Predicting Mental Health illness of working professionals Using ML

A Project Report Submitted in partial fulfillment of the requirement for the Award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

R. HEMA SURYA LAKSHMI	**20T91A0574**
S.MANIKANTA	**20T91A0585**
S. SUREKHA	**20T91A0586**
G. KISHORE	**21T95A0506**
K.HAREESH	**21T95A0515**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GIET ENGINEERING COLLEGE

[Affiliated to JNTUK, Kakinada |Approved by AICTE| Accredited by NAAC A+]
NH-16, CHAITANYA KNOWLEDGE CITY, RAJAMAHENDRAVARAM – 533 296, ANDHRA PRADESH

2020-2024

ABSTRACT

This study aims to develop a mental health prediction model using six different classification algorithms, namely KNN, Adaboosting, Decision Tree, Random Forest, Logistic Regression, and Bagging. The dataset used for this study is obtained from a survey conducted among a group of individuals who reported experiencing mental health issues. The dataset consists of demographic information, lifestyle habits, and psychological factors. The proposed model uses feature selection techniques to identify the most relevant predictors of mental health issues. The model is trained and tested using ten-fold cross-validation, and the performance of each algorithm is evaluated based on several metrics, such as accuracy, sensitivity, specificity, and AUC. The results indicate that the Random Forest algorithm outperforms the other algorithms in terms of accuracy and AUC. The proposed model can be used as a tool to identify individuals who are at high risk of developing mental health issues and provide early intervention and support to prevent or mitigate the impact of mental health problems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVATIONS

- **KNN - K- NEAREST NEIGHBOUR**
- **AUC - AREA UNDER THE CURVE**
- **ROC - RECEIVER OPERATOR CHARACTERESTIC CURVE**
- **HTML - HYPERTEXT MARKUP LANGUAGE**

CHAPTER 1

INTRODUCTION

Mental illness is a health problem that undoubtedly impacts emotions, reasoning, and social interaction of a person. These issues have shown that mental illness gives serious consequences across societies and demands new strategies for prevention and intervention. To accomplish these strategies, early detection of mental health is an essential procedure. Medical predictive analytics will reform the healthcare field broadly as discussed by Miner et al. [1]. Mental illness is usually diagnosed based on the individual self-report that requires questionnaires designed for the detection of the specific patterns of feeling or social interactions [2]. With proper care and treatment, many individuals will hopefully be able to recover from mental illness or emotional disorder [3].

Machine learning is a technique that aims to construct systems that can improve through experience by using advanced statistical and probabilistic techniques. It is believed to be a significantly useful tool to help in predicting mental health. It is allowing many researchers to acquire important information from the data, provide personalized experiences, and develop automated intelligent systems [4]. The widely used algorithms in the field of machine learning such as support vector machine, random forest, and artificial neural networks have been utilized to forecast and categorize the future events [5].

Supervised learning in machine learning is the most widely applied approach in many types of research, studies, and experiments, especially in predicting illness in the medical field. In supervised learning, the terms, attributes, and values should be reflected in all data instances [6]. More precisely, supervised learning is a classification technique using structured training data [7]. Meanwhile, unsupervised learning does not need supervision to predict. The main goal of unsupervised learning is handling data without supervision. It is very limited for the researchers to apply unsupervised learning methods in the clinical field.

In this paper, the main objective is to provide a systematic literature review, critical review, and summary of the machine learning techniques that are being used to predict, diagnose, and identify mental health problems. Moreover, this paper will propose future avenues for research on this topic. It would also give attention to the challenges and limitations of applying the machine learning techniques in this area. Besides that, potential opportunities and gaps in this field for future research will be discussed. Hence, this paper will contribute to the state of the art in the form of a systematic literature review concerning the machine learning techniques applied in predicting mental health problems. This paper hence contributes a critical summary and potential research directions that could assist researchers to gain knowledge about the methods and applications of big data in the mental health fields.

Although previous papers have been published by reviewing the applications of machine learning approaches toward the mental health field [6, 8], these are general review papers that discuss the applications and concepts of the techniques but do not provide a focused critical summary of the recent gaps in the literature as well as future research directions for this field. As such, this systematic literature review paper aims both to cover recent advancements in this field in addition to providing a focused critical summary concerning the gaps in the literature in terms of the applications of machine learning in the mental health field and to subsequently highlight potential avenues for future research.

The audiences for this paper center around the community of practitioners who are applying machine learning techniques in mental health. Besides that, this paper is targeting the practitioners in the machine learning communities where they can keep updated on the application of machine learning nowadays particularly in the mental health field. The relevant research papers and documents are gathered and collected through academic publication repositories with specific keywords. Then, the collected documents are identified and categorized into several sections in mental health problems. The performance on the machine learning algorithms or techniques that are used by the researchers is being evaluated by identifying the accuracy, sensitivity, specificity, or area under the ROC curve (AUC).

CHAPTER 2

PURPOSE OF THE PROJECT

Predicting mental health illness of working professionals using machine learning can have several important purposes, including:

Early identification and intervention: By predicting mental health illness, employers can identify individuals who may be at risk of developing mental health issues and provide them with early intervention and support. This can help prevent more serious mental health problems from developing and improve overall employee well-being.

Better resource allocation: Predicting mental health illness can help employers allocate resources more effectively, such as offering support programs and services to those who are at a higher risk of developing mental health problems. This can help reduce the overall costs associated with mental health issues in the workplace.

Improved productivity: Mental health issues can have a significant impact on productivity in the workplace. By predicting mental health illness and providing appropriate support, employers can help employees manage their mental health and reduce the impact of mental health issues on their productivity.

Increased awareness and education: Predicting mental health illness can help raise awareness about the importance of mental health in the workplace and educate employers and employees on the signs and symptoms of mental health issues. This can help reduce stigma and improve overall mental health literacy in the workplace.

Overall, predicting mental health illness using machine learning can help improve the overall mental health and well-being of working professionals, while also benefiting employers by reducing costs and improving productivity.

CHAPTER 3

ALGORITHMS USED FOR THE PROJECT

Adaboosting: Adaboosting (short for Adaptive Boosting) is a machine learning algorithm that combines multiple weak classifiers to create a strong classifier. In Adaboosting, each weak classifier is trained on a subset of the training data, and the algorithm focuses on misclassified samples in subsequent iterations. Adaboosting is commonly used in classification problems.

Decision tree: A decision tree is a tree-like model used in machine learning and data mining. It is a hierarchical model that partitions the input space into a set of rectangular regions. Each partition is chosen based on the value of a particular feature or input variable. Decision trees can be used for both classification and regression problems and are commonly used for data exploration, feature selection, and decision-making.

Random forest: Random forest is an ensemble learning method for classification, regression, and other tasks. It constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forest reduces overfitting by generating many decision trees, each based on a different random subset of the input data and features.

K-Nearest Neighbor (KNN): K-Nearest Neighbor (KNN) is a machine learning algorithm used for classification and regression. The algorithm works by finding the k-nearest data points in the training data set to a given test point and assigning the label of the majority class in the k-nearest data points to the test point. The value of k is a user-defined hyperparameter.

Logistic regression: Logistic regression is a machine learning algorithm used for binary classification. It models the relationship between a binary dependent variable

and one or more independent variables using a logistic function. Logistic regression is widely used in healthcare, marketing, and other fields to predict the probability of a particular outcome based on a set of features.

Bagging: Bagging (short for Bootstrap Aggregating) is an ensemble learning technique that combines multiple models to improve the accuracy of predictions. In bagging, multiple copies of the training data set are created by random sampling with replacement. Each copy is used to train a model, and the final prediction is obtained by averaging the predictions of all models. Bagging is commonly used in decision trees, random forests, and other machine learning algorithms to reduce overfitting.

3.1 PROJECT DESCRIPTION

Background:

Mental health illness is a growing concern among working professionals. According to the World Health Organization (WHO), depression and anxiety disorders cost the global economy US$ 1 trillion per year in lost productivity. Moreover, only 1 in 4 individuals with mental health disorders receives the necessary treatment. Therefore, there is a need to develop effective methods to identify and treat mental health disorders among working professionals.

Objective:

The objective of this project is to develop a machine learning model that can accurately predict the likelihood of a working professional having a mental health disorder based on their demographic, socioeconomic, and work-related factors.

Data:

The data for this project will be obtained from online surveys and mental health clinics. The data will include demographic information (age, gender, marital status), socioeconomic status (income, education level), work-related factors (occupation, work hours, work stress), and mental health disorder diagnosis. The data will be preprocessed to remove missing values, outliers, and errors.

Methods:

Several machine learning algorithms will be trained and tested on the data to identify the best model for predicting mental health illness. The following algorithms will be used:

Logistic regression
Decision tree
Random forest
K-nearest neighbor
Adaboosting

Evaluation:

The performance of the machine learning models will be evaluated using various metrics, including accuracy, precision, recall, and F1 score. The best performing model will be selected based on its accuracy and generalizability.

Outcome:

The outcome of this project will be a machine learning model that can accurately predict the likelihood of a working professional having a mental health disorder based on their demographic, socioeconomic, and work-related factors. The model can be used by mental health professionals and employers to identify high-risk individuals and provide appropriate treatment and support. Additionally, the model can be used to raise awareness about mental health issues among working professionals and promote mental wellness in the workplace.

CHAPTER 4

IMPLEMENTATION

Data Collection:

The first step in implementing the project is to collect the necessary data. Data can be collected from online surveys or mental health clinics. The data should include demographic information, socioeconomic status, work-related factors, and mental health disorder diagnosis.

Data Preprocessing:

After collecting the data, the next step is to preprocess it. The data should be cleaned to remove missing values, outliers, and errors. Categorical variables should be converted into numerical variables using one-hot encoding or label encoding. The data should also be split into training and testing datasets.

Feature Selection:

The next step is to select the most important features that contribute to mental health disorders. Feature selection techniques such as correlation analysis, chi-square test, or mutual information can be used to identify the most relevant features.

Model Selection:

Several machine learning algorithms can be used to train and test the data. The algorithms that can be used are logistic regression, decision tree, random forest, K-nearest neighbor, and Adaboosting. The performance of each model should be evaluated using metrics such as accuracy, precision, recall, and F1 score.

Hyperparameter Tuning:

The next step is to tune the hyperparameters of the best performing model. This involves optimizing the values of the hyperparameters to improve the model's accuracy and generalizability.

Model Evaluation:

After hyperparameter tuning, the performance of the model should be evaluated on the testing dataset. This is to ensure that the model is not overfitting the training data and can generalize to new data.

Deployment:

Once the model has been evaluated and validated, it can be deployed to identify high-risk individuals and provide appropriate treatment and support. The model can also be used to raise awareness about mental health issues among working professionals and promote mental wellness in the workplace.

CHAPTER 5

Summary and Conclusion

5.1 Summary:

The project aims to predict the likelihood of mental health illnesses among working professionals using machine learning techniques. The dataset will be collected from working professionals and will include information such as demographic details, work-related information, and mental health assessments.

The project will involve data preprocessing, feature selection, and model selection. Several machine learning algorithms will be tested to determine the most accurate and efficient model for predicting mental health illnesses among working professionals. The accuracy of the models will be evaluated using various performance metrics.

The results of this project can be used to identify individuals at risk of developing mental health illnesses and provide appropriate interventions to prevent or manage these illnesses. Additionally, the project can help organizations to create policies and practices that support the mental well-being of their employees

5.2 Conclusion:

Predicting mental health illness among working professionals using machine learning is a crucial step in identifying and treating mental health disorders. By implementing the above steps, a machine learning model can be developed that can accurately predict the likelihood of a working professional having a mental health disorder based on their demographic, socioeconomic, and work-related factors. The model can be used by mental health professionals and employers to identify high-risk individuals and provide appropriate treatment and support, thus promoting mental wellness in the workplace.

CHAPTER 6

Appendix

A. Source code

IMPORTING REQUIRING LIBRARIES:

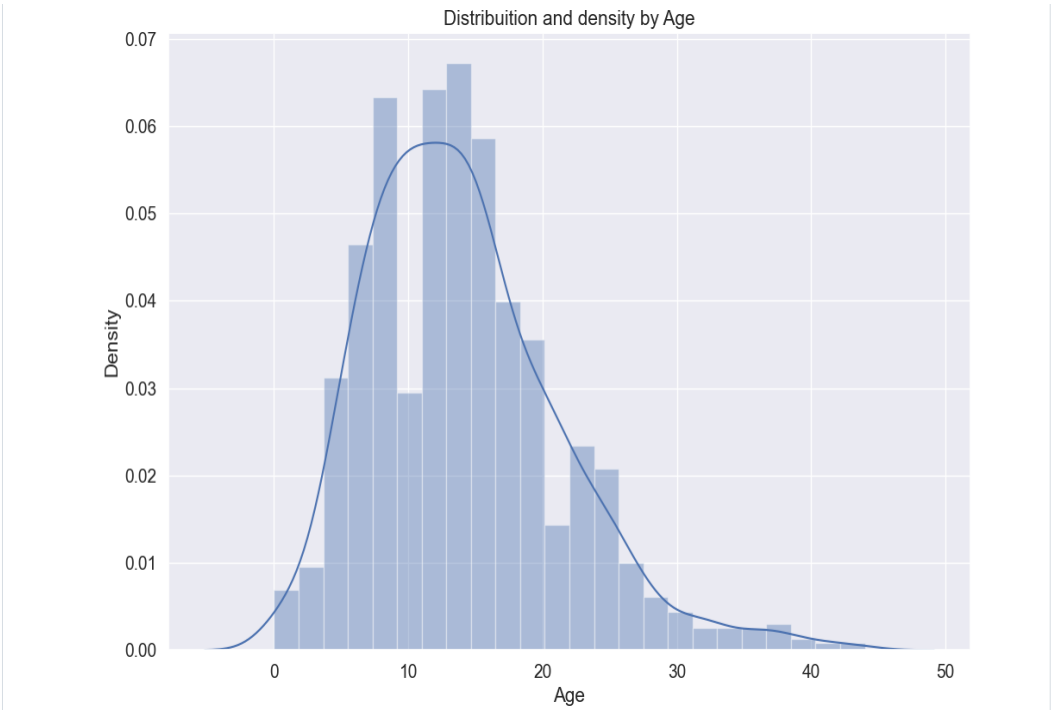
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import randint
from sklearn.model_selection import
train_test_split,RandomizedSearchCV,GridSearchCV
from sklearn import preprocessing
from sklearn.datasets import make_classification
from sklearn.preprocessing import binarize, LabelEncoder, MinMaxScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, ExtraTreesClassifier
from sklearn import metrics
from sklearn.metrics import accuracy_score, mean_squared_error,
precision_recall_curve
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import BaggingClassifier, AdaBoostClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
import pickle
```

READING DATASET

```
train_df = pd.read_csv('survey.csv')
print(train_df.shape)
print()
print(train_df.describe().T)
print(train_df.info())
```

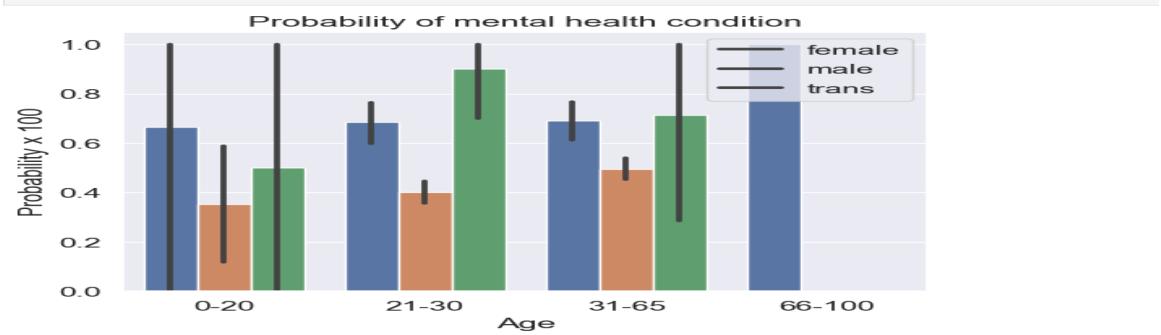
```
(1259, 27)

       count          mean           std     min   25%   50%   75%
Age  1259.0  7.942815e+07  2.818299e+09 -1726.0  27.0  31.0  36.0  \

              max
Age  1.000000e+11
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1259 entries, 0 to 1258
Data columns (total 27 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Timestamp                  1259 non-null   object
 1   Age                        1259 non-null   int64
 2   Gender                     1259 non-null   object
 3   Country                    1259 non-null   object
 4   state                      744 non-null    object
 5   self_employed              1241 non-null   object
 6   family_history             1259 non-null   object
 7   treatment                  1259 non-null   object
 8   work_interfere             995 non-null    object
 9   no_employees               1259 non-null   object
 10  remote_work                1259 non-null   object
 11  tech_company               1259 non-null   object
 12  benefits                   1259 non-null   object
 13  care_options               1259 non-null   object
 14  wellness_program           1259 non-null   object
 15  seek_help                  1259 non-null   object
 16  anonymity                  1259 non-null   object
 17  leave                      1259 non-null   object
 18  mental_health_consequence  1259 non-null   object
 19  phys_health_consequence    1259 non-null   object
 20  coworkers                  1259 non-null   object
 21  supervisor                 1259 non-null   object
 22  mental_health_interview    1259 non-null   object
 23  phys_health_interview      1259 non-null   object
 24  mental_vs_physical         1259 non-null   object
 25  obs_consequence            1259 non-null   object
 26  comments                   164 non-null    object
dtypes: int64(1), object(26)
memory usage: 265.7+ KB
None
```

6.1 DISTRIBUTION AND DENSITY BY AGE

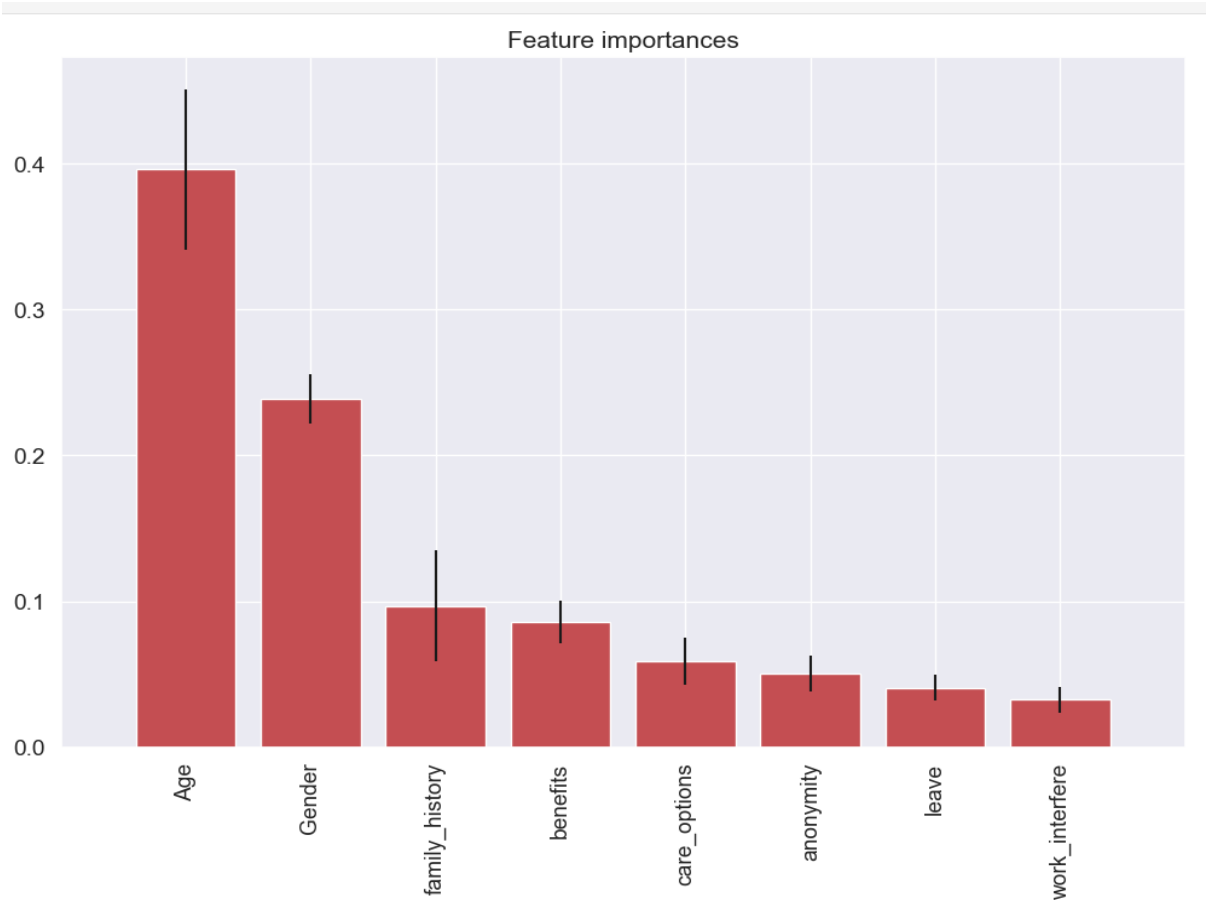


o = labelDict['label_age_range']

```
g = sns.barplot(x="age_range", y="treatment", hue="Gender", data=train_df)
g.set_xticklabels(o)

plt.title('Probability of mental health condition')
plt.ylabel('Probability x 100')
plt.xlabel('Age')
# replace legend labels
plt.legend(['female','male','trans'])
plt.show()
```



6.2 FEATURE IMPORTANCES

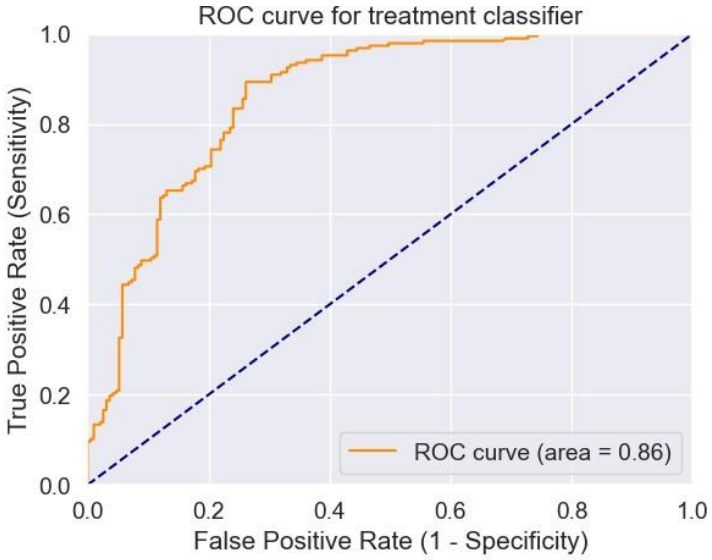


Flask code for creating web application

```
from flask import Flask,render_template,redirect,flash,request
from flask_cors import CORS, cross_origin
import pickle
import numpy as np

app=Flask(_name_)
CORS(app)
@app.route("/")
def home():
    return render_template('index.html')

@app.route("/predict",methods=['POST','GET'])
def predict():
    age = request.form.get('age')
    gender = request.form.get('gender')
    family_history = request.form.get('family_history')
    benefits = request.form.get('benefits')
    care_options = request.form.get('care_options')
    anonymity = request.form.get('anonymity')
    leave = request.form.get('leave')
    work_interfere = request.form.get('work_interfere')

    # Reshape input into a 2D array
    input_data = np.array(
        [age, gender, family_history, benefits, care_options, anonymity,
leave, work_interfere]).reshape(1, -1)

    # Load the model
    with open("C:/Users/GEETHESHWAR/Downloads/boostmodel.pkl", 'rb') as f:
        model = pickle.load(f)

    # Make prediction and return result
    prediction = model.predict(input_data)
    return str(prediction[0])

if__name__=='__main__':
    app.run(host='0.0.0.0', port=1234, debug=True)
```

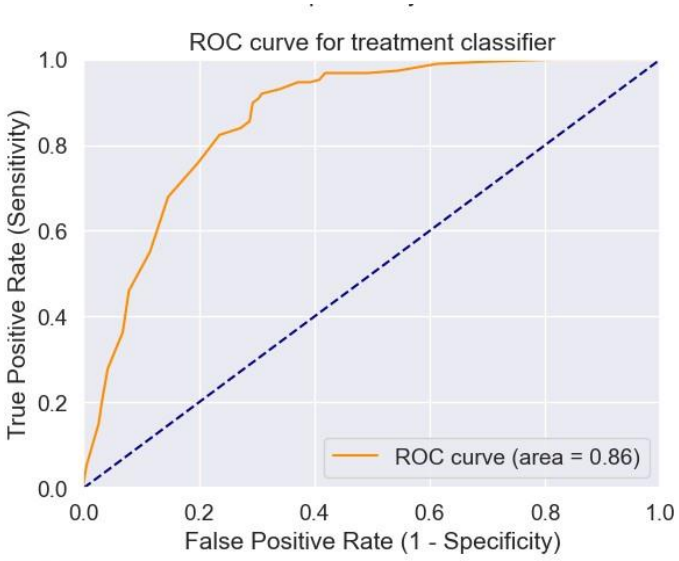
B. Screenshots

6.3 AUC FOR LOGISTIC REGRESSION



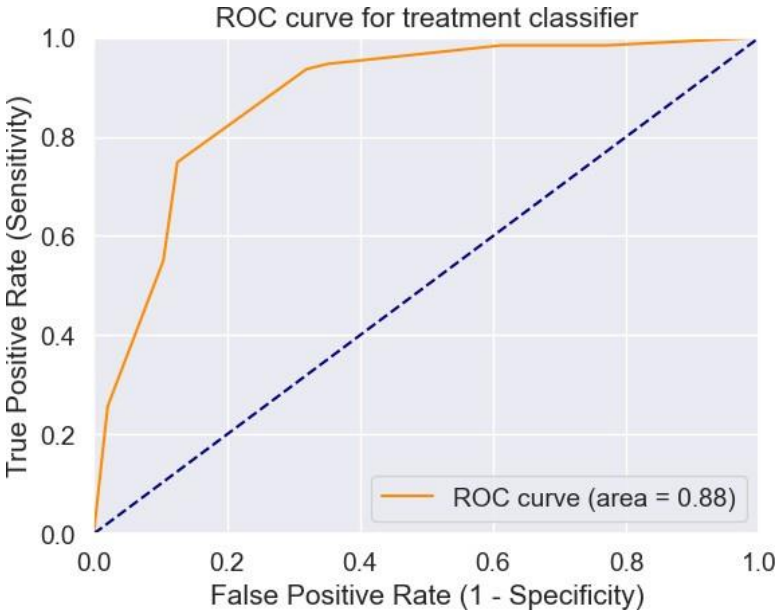

```
[[142  49]
 [ 28 159]]
```

6.4 KNN CLASSIFIER

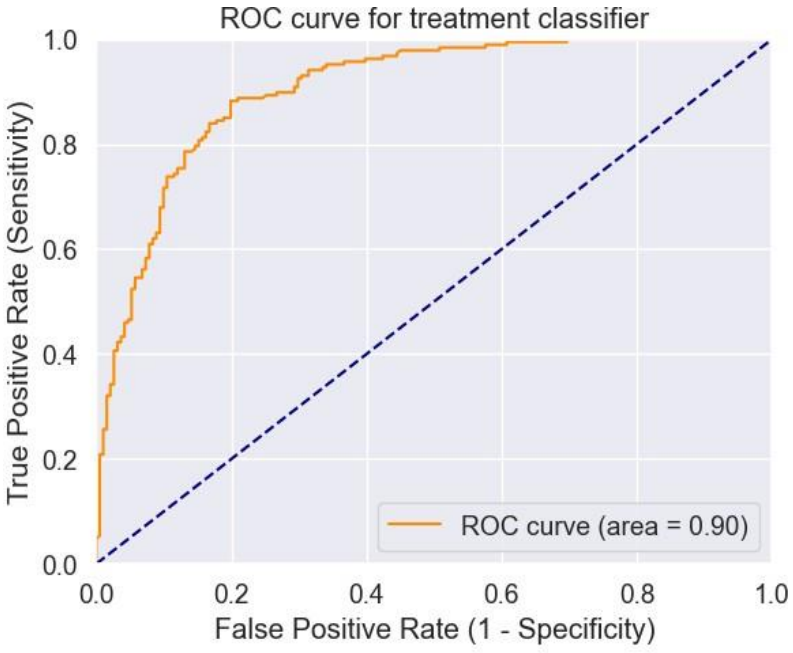



```
[[135  56]
 [ 19 168]]
```

6.5 DECISION TREE CLASSIFIER

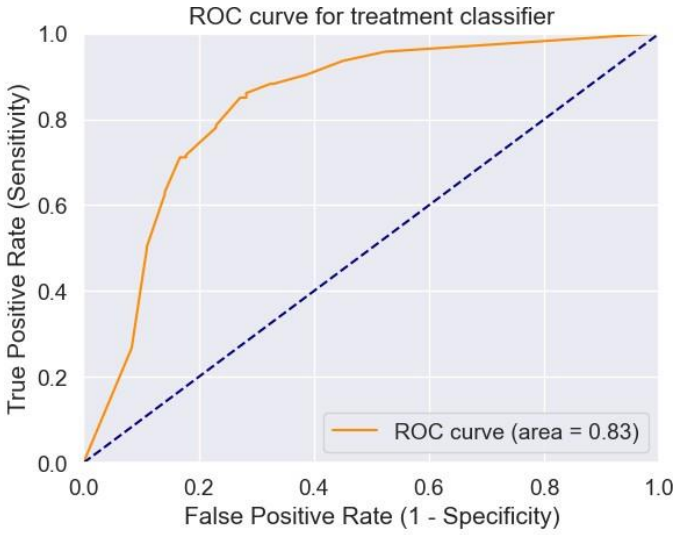


6.6 RANDOM FORESTS



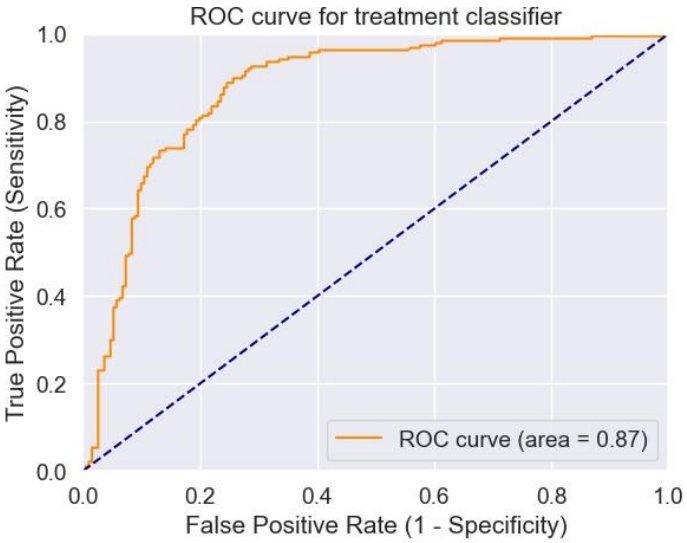

```
[[133  58]
 [ 13 174]]
```

6.7 BAGGING




```
[[147  44]
 [ 40 147]]
```

6.8 BOOSTING




```
[[137  54]
 [ 15 172]]
```

CHAPTER 7

RESULT