International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Artificial Intelligence as a Tool to Provide Agility in Health Care

Heena Rijhwani[1], Disha Motwani[2], Thackur Sahijwala[3]

[1-3]*Student, Thadomal Shahani Engineering College, University of Mumbai, Maharashtra, India*

Abstract - *Data Science holds great potential to transform healthcare and overcome its pitfalls and cases of ambiguity. In this paper, we are classifying whether a patient is affected by any of the five major diseases including Heart Conditions, Chronic Kidney Disease, Liver Disease, Breast Cancer and Diabetes by the use of machine learning models. We implemented six classification models including ensemble methods, logistic regression, support vector classifier and K nearest neighbor, and evaluated them based on confusion matrices, classification reports, receiver operating characteristics curves and overall accuracy. We used datasets collected from the UCI Machine learning repository for the purpose of this study.*

Key Words: *Adaboost Classifier, Decision Trees, Gradient Boosting Classifier, Logistic Regression, Naive Bayes, Random Forest, Support Vector Classifier*

1. INTRODUCTION

The WHO (World Health Organization) reported in 2019 that the top ten leading causes of death (accountable for 55% of the 55.4 million deaths worldwide) included Heart conditions (Ischaemic heart disease and stroke in particular), Cancer, Diabetes, Kidney diseases and Liver diseases (Cirrhosis of the liver). In a densely populated country like India, the number of patients is far more than the number of specialists. The 80,000 cardiology patients far outweigh the 4000 cardiologists in the country. Machine learning can be a reliable way of making health care more accessible, affordable and improving the reach of health care systems. By early detection of symptoms and diseases, it can help increase chances of survival. By preprocessing clinical data, selecting relevant features and applying Machine learning algorithms to automate the process of prediction of the five major diseases, we identify hidden patterns in data and make reliable clinical decisions with high accuracy.

2. LITERATURE REVIEW

Various surveys have been conducted and disease prediction models have been developed using data mining, machine learning, deep learning, and big data analytics techniques.

[1] surveyed various classification and association techniques to predict Chronic Kidney Disease. Ten best selection rules were chosen using the Apriori association algorithm and classification algorithms including ZeroR, OneR, IBk, J48, naive Bayes, and K nearest neighbor were compared in order to predict CKD.

[2[compared the performance of various machine learning algorithms on two datasets and achieved a highest accuracy of 98.8% using Adaboost classifier in order to predict if a patient is affected by diabetes.

[3] performed predictive analysis on the PIMA Diabetes dataset to predict the type of diabetes and related future risks using Hadoop and MapReduce techniques.

[4] used classification algorithms, Decision Tree and Naive Bayes to get insights from the heart disease dataset. Comparison was made for performance of both algorithms and on the basis of their accuracies.

[5] used support vector machine (SVM), K nearest neighbors, random forests, artificial neural networks (ANNs) and logistic regression to predict breast cancer using the Wisconsin Breast Cancer dataset and measured their performance with respect to accuracy, sensitivity, specificity, precision, negative predictive value, false-negative rate, false-positive rate, F1 score, Matthews Correlation Coefficient, precision–recall area under curve and receiver operating characteristic curve.

[6] performed feature selection and classification techniques and proposed an intelligent liver disease prediction software (ILDPS) based on software engineering model.

[7] compared algorithms like CHAID, Boosted c5.0, H-ANN and used ANN in order to gain actionable insights on a liver dataset.

[8] performed data visualizations to accurately represent comparisons between classification algorithms and used Logistic Regression, with an accuracy of 75%, for early prediction of liver disease.

[9] surveyed the challenges faced in finding patterns in electronic health records using natural language processing techniques and proposed a uniform and reproducible protocol for the same.

3. PROPOSED METHODS

3.1 Cancer

Breast Cancer is characterized by the uncontrolled growth of cells, which results in the formation of lumps within the breast. It is a treatable form of cancer but if not detected early, it can spread to other parts of the body and be life threatening. It can occur in both men and women, although male breast cancer is rare. Those who are more at risk include women above the age of 55, early menstruation or menopause, excessive consumption of alcohol, giving birth at an older age, hormone therapies, and genetics. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS) and invasive carcinoma. Others, like phyllodes

tumors and angiosarcoma are less common. Warning signs of breast cancer comprise of presence of lumps in the breast, swelling in areas near the breast or under the arms, pain in the breast, discharge from the nipples, other than breast milk, swelling under the arms, and change in the size or shape of nipples.

Data Description

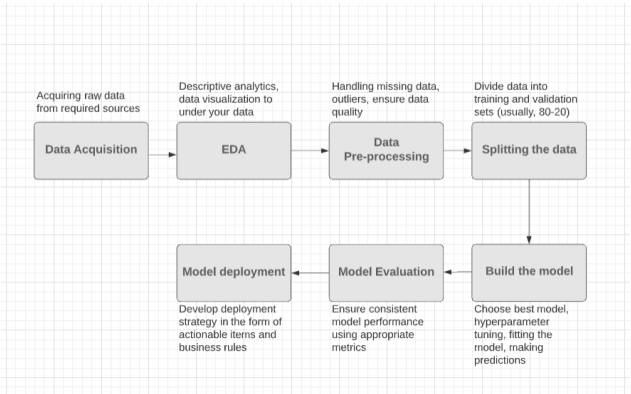
Total records: 569

The Wisconsin Breast Cancer data used in this study is taken from the UCI Machine Learning Repository, having 569 instances, 2 classes (malignant and benign), and 30 attributes for each cell nucleus including radius, texture, perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry and fractal dimension.

Class distribution: benign: 357 , malignant: 212.

For the purpose of the study, Jupyter notebook was used for implementation and Python programming language was used for coding.

Methodology



I.Data acquisition

II.Exploratory Data Analysis

III.Data Pre-processing

IV.Splitting the data

V.Building the model

VI.Evaluation of the model

VII.Model deployment

Data Pre-processing

We removed the unnecessary column of id, assigned numeric values to the class labels by replacing 'M'(malignant) and 'B'(Benign) in the diagnosis column to 1 and 0 respectively. We checked the dataset for outliers. We found missing values in four columns and handled them by replacing with the median of the particular column with respect to class label. We also performed feature scaling to normalize the range of the data and prevent incorrect predictions due to the tremendous difference between the magnitudes of the features. For this, we used the standard scalar library.

Standardization (also called z-score normalization) replaces values of the features with their z scores.

$$x' = (x - \bar{x}) / \sigma \quad (1)$$

where x is the data point, $\overline{X}$ is the mean of the entire column for a particular feature, and σ is the standard deviation. After applying this technique, distribution is converted in such a way that mean =0 and std deviation=1.

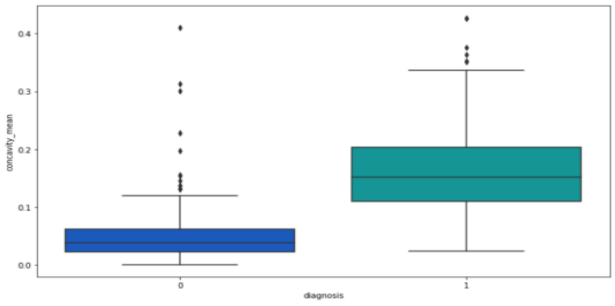


Fig. 1. In case of concavity mean, missing values were filled with 0.05 for class label 0 (Benign tumor) and 0.15 for class label 1(Malignant tumor).

Data Visualization

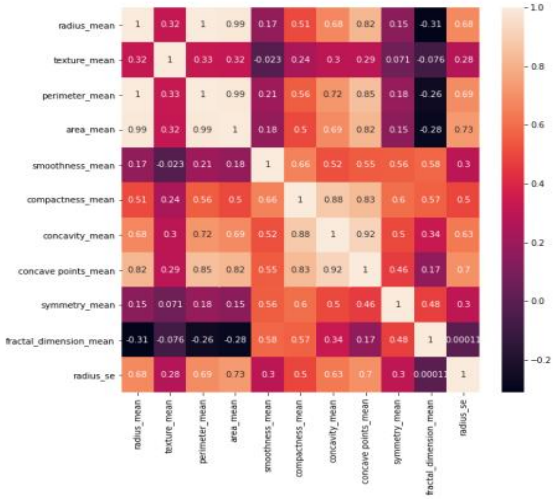


Fig. 2. Correlation between variables- A strong relation was found between area mean and radius mean, perimeter mean and area mean, concave points mean and concavity mean, and compactness mean and concavity mean.

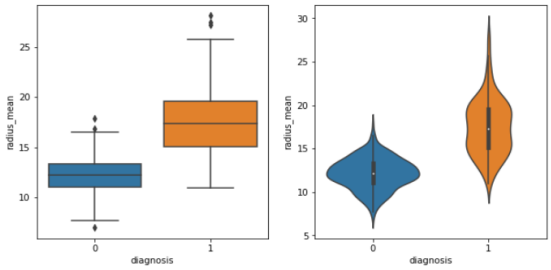


Fig. 3. Understanding correlation between radius and diagnosis of tumor- Malignant tumor can be associated with a higher radius mean.

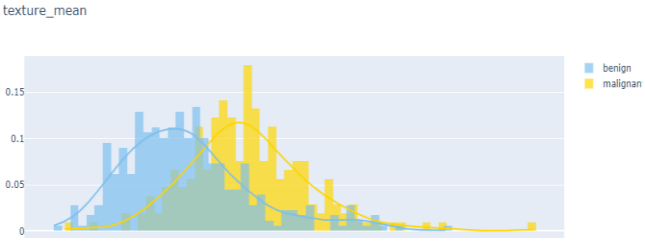


Fig. 4. Distplot to show histogram and density curve of texture mean for patients with benign and malignant tumors

Algorithms used

We compared the performance of various machine learning classification algorithms to diagnose a patient with breast cancer.

- Support Vector Machine
- Decision Tree Classifier
- K Nearest Neighbor
- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier
- Ada Boost Classifier

Initial comparison of Machine Learning Algorithms

Model	Accuracy_Score
DecisionTreeClassifier	0.92
GradientBoostingClassifier	0.93
KNeighborsClassifier	0.94
LogisticRegression	0.96
AdaBoostClassifier	0.97
SVC	0.973
RandomForestClassifier	0.973

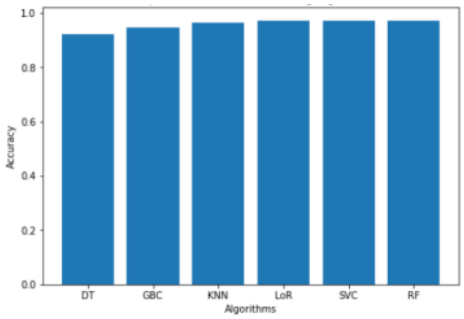


Fig. 5. Comparison of algorithms

We got the highest accuracy for Logistic Regression, Support Vector Classifier and Random Forest, thus we explored them further. For LoR and SVC, we applied Grid Search for choosing optimal parameters and Repeated Stratified K fold Cross Validation. We chose K value as 10. In 10-fold cross validation, our dataset will be divided into 10 blocks. Then in the first iteration, out of the 570 records in our dataset, 570/10= 57 will be test data and the model will be trained on the remaining 513 samples, giving us accuracy 1. In iteration 2, the next 57 samples will act as test data and the model will be trained on the remaining samples, giving us accuracy 2. A similar process will be followed for the next 8 iterations. The final accuracy of our machine learning model will be the mean of the 10 accuracies. Stratified cross validation was chosen since it makes sure that for each iteration, the number of instances of each class, for train and test data, is taken in an efficient manner.

Logistic Regression optimal parameters: {'C': 1, 'penalty': 'l2', 'solver': 'newton-cg'}

Support vector Classifier optimal parameters: {'C': 10, 'gamma': 0.01, 'kernel': 'rbf'}

For Random Forest we used Randomized Search for hyperparameter tuning.

Random Forest best parameters: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': 55, 'bootstrap': True}

Model Evaluation

Models were compared based on:

1.Accuracy- It is equal to the number of correct predictions / total predictions. It has a range from 0 to 1. The Accuracy measure merely tells how good or bad our model is but gives no information regarding what is wrong with it or where it is making errors. Hence we evaluated our models based on other metrics as well.

2.Confusion matrix is a table that gives insights into how many correct predictions our model has made. It correlates actual and predicted output values. True Positive(TP) values are those which are predicted correctly as positive. False Positive (FP) values are incorrectly predicted as positive. False Negative(FN) values should have been predicted as positive but were not. Lastly, True Negative(TN) values were correctly predicted as negative.

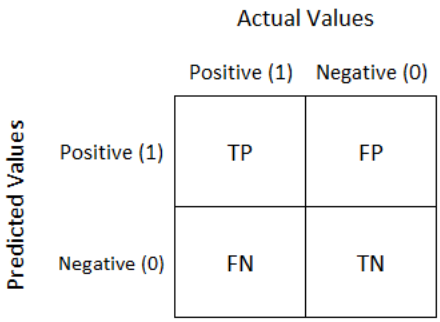


Fig. 6. Type 1 Error=FP Rate= FP/FP+TN, Type 2 Error= FN Rate=FN/FN+TP. In these terms we can say Accuracy= TN+TP / (TP+TN+FP+FN).

3.Precision or positive prediction value- It is the ratio of true positive to the total positives, i.e., ratio of patients who have been correctly identified with cancer to the total number of patients who have been detected with cancer.

Precision= TP/TP+FP

4.Recall or Sensitivity- It is the ratio of true positives to the sum of true positives and false negatives. In simple terms, it is the ratio of correctly predicted values to the total number of actual positive values.

Recall= TP/TP+FN

5.F Beta- It is the harmonic mean of precision and recall. If Beta value= 1, it is called F1 score.

$$F_\beta = \frac{(1+\beta^2)\cdot(\text{precision}\cdot\text{recall})}{(\beta^2\cdot\text{precision}+\text{recall})} \quad (2)$$

6.F1 Score-It takes into account true positive, false positive and false negative but not true negative. A good F1 Score indicates a good precision as well as recall and this way, we don't need to individually focus on any one of them.

F1 Score= 2 * (Precision*Recall / Precision+Recall)

=2*TP/2* TP+FP+FN

7.Matthews correlation coefficient- This uses all four categories of the confusion matrix.

MCC

=TP*TN-FP*FN/sqrt((TP+FP)(TP+FN)(TN+FP)(TN+FN))

If a person has cancer but has been predicted to not have it (false negative) would be the worst case scenario and could have disastrous ramifications. Thus we will focus on Recall as we want to reduce the number of false negatives.

Table- 1

	SVC	LoR	RF
Accuracy	98.25%	97.37%	97.37%
Confusion Matrix	[[70, 2], [0, 42]]	[[69, 0], [2, 41]]	[[70, 2], [1, 41]]
Precision	0.9832	0.9826	0.9739
Recall	0.9824	0.9821	0.9736
F Beta (F 2)	0.990	0.9624	0.971
MCC	0.963	0.9626	0.943

Since Support Vector Classifier gave us the highest accuracy as well as highest recall, we deployed it to classify whether a patient has Breast Cancer or not.

Mean squared error= 0.132

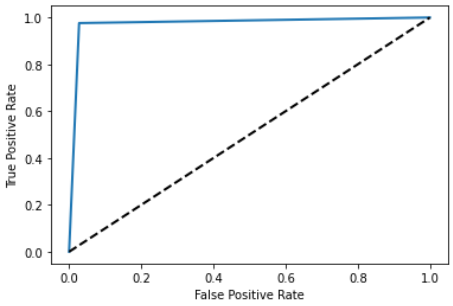


Fig. 7. ROC Curve

	precision	recall	f1-score	support
0	1.00	0.97	0.99	72
1	0.95	1.00	0.98	42
accuracy			0.98	114
macro avg	0.98	0.99	0.98	114
weighted avg	0.98	0.98	0.98	114

Fig. 8. Classification Report

3.2 Heart disease

Heart disease encompasses a wide range of cardiovascular problems. Several diseases and conditions fall under the umbrella of heart disease including Arrhythmia, Atherosclerosis, Cardiomyopathy, Congenital heart defects, Coronary artery disease (CAD), and Heart infections. Factors that increase the risk of heart diseases include High blood pressure, High cholesterol, Smoking, Diabetes, Dietary choices and High stress and anxiety. Symptoms include Chest pain, tightness, pressure and discomfort, Shortness of breath, Pain, Numbness, weakness or coldness in legs or arms and/or Pain in the neck, jaw, throat, upper abdomen or back.

Data Description

Total records:303

The UCI Machine learning repository dataset we have used has 2 classes (0 = no disease and 1 = heart disease), and 13 attributes.

Class distribution: heart disease: 165 , no disease: 138.

Sr. No.	Attribute	Details
1	Age	Patient's age in yrs
2	Sex	0:female, 1:male
3	Cp	4 types of chest pain (1: typical angina, 2: atypial angina,3: non-anginal pain, 4: asymptomatic)
4	Trestbps	Resting systolic blood pressure (in mm Hg)
5	Chol	Serum cholesterol (in mg/dl)
6	Fbs	Fasting blood sugar
7	Restecg	Rest electrocardiograph(0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy)
8	Thalch	Maximum heart rate achieved
9	Exang	Exercised induced angina (0: no, 1:yes)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (up sloping/ flat/ down sloping)
12	Ca	Number of major vessels (0-3)
13	Thal	Defect types (normal/ fixed defect/ reversible defect)

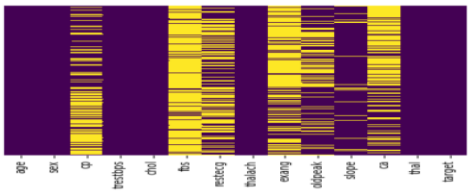


Fig. 9. Missingness map displaying missing values in columns

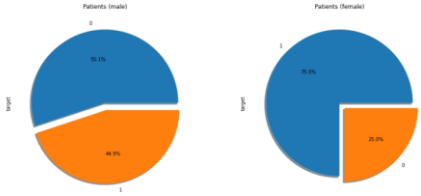


Fig. 10. A larger proportion of female patients were affected

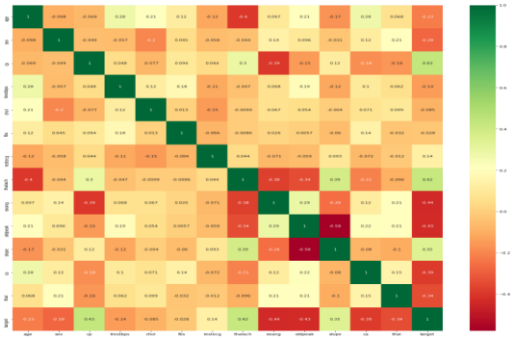


Fig. 11. Correlation matrix- Maximum heart rate achieved and slope of the peak exercise ST segment were seen to be strongly correlated with the target variable.

Data Pre-processing

Data types of certain columns such as sex, chest pain type, fasting blood sugar and rest ecg were changed from integer to categorical (object). For the categorical variables, one hot encoding was done by creating dummy variables. The first category of each was dropped to avoid the dummy variable trap since it can be derived using the rest (n-1) columns, where n is the number of classes in the particular categorical variable. For example, in case of type of chest pain (cp column) with 4 labels: 1—typical angina; 2—atypical angina; 3—non-anginal pain; 4—asymptomatic), we dropped typical angina. We checked for missing values and removed the columns of fbs, exang and ca since majority of their values were missing. For the rest of the columns with missing values (cp, restecg, oldpeak, slope, thal), we replaced them with the median of the respective column with respect to the target variable. We checked the dataset for outliers and performed feature scaling to normalize the range of the data and prevent incorrect predictions.

Algorithms used

We compared the performance of various machine learning classification algorithms to diagnose a patient with heart disease.

- Support Vector Machine
- Decision Tree Classifier
- K Nearest Neighbor
- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier

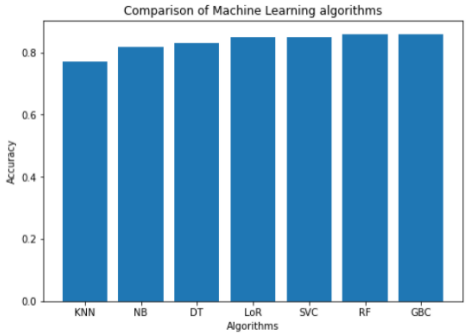


Fig. 12. Initial Comparison Of Machine Learning Algorithms

Model	Accuracy_Score
KNeighborsClassifier	0.77
NaiveBayes	0.81
DecisionTreeClassifier	0.83
SVC	0.85
LogisticRegression	0.84
RandomForestClassifier	0.85
GradientBoostingClassifier	0.85

We get the highest accuracy using Logistic Regression, Random Forest and Gradient Boosting Classifier, thus we explored them further. For LoR, we applied Grid Search for choosing optimal parameters and Repeated Stratified K fold Cross Validation for consistent results. For Random Forest and GBC we used Randomized Search for hyperparameter tuning.

Logistic Regression optimal parameters: {'C': 1, 'penalty': 'l2', 'solver': 'newton-cg'}

Random forest best parameters: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1,'max_features': 'auto', 'max_depth': 44, 'bootstrap': True}

Gradient Boosting Classifier optimal parameters: {'n_estimators':200,'min_samples_split':2,'min_samples_leaf': 1, 'max_features': 'auto','max_depth': 66}

Model Evaluation

Table -2

	GBC	LoR	RF
Accuracy	86.89	85.25%	85.25%
Confusion Matrix	[[25, 3], [5, 28]]	[[23, 5], [4, 29]]	[[23, 5], [4, 29]]
Precision	0.871	0.85244	0.85244
Recall	0.868	0.85245	0.85245
F Beta (F 2)	0.858	0.873	0.873
MCC	0.739	0.702	0.702

Since Gradient Boosting Classifier gave us the highest accuracy, as well as highest recall, we deployed it to classify whether a patient has Heart disease or not.

Mean squared error= 0.36214

	precision	recall	f1-score	support
0	0.83	0.89	0.86	28
1	0.90	0.85	0.88	33
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.87	0.87	0.87	61

Fig. 13. Classification report

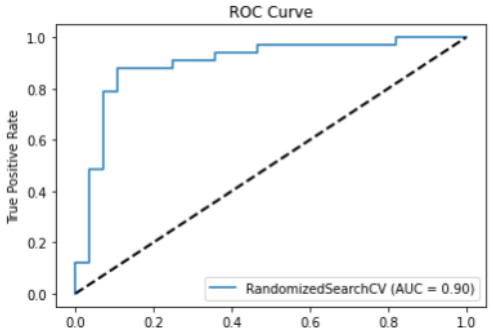


Fig. 14. ROC Curve

3.3 Diabetes

Diabetes mellitus is a chronic condition characterized by high levels of sugar (glucose) in the blood. It affects how the body controls and uses glucose, and people with diabetes have chronic high blood glucose levels (hyperglycaemia). Without ongoing careful management, diabetes can lead to a buildup of sugars in the blood, which can increase the risk of dangerous complications, including stroke and heart disease. There can be different types of diabetes, and managing the condition often depends on the type. Type 1 diabetes makes up for 10% of the cases and is a chronic autoimmune disease that develops when the immune system destroys the insulin-producing cells of the pancreas. People with type 1 diabetes require insulin replacement every day. Type 2 diabetes makes up for 80% of diagnosed cases and is usually caused by lifestyle factors. It occurs when the body becomes resistant to insulin and/or the body does not produce adequate insulin and is often preventable by keeping a healthy lifestyle. General symptoms of diabetes include Increased hunger and fatigue, Unplanned weight loss, Frequent urination and increased thirst, Dry mouth and itchy skin and Increased time for healing of wounds.

Data Description

Total records: 768

The dataset sourced from National Institute of Diabetes and Digestive and Kidney Diseases contains records of patients bifurcated based on Pregnancies ,Glucose, BloodPressure,

SkinThickness, Insulin, BMI, Diabetes Pedigree Function and Age.

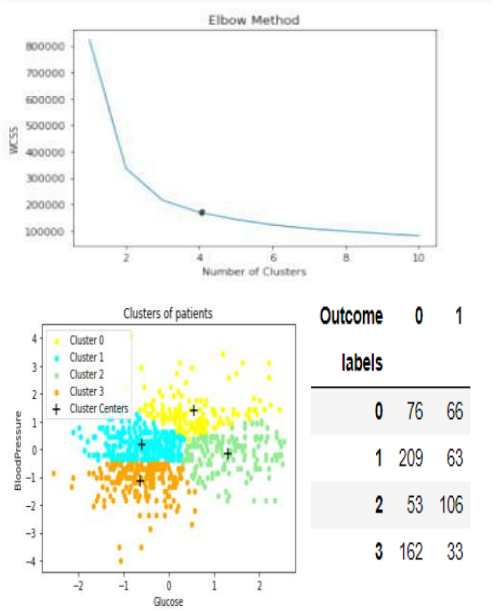
Class distribution: non diabetic: 500 , diabetic: 268.

Table - 3

Sr No.	Attribute
1	Pregnancies
2	Glucose
3	Blood pressure
4	Skin thickness
5	Insulin
6	BMI
7	Diabetes Pedigree Function
8	Age

Clustering

We performed Kmeans clustering on the data based on attributes Glucose and Blood Pressure. We used the Within-Cluster-Sum-of-Squares (WCSS) method and found optimum value for K using an Elbow point graph. After implementation of clustering, we looked at the distribution of people with and without diabetes with respect to each cluster.



Outcome	0	1
labels		
0	76	66
1	209	63
2	53	106
3	162	33

Fig. 15, 16 and 17. High proportion of diabetic patients was observed in Cluster 2, having higher glucose levels and relatively high blood pressure.

Algorithm:

1.Specify number of clusters K.

2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters is not changing.

Data Preprocessing

We checked for missing values. These were replaced with the median of the respective column based on its corresponding Outcome value.

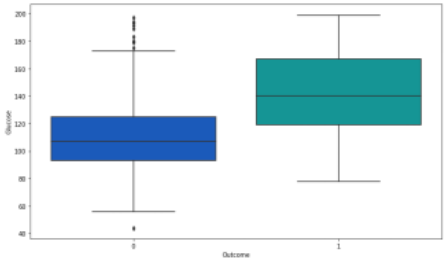


Fig. 18. This box plot has Outcome on the x-axis, depicting whether a patient has or does not have diabetes (0 or 1) and Glucose levels on the y-axis. It shows the median Glucose level for patients in both these classes. For records with missing values, we replaced them with the median of the corresponding class they belonged to instead of taking a more arbitrary approach and replacing null values with the median of the entire column. No outliers were found in the dataset. We also performed feature scaling to normalize the range of the data.

Data Visualization

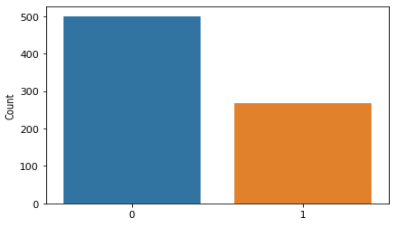


Fig. 19. The bar graph gives us the representation of presence of diabetes in a patient wherein “0” marks the absence of Diabetes and “1” marks the presence of it. More number of patients were detected with diabetes.

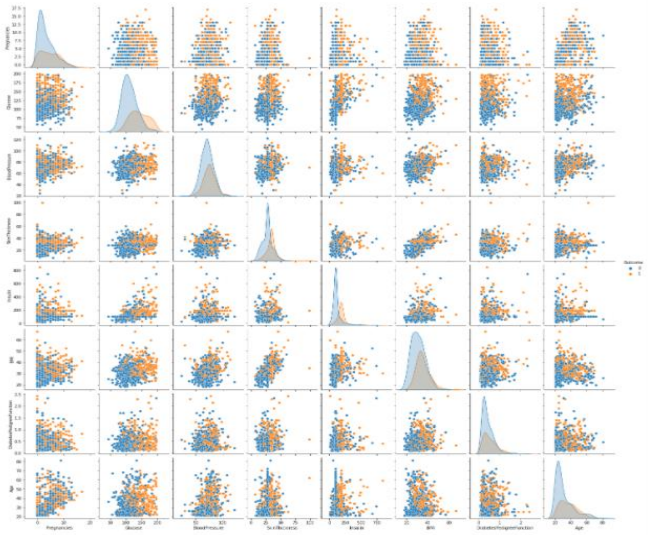


Fig. 20. The diagonal shows the distribution of the the dataset with the kernel density plots. The scatter-plots shows the relation between each and every attribute or features taken pairwise.

Algorithms used

- Support Vector Machine
- Decision Tree Classifier
- K Nearest Neighbor
- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier
- Ada Boost Classifier

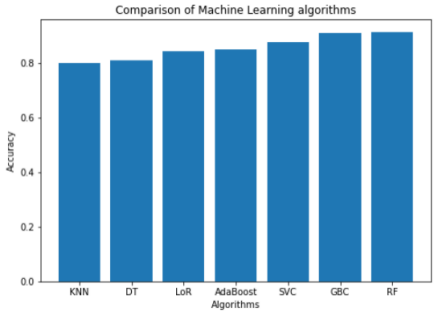


Fig. 21. Initial Comparison Of Machine Learning Algorithms

Model	Accuracy_Score
KNeighborsClassifier	0.83
DecisionTreeClassifier	0.83
LogisticRegression	0.85
AdaBoostClassifier	0.852
SVC	0.86
GradientBoostingClassifier	0.896
RandomForestClassifier	0.889

We performed Grid Search and Repeated Stratified K fold Cross Validation for SVC, and Randomized Search for RF and GBC.

Random Forest best parameters:{'n_estimators': 2000, 'min_samples_split': 2,'min_samples_leaf': 2,'max_features': 'sqrt','max_depth': 33,'bootstrap': True}

Logistic Regression best parameters:

{'C': 1, 'penalty': 'l2', 'solver': 'newton-cg'}

SVC best parameters:

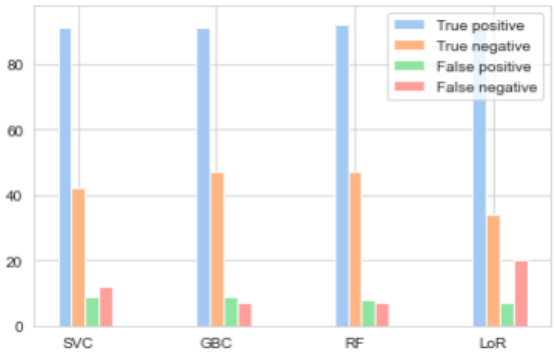
{'C': 1000, 'gamma': 0.1, 'kernel': 'linear'}

GBC best parameters: {'n_estimators': 200, 'min_samples_split':10, min_samples_leaf': 2, 'max_features': 'sqrt','max_depth': 55}

Model Evaluation

Table 4

	SVC	GBC	RF
Accuracy	86.36%	89.61%	90.259%
Confusion Matrix	[[91, 9], [12, 42]]	[[91, 9], [7, 47]]	[[92, 8], [7, 47]]
Precision	0.86	0.897	0.903
Recall	0.86	0.896	0.9025
F Beta (F 2)	0.786	0.8639	0.867
MCC	0.697	0.774	0.787



TP	91	91	92	90
FP	9	9	8	6
FN	12	7	7	20

TN	42	47	47	36

Fig. 22. illustrates the combined confusion matrix which depicts that the LoR model gives highest number of false negatives, which can result in a non-diabetic patient not being detected with the disease, whereas RF has the least FN.

Since Random Forest Classifier gave us the highest accuracy as well as highest recall, we deployed it to classify whether a patient has Heart disease or not.

Mean squared error= 0.31209

	precision	recall	f1-score	support
0	0.93	0.92	0.92	100
1	0.85	0.87	0.86	54
accuracy			0.90	154
macro avg	0.89	0.90	0.89	154
weighted avg	0.90	0.90	0.90	154

Fig. 23. Classification report

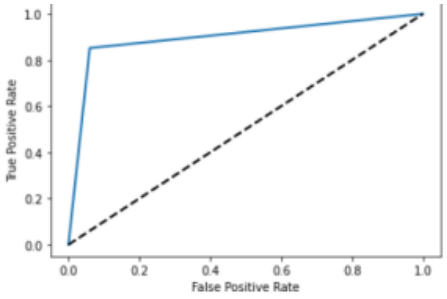


Fig. 24. ROC Curve

3.4 Liver disease

The liver plays an important role in many bodily functions ranging from protein production and blood clotting to cholesterol, glucose (sugar), and iron metabolism. Liver disease is a general term that refers to any condition affecting your liver. These conditions may develop for different reasons, but they can all damage your liver and impact its function. Liver disease refers to any condition that causes inflammation or damage to your liver. Liver disease can affect the overall function of your liver. Liver failure is when your liver has lost some or all of its functionality. It can occur due to the damage that's caused by liver disease. Symptoms include Yellow skin and eyes, known as jaundice, Swollen ankles, legs, or abdomen, Urine develops a dark yellow colouration, Stools look pale, bloody or black (tar-like), Easy bruising, Nausea or vomiting and Decreased appetite.

Data Description

Total records: 583

The dataset contains records collected from North East of Andhra Pradesh, India and has been sourced from the UCI Machine Learning repository.

Class distribution: Liver disease: 416 , No liver disease: 167.

Table- 5

Sr No.	Attribute
1	Age of the patient
2	Gender of the patient
3	Total Bilirubin
4	Direct Bilirubin
5	Alkaline Phosphotase
6	Alamine Aminotransferase
7	Aspartate Aminotransferase
8	Total Protiens
9	Albumin
10	Albumin and Globulin Ratio

Data Pre-processing

We use one hot encoding on gender using dummy variables. The first category was dropped, leaving us with Gender_male (0: female, 1:male). Null values found in the Albumin_and_Globulin_Ratio column were replaced by median 0.8 for patients not affected by liver disease and 1.0 for patients affected by liver disease. An imbalance was found in the data.

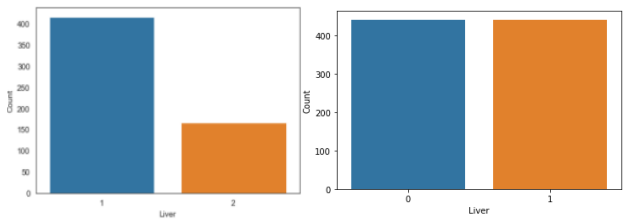


Fig. 25. Before and after oversampling

We performed random oversampling to overcome this problem. It is achieved by keeping the number of majority class samples the same and working on increasing frequency of samples belonging to the minority class by duplicating the samples at random. Value counts for Outcome column after random oversampling:

2 416

1 416

Data Visualization

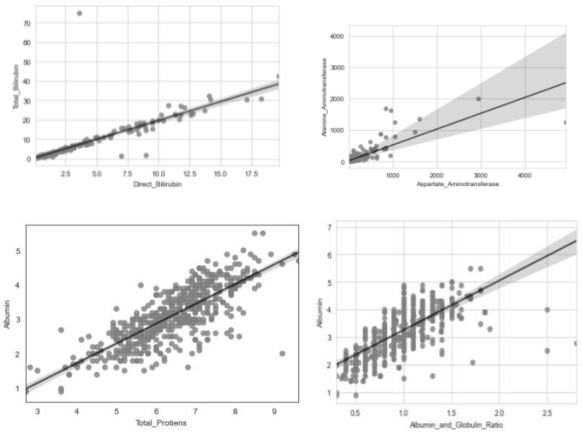


Fig. 27. Direct relationship found between Direct_Bilirubin and Total_Bilirubin, Aspartate_Aminotransferase and Alamine_Aminotransferase, Total_Protiens and Albumin, and Albumin_and_Globulin_Ratio and Albumin.

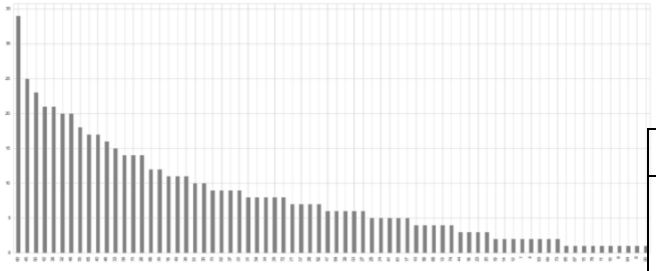


Fig. 28. Greatest number of patients were 60 years old.

Algorithms used

- Support Vector Machine
- Decision Tree Classifier
- K Nearest Neighbor
- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier

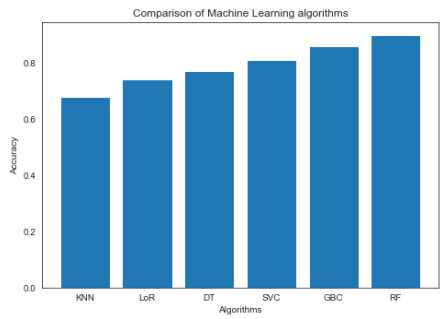


Fig. 29. Initial comparison of Machine Learning Algorithms

Model	Accuracy_Score
KNeighborsClassifier	0.68
LogisticRegression	0.74
DecisionTreeClassifier	0.75
SVC	0.80
GradientBoostingClassifier	0.84
RandomForestClassifier	0.91

We performed Grid Search and Repeated Stratified K fold Cross Validation for SVC, and Randomized Search for Random Forest and Gradient Boosting Classifier. Random Forest Optimal parameters: {'n_estimators': 200, 'min_samples_split': 7, 'min_samples_leaf': 3, 'max_features': 'auto', 'max_depth': 110, 'bootstrap': False}

Support Vector Classifier Optimal parameters:{'C': 100, 'gamma': 1, 'kernel': 'rbf'}

Gradient Boosting Classifier Optimal parameters:

{'n_estimators': 800, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'sqrt','max_depth': 88}

Model Evaluation

Table -6

	GBC	SVC	RF
Accuracy	87.01%	84.18%	93.22%
Confusion Matrix	[[82, 6], [17, 72]]	[[77, 11], [17, 72]]	[[82, 6], [6, 83]]
Precision	0.8579	0.8434	0.9322
Recall	0.8700	0.8418	0.9322
F Beta (F 2)	0.8294	0.8200	0.9325
MCC	0.7460	0.6853	0.8624

Since Random Forest gave us the highest accuracy, recall and lowest false negatives, we deployed it to classify whether a patient has liver disease or not.

	precision	recall	f1-score	support
0	0.93	0.93	0.93	88
1	0.93	0.93	0.93	89
accuracy			0.93	177
macro avg	0.93	0.93	0.93	177
weighted avg	0.93	0.93	0.93	177

Fig. 29. Classification report

Mean squared error: 0.2603

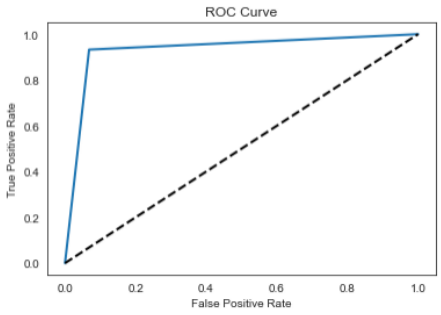


Fig. 30. ROC Curve

3.5 Chronic kidney disease

Kidney failure occurs when one's kidneys lose the ability to sufficiently filter waste from your blood. Many factors can interfere with your kidney health and function, such as Toxic exposure to environmental pollutants or certain medications, Certain acute and chronic diseases, Severe dehydration and/or Kidney trauma. Symptoms include Swelling of legs, ankles, and feet from retention of fluids caused by the failure of the kidneys to eliminate water waste, Unexplained shortness of breath, A reduced amount of urine, Blood during urination, Fatigue, Erectile dysfunction in men, Unexplained rise in blood pressure, High blood potassium (hyperkalemia) and pain or pressure in the chest.

Data Description

The dataset sourced the UCI Machine Learning repository contains records of patients over a period of 2 months.

Total records: 400

Table -7

Sr No.	Attribute	Description
1	age	Age of the patient
2	id	Patient id
3	bp	blood pressure
4	sg	specific gravity
5	al	albumin
6	su	sugar
7	rbc	red blood cells
8	pc	Pus cells
9	pcc	Pus cell clumps
10	ba	bacteria
11	bgr	blood glucose random
12	bu	Blood urea
13	sc	Serum creatinine
14	sod	sodium
15	pot	potassium
16	hemo	hemoglobin
17	pcv	Packed cell volume
18	wc	White blood cell count
19	rc	Red blood cell count
20	htn	hypertension
21	dm	diabetes mellitus
22	cad	coronary artery disease
23	appet	appetite
24	pe	pedal edema
25	ane	anemia
26	classification	0: no ckd, 1: ckd

Class distribution: Chronic kidney disease: 250 , No chronic kidney disease: 150.

Data Pre-processing

Unique values in the classification column were found to be ['ckd', 'ckd\t', 'notckd']. These were to mapped to numeric variables ({'ckd':1.0,'ckd\t':1.0,'notckd':0.0,'no':0.0}). We replaced values such as 'yes','present','good','normal' with 1 and 'no','notpresent','poor','abnormal' with 0.

```
row_rep = {'yes':1,'no':0,
           'normal':1,'abnormal':0,
           'present':1,'notpresent':0,
           'good':1,'poor':0,
           '\tno':0,'\tyes':1,
           'yes':1}
df1 = df.replace(row_rep)
```

The unique identifier (id) column was removed and missing values were checked for. The 'rbc' and 'rc' had a vast umber of missing records and were hence dropped. The rest of the missing values were deal with by removing the particular records, which left us with lesser number of records but more reliable data.

Data Visualizations

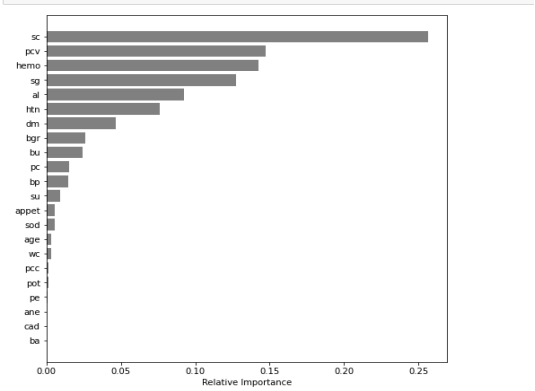


Fig. 31. Serum creatinine, packed cell volume and hemoglobin were seen to be the most relevant features.

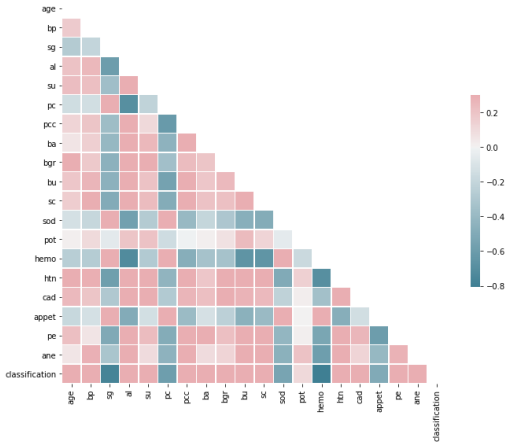


Fig. 32. Lower Triangle Heatmap shows negative relation between hemoglobin and Albumin, blood urea, hypertension and serum creatinine.

Algorithms used

- Support Vector Machine
- Decision Tree Classifier
- K Nearest Neighbor
- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier

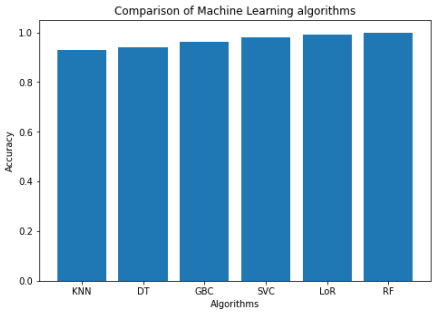


Fig. 33. Initial comparison of Machine Learning Algorithms

Model	Accuracy_Score
KNeighborsClassifier	0.93
DecisionTreeClassifier	0.94
GradientBoostingClassifier	0.96
SVC	0.98
LogisticRegression	1.00
RandomForestClassifier	1.00

We used Randomized Search and found the best parameters for Random Forest as: {'n_estimators': 800,

'min_samples_split': 3,

'min_samples_leaf': 2,

'max_features': 'sqrt',

'max_depth': 55,

'bootstrap': False}

Model evaluation

Accuracy was seen to be 100% and values for precision, recall, F2 score, matthews correlation coefficient were 1.0.

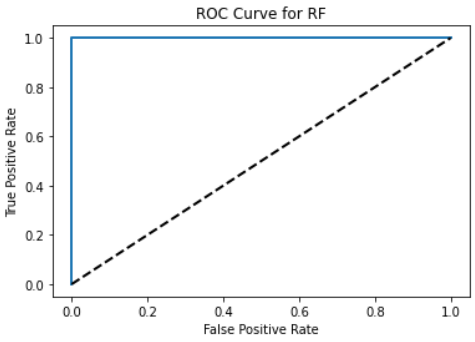


Fig. 34. ROC curve (receiver operating characteristic curve)

4. Conclusion

For each of the major diseases, we researched the performance of various classifiers and deployed the one with the lowest number of false negatives and highest recall, after finding the optimal parameters. Further, a Machine Learning web app was created using Flask and deployed using Heroku in order to make these predictions gainful to every user. Additionally, this work can be scaled to include detection and diagnosis of a myriad of other diseases (including Covid-19) in order to supplement the clinical decision making process.

5. Acknowledgement

We thank Professor Mukesh Israni (Thadomal Shahani Engineering College, University of Mumbai) for his guidance during the course of this research.

6. References

[1]www.sciencepubco.com/index.php/IJET

[2]Aishwarya Mujumdar, V Vaidehi, "Diabetes Prediction using Machine Learning Algorithms",Procedia Computer Science,Volume 165,2019

[3]Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar, "Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017

[4]S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019

[5]Islam, M.M., Haque, M.R., Iqbal, H. et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN COMPUT. SCI. 1, 290 (2020)

[6]Jagdeep Singh, Sachin Bagga, Ranjodh Kaur,"Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques",Procedia Computer Science,Volume 167,2020

[7]Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D. Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, Wendy Chapman, Rina Dutta,"Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances",Journal of Biomedical Informatics,Volume 88,2018.