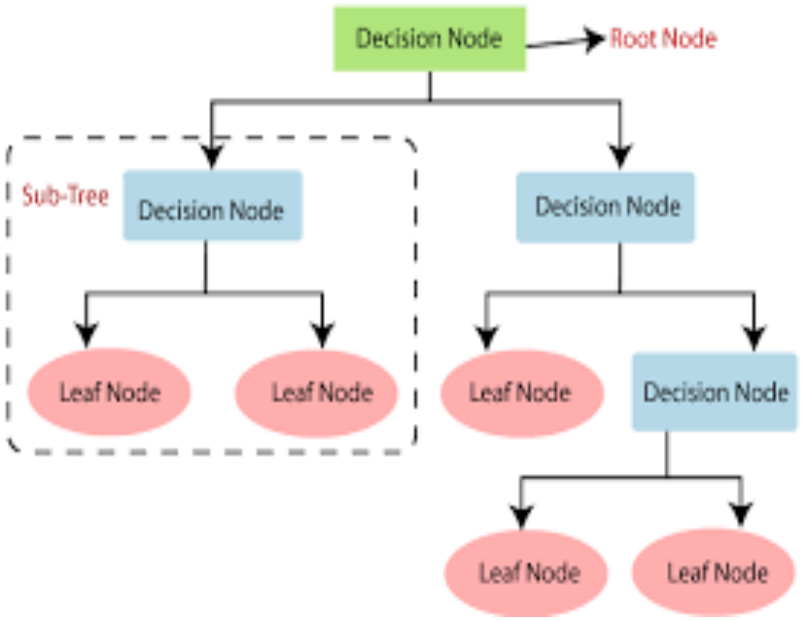
DAYANANDA SAGAR COLLEGE OF ENGINEERING
COMPUTER SCIENCE & ENGINEERING

Minor Project- Report
Aug-2021-2022

Course Faculty: Dr. Vinothini C
Course Name & Code: Cloud & Big Data Laboratory with Minor Project & 18CS7DLCBL
Semester: 7
Date: 19-1-2022

TITLE OF THE PROJECT	Heart Disease Analysis		
STUDENT NAME	Gautam Kumar	Rithvik K Bhat	Vignesh K
USN	1DS18CS710	1DS18CS732	1DS18CS744
INDIVIDUAL CONTRIBUTION	AWS Random Forest Algorithm	Decision Tree Algorithm Random Forest Algorithm	AWS Decision Tree Algorithm
GUIDE	Dr. Vinothini C		
PROJECT ABSTRACT	Heart Disease prediction is one of the most complicated tasks in medical field. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This project uses the heart disease data set available on the AWS cloud S3 platform. The project predicts the chances of heart disease and classifies patient's risk level by implementing two data mining techniques they are Decision Tree and Random Forest. Thus, this project analyses the performance of the two machine learning algorithms. The trial results verify that Random Forest algorithm has achieved the highest accuracy compared to Decision Tree.		
PLATFORM USED (H/W & S/W TOOLS TO BE USED	Jupyter Notebook & Amazon Web Services S3		

INTRODUCTION	Human heart is the main part of the human body which regulates blood flow throughout our body. Any irregularity to the heart can cause distress in other parts of the body and can be classified as a heart disease. Today, heart disease is one of the primary reasons for the occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension. The main challenge in today's healthcare is provision of best quality services and accurate diagnosis. The accuracy in management of a disease lies in the proper time of detection of that disease. The project makes an attempt to detect these heart diseases at an early stage to avoid disastrous consequences. Records of large set of medical data are available for analyzing and extracting valuable knowledge from it. Data mining techniques are used for extracting valuable information from the large data available. The medical database consists of discrete information which makes decision making a complex and tough task. Machine Learning (ML) which is a subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of our project is to provide a tool for doctors to detect heart disease at an early stage. This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and analyze the given data. This project shows the performance analysis of two ML techniques: Decision Tree and Random Forest for predicting heart disease at an early stage.
DESIGN	 Fig. 1 – Decision Tree Algorithm

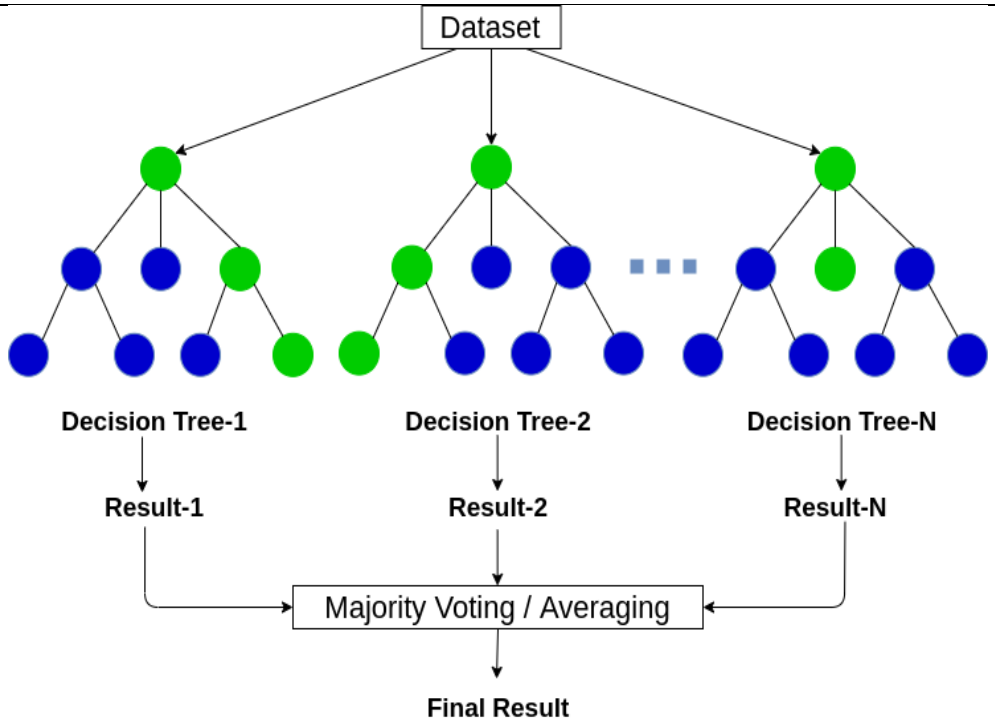
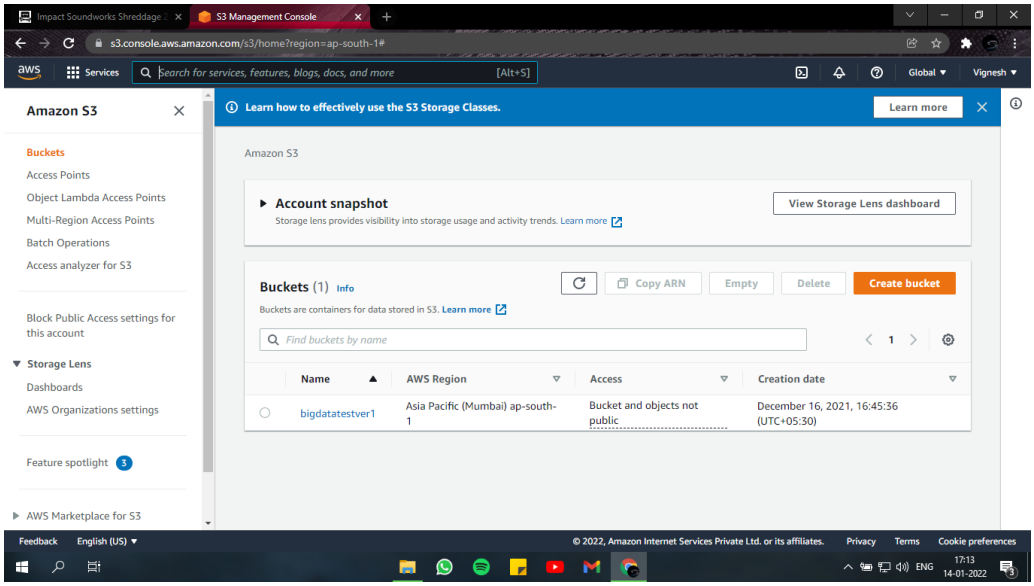
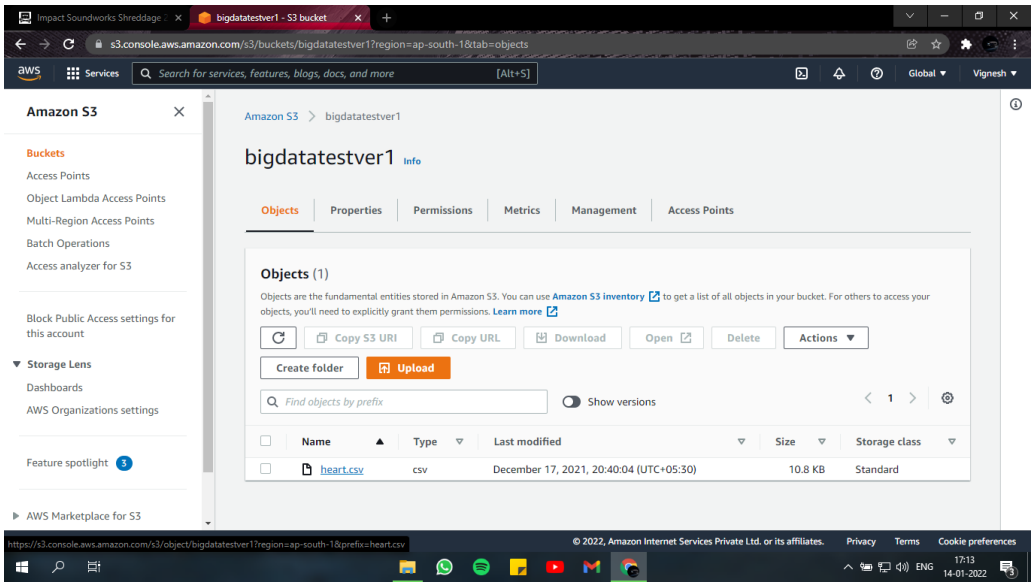
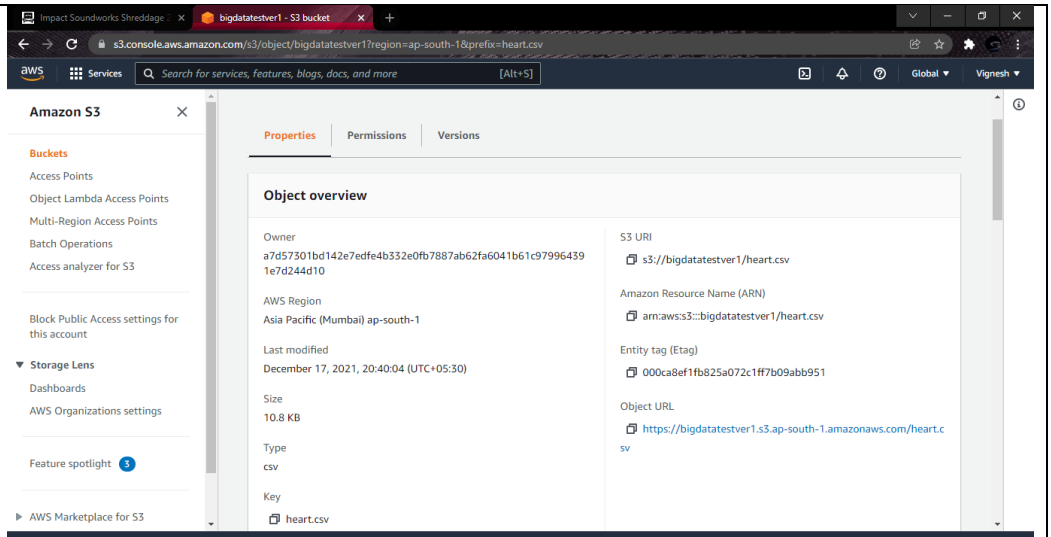
<table>
<tr><td></td><td> 

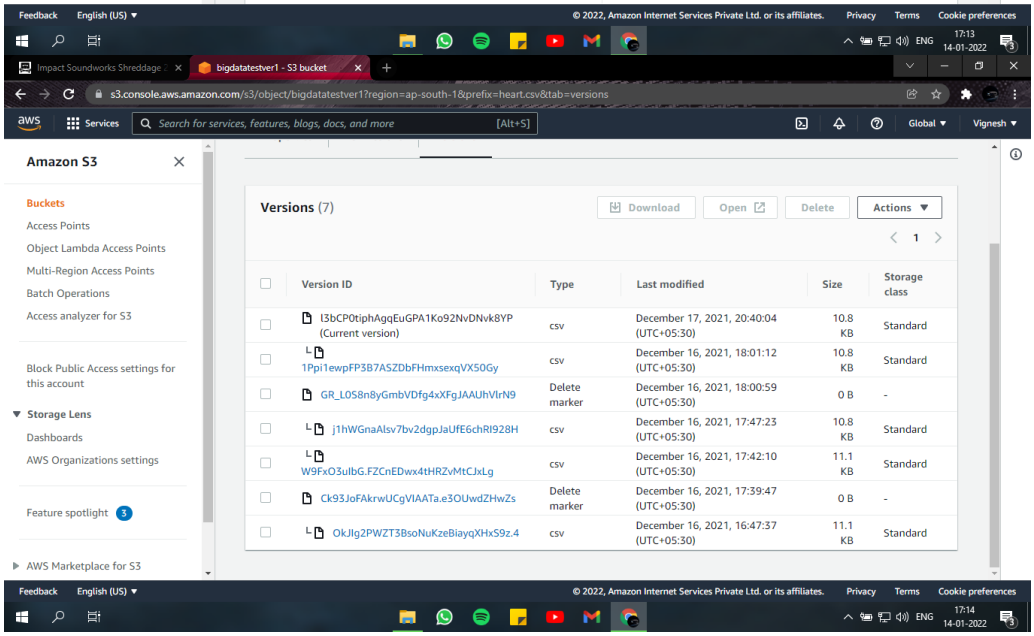
Fig.2 – Random Forest Algorithm</td></tr>
<tr><td></td><td></td></tr>
<tr><td>PROJECT SOURCE CODE LINK (GITHUB/ GOOGLE DRIVE)</td><td>https://github.com/gautamK007/Heart-Disease-Analysis</td></tr>
<tr><td></td><td></td></tr>
<tr><td>CONCLUSION /FUTURE ENHANCEMENT</td><td>Predicting heart disease is one of the most difficult problems in medicine. In the realm of healthcare, data science is critical for examining massive amounts of data. Because predicting cardiac illness is a difficult undertaking, it is necessary to automate the process in order to avoid the risks connected with it and to inform the patient well in advance. This project makes use of the AWS cloud S3 platform's heart disease data set. The study uses two data mining approaches, Decision Tree and Random Forest, to predict the likelihood of heart disease and classify the risk level of patients. As a result, the performance of the two machine learning algorithms is examined in this study. The trial results show that the Random Forest algorithm gives the best results.
For future enhancements, the project can be used to a variety of machine algorithms. Using visuals to detect cardiac disease could be a future improvement as well.</td></tr>
<tr><td></td><td></td></tr>
</table>

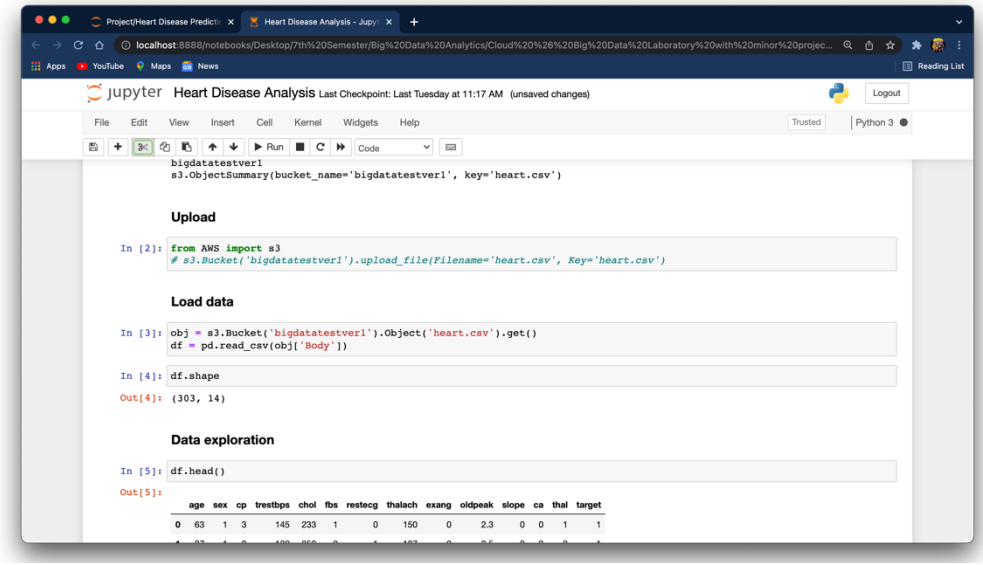
UI SCREENSHOTS

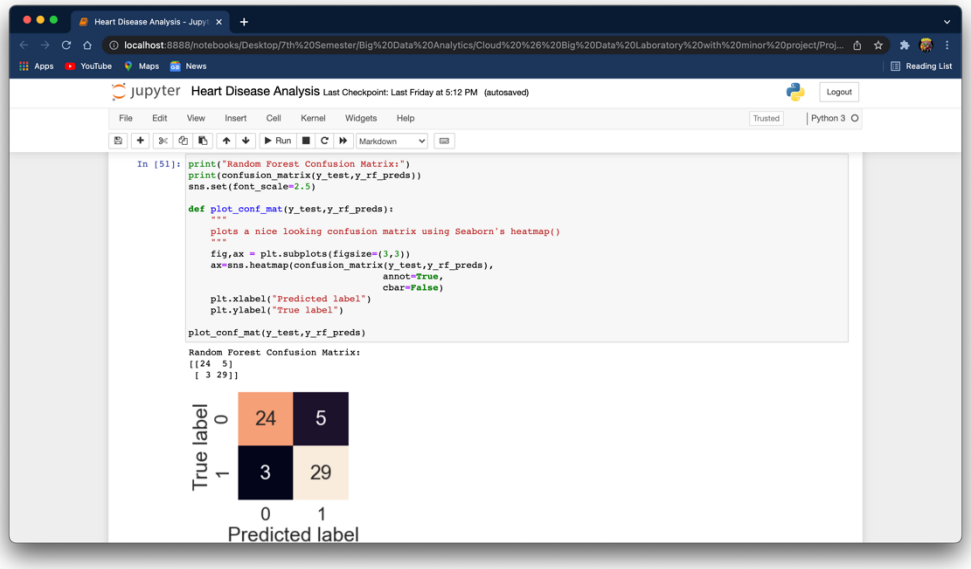


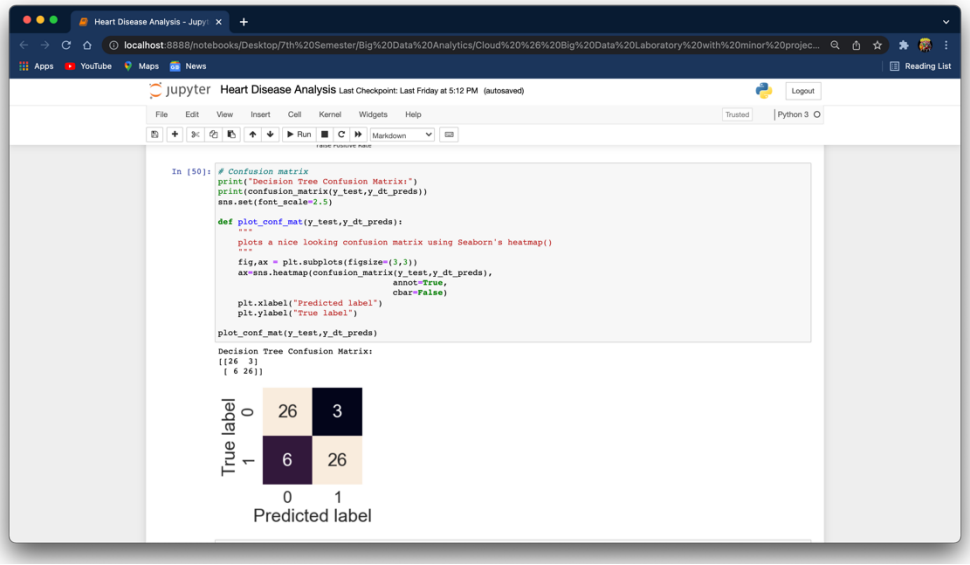


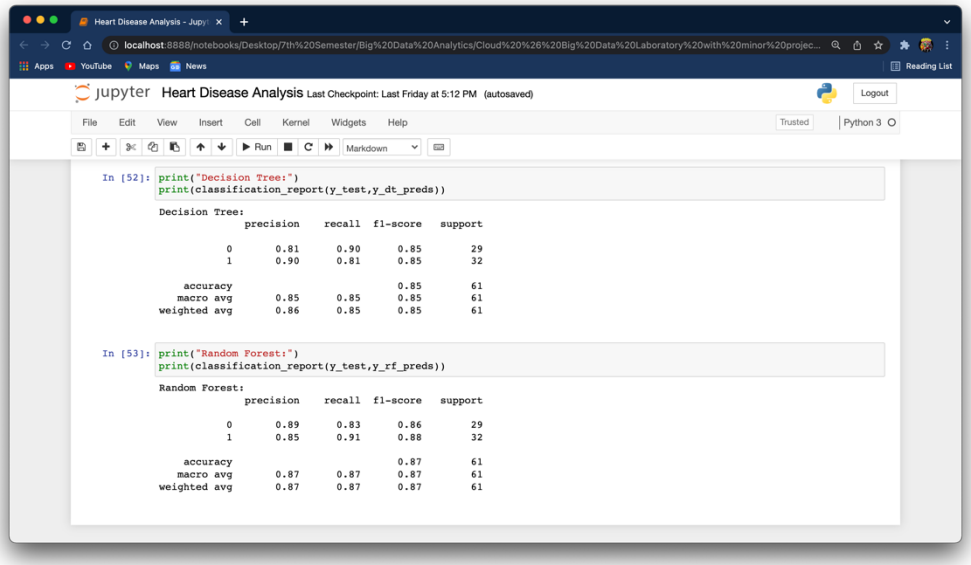

bigdatatestver1 - S3 bucket
s3.console.aws.amazon.com/s3/object/bigdatatestver1?region=ap-south-1&prefix=heart.csv
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Properties
Permissions
Versions
Object overview
Owner
a7d57301bd142e7edfe4b332e0fb7887ab62fa6041b61c979964391e7d244d10
AWS Region
Asia Pacific (Mumbai) ap-south-1
Last modified
December 17, 2021, 20:40:04 (UTC+05:30)
Size
10.8 KB
Type
csv
Key
heart.csv
S3 URI
s3://bigdatatestver1/heart.csv
Amazon Resource Name (ARN)
arn:aws:s3:::bigdatatestver1/heart.csv
Entity tag (Etag)
000ca8ef1fb825a072c1ff7b09abb951
Object URL
https://bigdatatestver1.s3.ap-south-1.amazonaws.com/heart.csv
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


bigdatatestver1 - S3 bucket
s3.console.aws.amazon.com/s3/object/bigdatatestver1?region=ap-south-1&prefix=heart.csv&tab=versions
Amazon S3
Versions (7)
Download
Open
Delete
Actions
Version ID
Type
Last modified
Size
Storage class
I3bCP0tiphAgqEuGPA1Ko92NvDNvk8YP (Current version) csv December 17, 2021, 20:40:04 (UTC+05:30) 10.8 KB Standard
1Ppi1ewpFP3B7ASZDbFHmxsexqVX50Gy csv December 16, 2021, 18:01:12 (UTC+05:30) 10.8 KB Standard
GR_L0S8n8yGmbVDfg4xXFgJAAUhVlrN9 Delete marker December 16, 2021, 18:00:59 (UTC+05:30) 0 B -
j1hWGnaAlsv7bv2dgpJaUfE6chRI928H csv December 16, 2021, 17:47:23 (UTC+05:30) 10.8 KB Standard
W9FxO3uIbG.FZCnEDwx4tHRZvMtCJxLg csv December 16, 2021, 17:42:10 (UTC+05:30) 11.1 KB Standard
Ck93JoFAkrwUCgVIAATa.e3OUwdZHwZs Delete marker December 16, 2021, 17:39:47 (UTC+05:30) 0 B -
OkJIg2PWZT3BsoNuKzeBiayqXHxS9z.4 csv December 16, 2021, 16:47:37 (UTC+05:30) 11.1 KB Standard


Heart Disease Analysis - Jupyter
localhost:8888/notebooks/Desktop/7th%20Semester/Big%20Data%20Analytics/Cloud%20%26%20Big%20Data%20Laboratory%20with%20minor%20projec...
jupyter Heart Disease Analysis Last Checkpoint: Last Tuesday at 11:17 AM (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
bigdatatestver1
s3.ObjectSummary(bucket_name='bigdatatestver1', key='heart.csv')
Upload
In [2]:
from AWS import s3
# s3.Bucket('bigdatatestver1').upload_file(Filename='heart.csv', Key='heart.csv')
Load data
In [3]:
obj = s3.Bucket('bigdatatestver1').Object('heart.csv').get()
df = pd.read_csv(obj['Body'])
In [4]:
df.shape
Out[4]:
(303, 14)
Data exploration
In [5]:
df.head()
Out[5]:
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
0 63 1 3 145 233 1 0 150 0 2.3 0 0 1 1


Heart Disease Analysis - Jupyter
localhost:8888/notebooks/Desktop/7th%20Semester/Big%20Data%20Analytics/Cloud%20%26%20Big%20Data%20Laboratory%20with%20minor%20project/Proj...
jupyter Heart Disease Analysis Last Checkpoint: Last Friday at 5:12 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown
In [51]:
print("Random Forest Confusion Matrix:")
print(confusion_matrix(y_test,y_rf_preds))
sns.set(font_scale=2.5)
def plot_conf_mat(y_test,y_rf_preds):
"""
plots a nice looking confusion matrix using Seaborn's heatmap()
"""
fig,ax = plt.subplots(figsize=(3,3))
ax=sns.heatmap(confusion_matrix(y_test,y_rf_preds),
annot=True,
cbar=False)
plt.xlabel("Predicted label")
plt.ylabel("True label")
plot_conf_mat(y_test,y_rf_preds)
Random Forest Confusion Matrix:
[[24 5]
[3 29]]
True label
0
1
24
5
3
29
0
1
Predicted label


Heart Disease Analysis - Jupyter
localhost:8888/notebooks/Desktop/7th%20Semester/Big%20Data%20Analytics/Cloud%20%26%20Big%20Data%20Laboratory%20with%20minor%20projec...
jupyter Heart Disease Analysis Last Checkpoint: Last Friday at 5:12 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown
In [50]:
# Confusion matrix
print("Decision Tree Confusion Matrix:")
print(confusion_matrix(y_test,y_dt_preds))
sns.set(font_scale=2.5)
def plot_conf_mat(y_test,y_dt_preds):
"""
plots a nice looking confusion matrix using Seaborn's heatmap()
"""
fig,ax = plt.subplots(figsize=(3,3))
ax=sns.heatmap(confusion_matrix(y_test,y_dt_preds),
annot=True,
cbar=False)
plt.xlabel("Predicted label")
plt.ylabel("True label")
plot_conf_mat(y_test,y_dt_preds)
Decision Tree Confusion Matrix:
[[26 3]
[6 26]]
True label
0
1
26
3
6
26
0
1
Predicted label


Heart Disease Analysis - Jupyter
localhost:8888/notebooks/Desktop/7th%20Semester/Big%20Data%20Analytics/Cloud%20%26%20Big%20Data%20Laboratory%20with%20minor%20projec...
jupyter Heart Disease Analysis Last Checkpoint: Last Friday at 5:12 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown
In [52]:
print("Decision Tree:")
print(classification_report(y_test,y_dt_preds))
Decision Tree:
precision recall f1-score support
0 0.81 0.90 0.85 29
1 0.90 0.81 0.85 32
accuracy 0.85 61
macro avg 0.85 0.85 0.85 61
weighted avg 0.86 0.85 0.85 61
In [53]:
print("Random Forest:")
print(classification_report(y_test,y_rf_preds))
Random Forest:
precision recall f1-score support
0 0.89 0.83 0.86 29
1 0.85 0.91 0.88 32
accuracy 0.87 61
macro avg 0.87 0.87 0.87 61
weighted avg 0.87 0.87 0.87 61