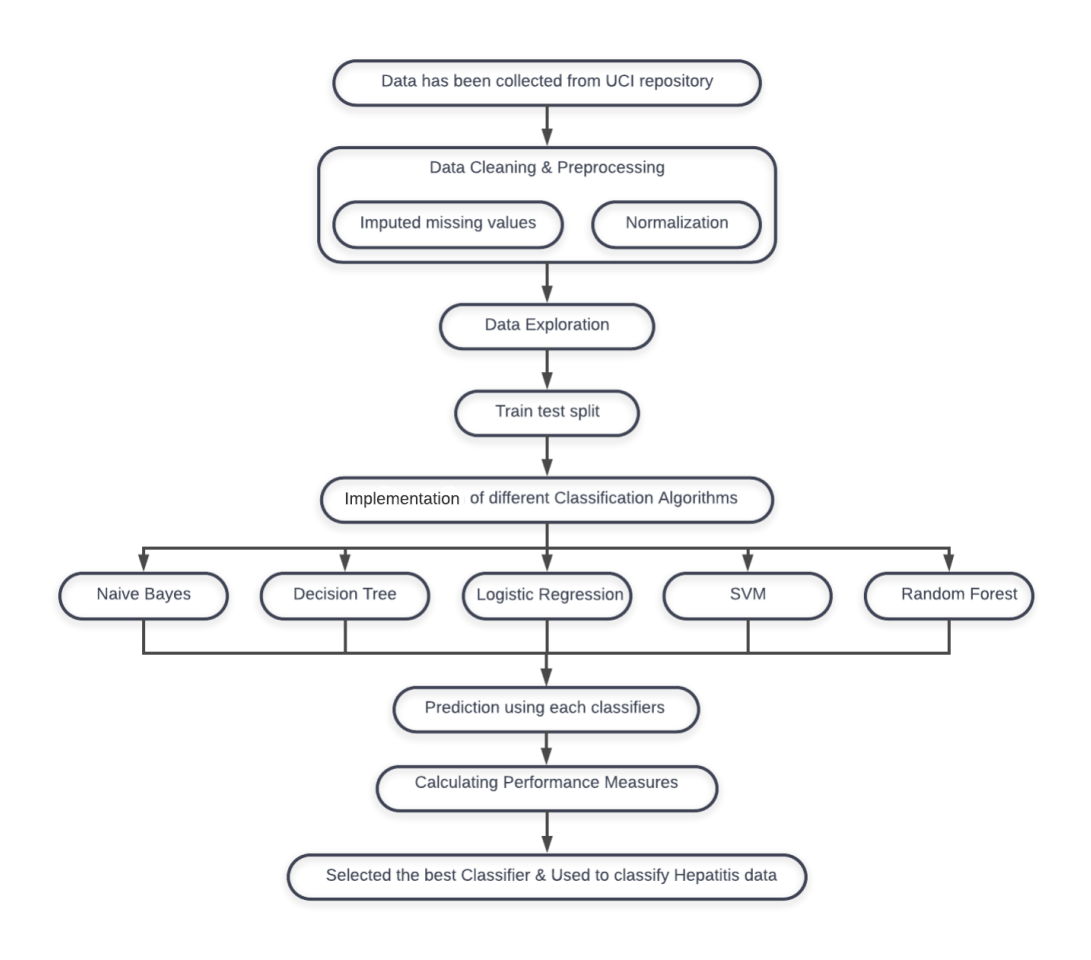
Related Works

Paper Name	Models Used	Performance Measures Used	Result
Ramaswamy et al [3]	Decision Stump, Random Forest, etc	Accuracy, Precision, Recall and F1-Measure	• Random Forest • Accuracy 87.5%
A. H. Roslina et al [4]	SVM (with & without feature selection)	Accuracy	• Feature selected • Accuracy 74.55%
S.Ekiz et al [5] (Heart Disease Dataset)	Decision Tree, SVM and Ensemble Learning in Weka and MATLAB	Accuracy	• Decision Tree (WEKA) • Accuracy 67.7%
K. Santhosh Bhargav et al [12]	SVM, Decision Tree, Logistic Regression, Naïve Bayes	Accuracy, Precision, Recall, and F-Measure	• Logistic Regression • Accuracy 87.17%

Methodology – Dataset Description

- Dataset is collected from UCI repository
- 155 tuples, 19 self-dependent attributes and label named “Class”
- **Numerical Attributes:**
  - Age, Bilirubin, Alk Phosphate, SGOT, Albumin & Protime
- **Categorical Attributes:**
  - Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices, Histology, **Class**

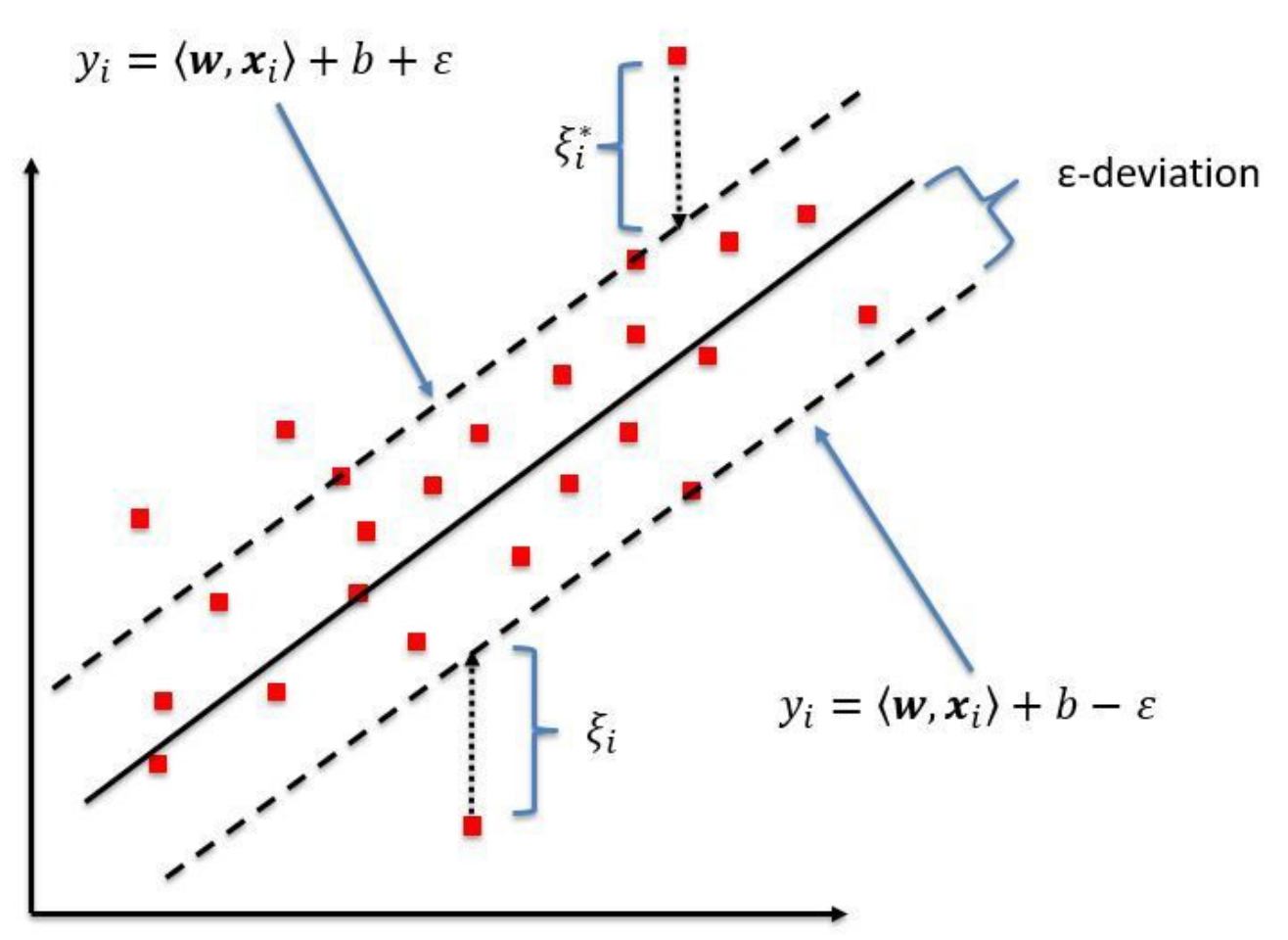
Methodology – Process Flow



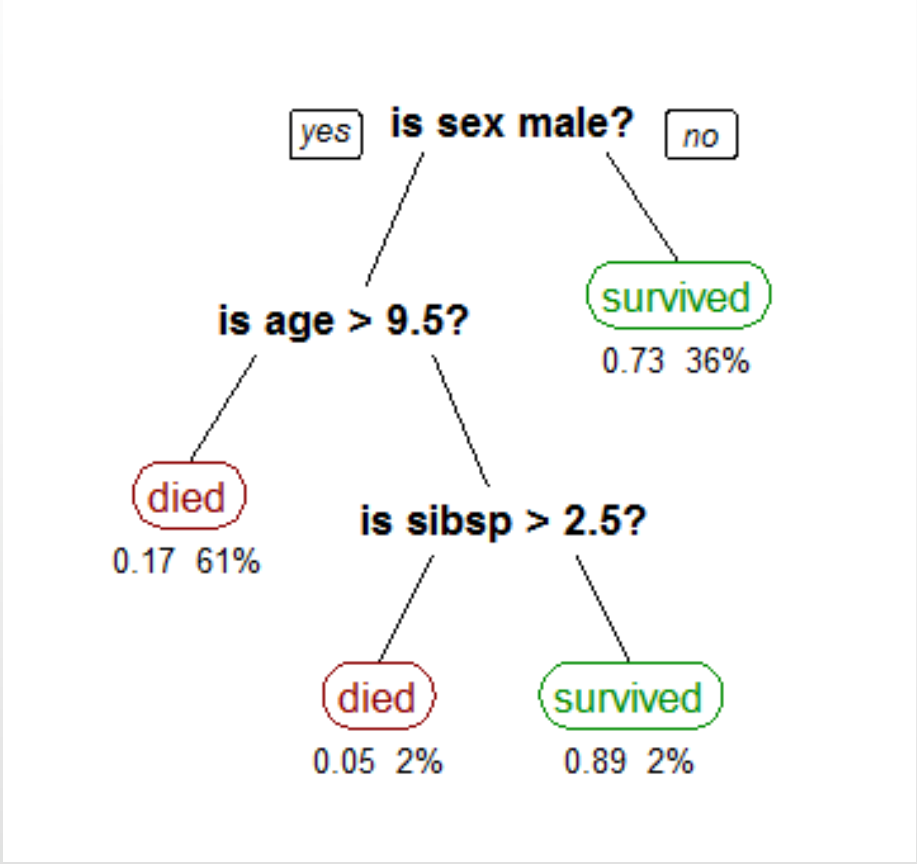
Methodology – Classifier Models

Support Vector Machine (SVM):	Naïve Bayes:	Decision Tree	Random Forest	Logistic Regression
• Classifies instances with high efficiency if they are in the form of vector [4] • Used to find optimal dividing hyperplane between classes [4]	• Based on implementation of Bayes Theorem [8] • Works on the assumption that attributes are independent of each other	• Structure similar to flowchart [9] • Consists of Decision Nodes, Chances Nodes and End Nodes [9]	• Ensemble learning method [10] • Builds swarm of decision trees during training • Overfit the model to training set [10]	• Logistic function used to model binary class variable • Label should be numerical (0/1) [11]

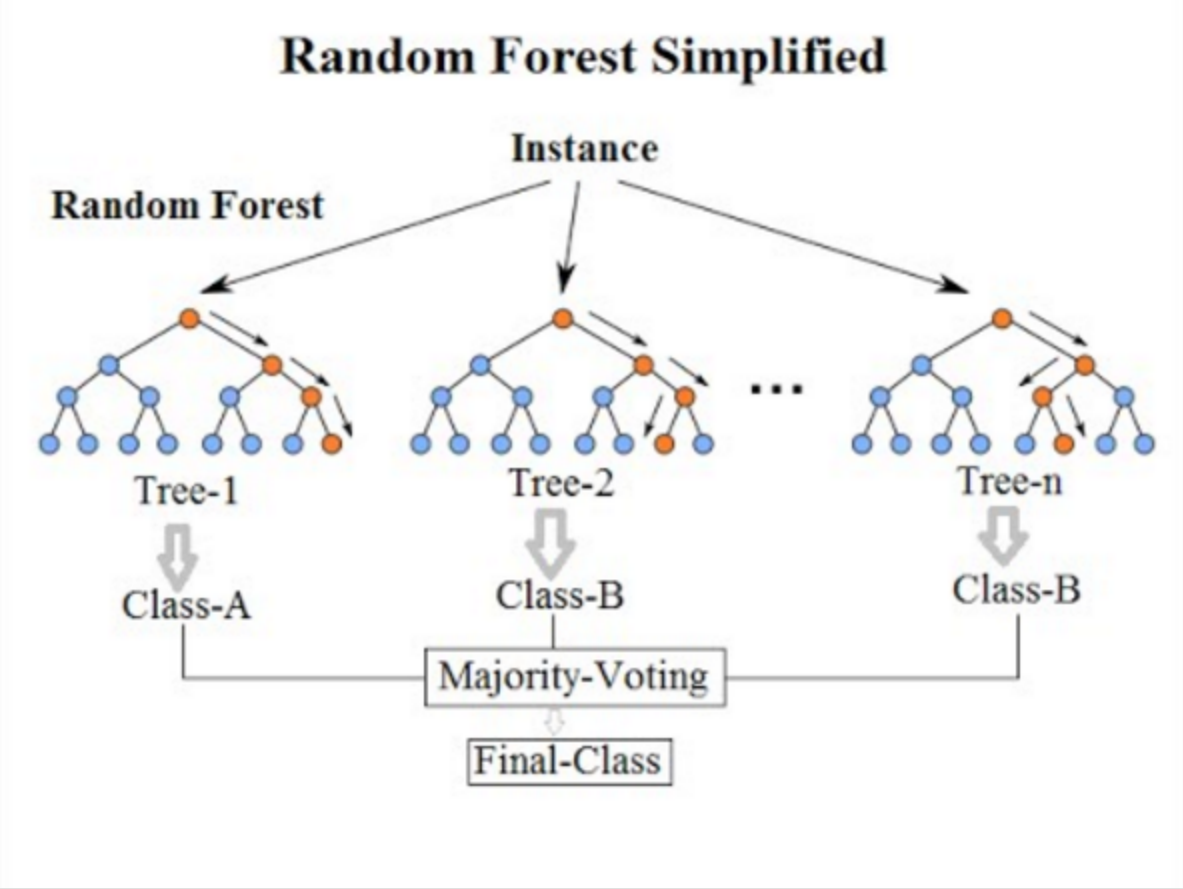
Methodology - SVM



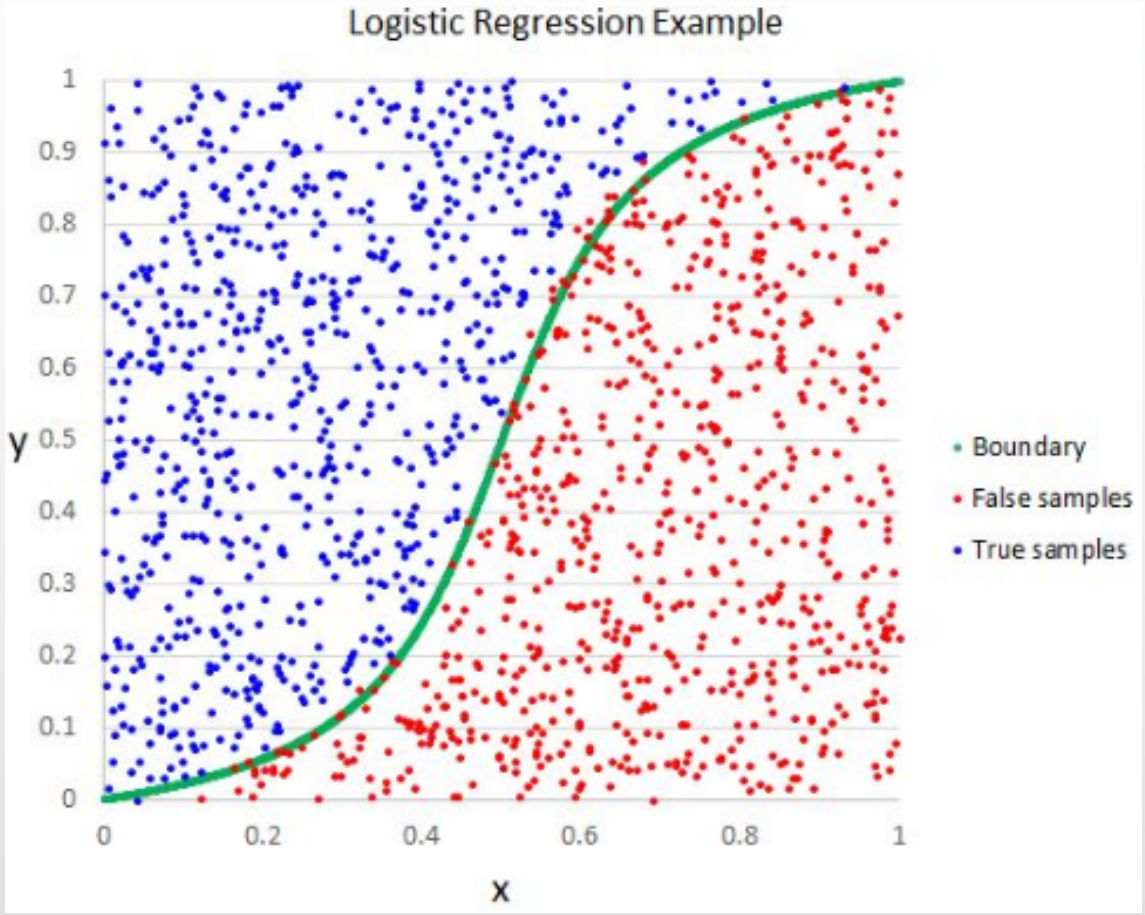
Methodology – Decision Tree



Methodology – Random Forest



Methodology – Logistic Regression



Methodology – Performance Measures

- Confusion Matrix / Contingency Table:

Confusion Matrix	Actual Class		
Predicted Class	Class	Class A	Class B
	Class A	True Positive (TP)	False Positive (FP)
	Class B	False Negative (FN)	True Negative (TN)

Methodology – Performance Measures

Accuracy	(TP + TN) / (TP + TN + FP + FN)
Balanced Accuracy	((TP / P) + (TN + N)) / 2
Recall (R) / Sensitivity (SN)	TP / (FP + TN)
Specificity (SP)	TN / (TN + FP)
Precision (P)	TP / (TP + FP)
Negative Predictive Value (NPV)	TN / (TN + FN) [6]
Fall-out	FP / (FP + TN)

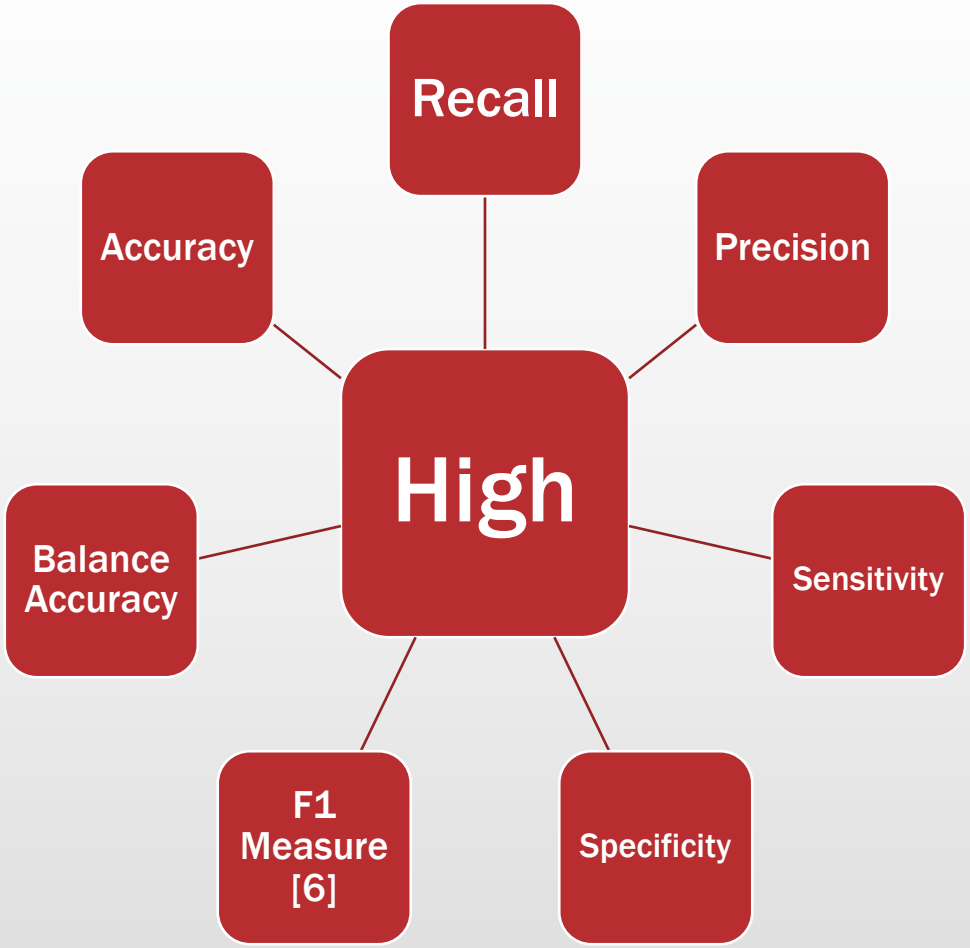
Methodology – Performance Measures

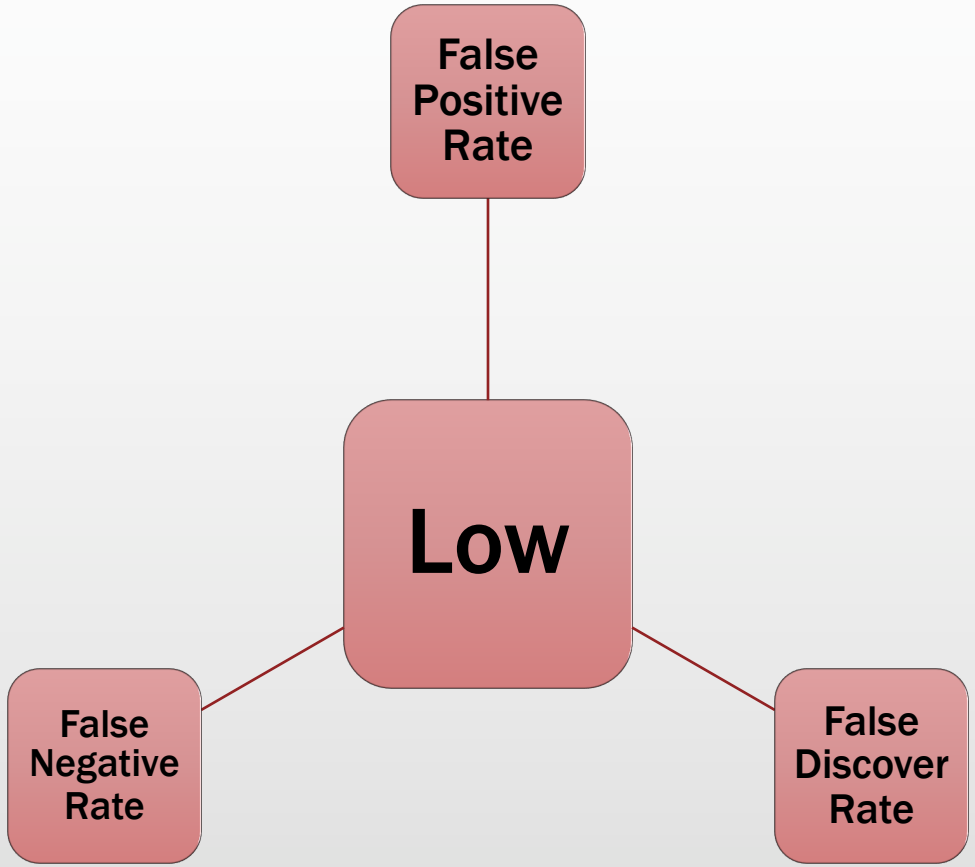
False Discovery Rate	FP / (FP + TP)
False Negative Rate	FN / (FN + TP)
F-Measure	2 * P * R / (P + R) [6]
Mathews Correlation Coefficient	((TP * TN) – (FP * FN)) / sqrt((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))
Informedness [7]	SP + SN - 1
Markedness	P + NPV - 1

Results & Discussion

- Missing values in the dataset has been imputed using Predictive Mean Matching
- Numerical attributes normalized using Z-score Normalization
- Holdout method used iteratively to split data into Train and Test where each iteration has different set of instances

Results & Discussion – Good Classifier





Results & Discussion

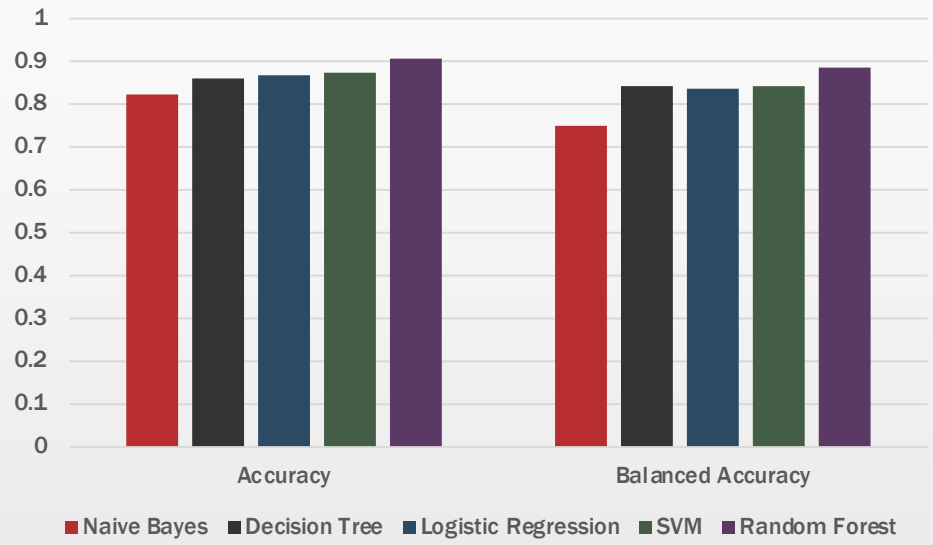
Performance Measure	Naïve Bayes	Decision Tree	Logistic Regression	SVM	Random Forest
Accuracy	0.823	0.86	0.867	0.873	**0.907**
Balanced Accuracy	0.749	0.843	0.836	0.842	**0.885**
Recall	0.536	0.808	0.702	0.788	**0.845**
Specificity	**0.962**	0.878	0.904	0.896	0.926
Precision	**0.867**	0.45	0.717	0.55	0.683
Negative Predictive Value [6]	0.813	**0.963**	0.904	0.954	**0.963**
Fall-out	**0.038**	0.122	0.096	0.104	0.074
False Discovery Rate [6]	**0.133**	0.55	0.283	0.45	0.317
False Negative Rate	0.464	0.192	0.298	0.212	**0.155**
F-Measure [6]	0.66	0.535	0.68	0.622	**0.734**

Results & Discussion

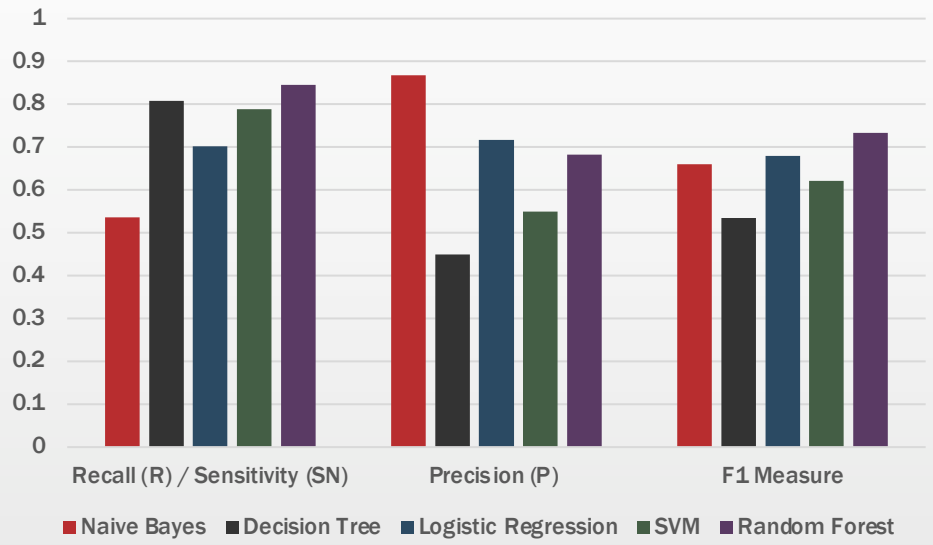
Performance Measure	Naïve Bayes	Decision Tree	Logistic Regression	SVM	Random Forest
Matthews Correlation Coefficient	0.116	**0.206**	0.158	0.189	0.189
Informedness [7]	0.499	0.687	0.606	0.684	**0.771**
Markedness	**0.679**	0.413	0.621	0.504	0.646

Results & Discussion

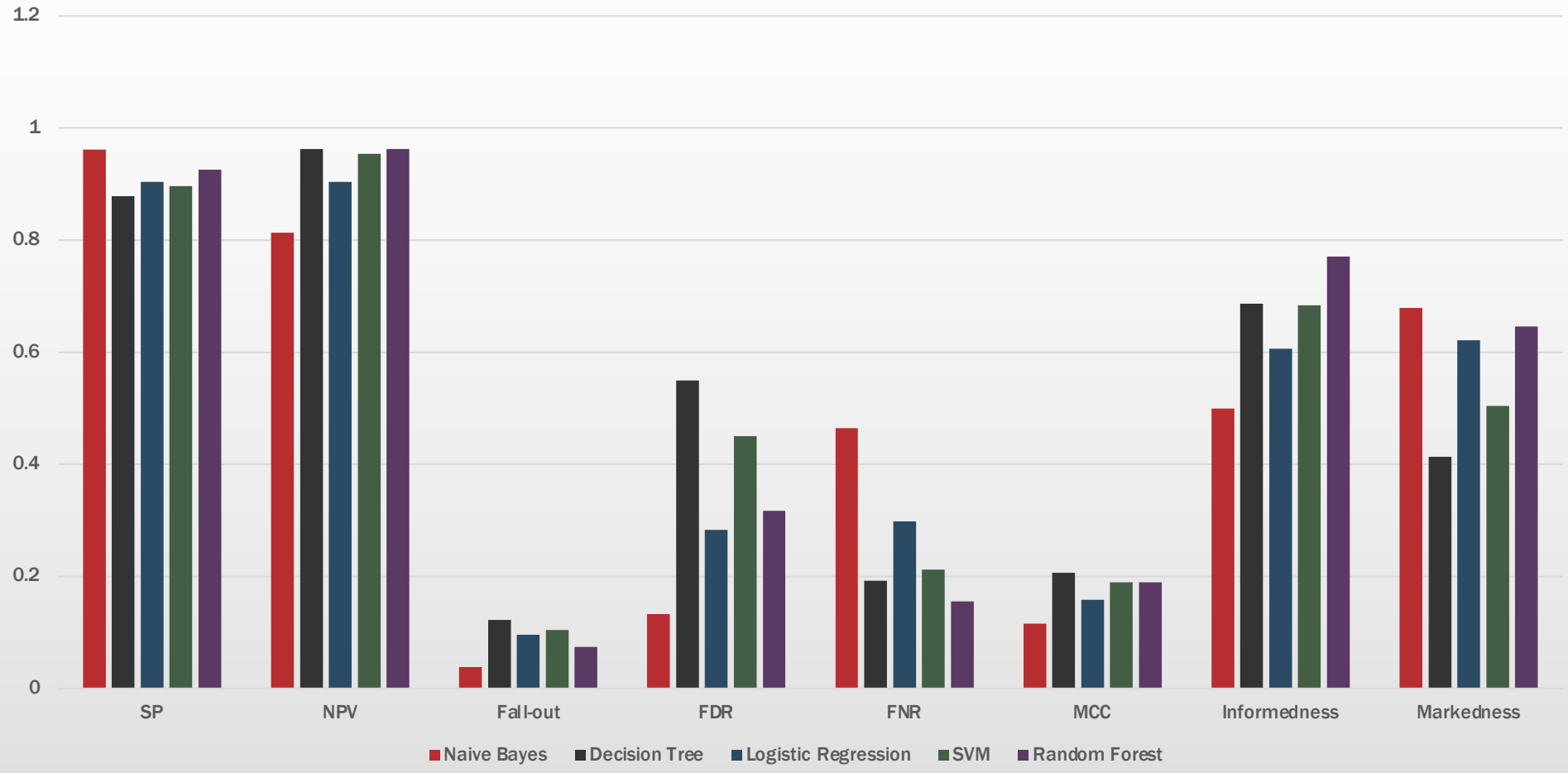
Accuracy & Balanced Accuracy



Recall, Precision & F1 Measure



Results & Discussion



Conclusion

- Inspected the performance of different classifiers modeled on the Hepatitis Data from UCI repository
- Random Forest works best with accuracy of 90.7%
- Also this model can be chosen because it worked well on the sparse data

References

1. Huda Yasin, Tahseen A Jilani and Madiha Danish. Article: Hepatitis-C Classification using Data Mining Techniques. International Journal of Computer Applications 24(3):1–6, June 2011.

2. http://www.who.int/news-room/fact-sheets/detail/hepatitis-c

3. M. Ramasamy, S. Selvaraj and M. Mayilvaganan, "An empirical analysis of decision tree algorithms: Modeling hepatitis data," 2015 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, 2015, pp. 1-4.

4. A. H. Roslina and A. Noraziah, "Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method," 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, 2010, pp. 2209-2211.

5. S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," 2017 Electric Electronics, Computer Science, Biomedical Engineering' Meeting (EBBT), Istanbul, 2017, pp. 1-4.

6. Wikipedia contributors. (2018, May 26). Evaluation of binary classifiers. In Wikipedia, The Free Encyclopedia

References

7. Powers, D.M.W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation. Technical Report SIE-07-001. School of Informatics and Engineering, Flinders University Adelaide, South Australia

8. Wikipedia contributors. (2018, September 23). Naive Bayes classifier. In Wikipedia, The Free Encyclopedia

9. Wikipedia contributors. (2018, September 20). Decision tree. In Wikipedia, The Free Encyclopedia

10. Wikipedia contributors. (2018, August 30). Random forest. In Wikipedia, The Free Encyclopedia.

11. Wikipedia contributors. (2018, October 3). Logistic regression. In Wikipedia, The Free Encyclopedia

12. Bhargav, K. Santosh, et al. "Application of Machine Learning Classification Algorithms on Hepatitis Dataset." *International Journal of Applied Engineering Research* 13.16 (2018): 12732-12737.