[HS-MTCS-083]

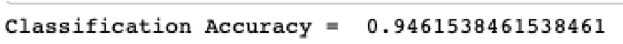
[Predicting Alzheimer’s Disease]

Q1: Problem & Objectives

- The Problem that I tackled was developing a machine learning model that can accurately classify the condition of a patient and their future MMSE score with data that doesn’t require visiting hospitals (Ex: MRI scans).
- There are several models that can classify Alzheimer's Disease, but the accurate ones require brain imaging (MRI) and advanced data.
- Thus, my model is unique as it can accurately classify the state of a patient and predict their future MMSE score with easily accessible data.

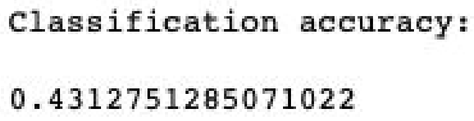
Q3: Data Analysis & Results

- The optimal machine learning model for classifying Alzheimer's Disease was based on Naive Bayes with an accuracy of around 95% with an AUC = 0.9



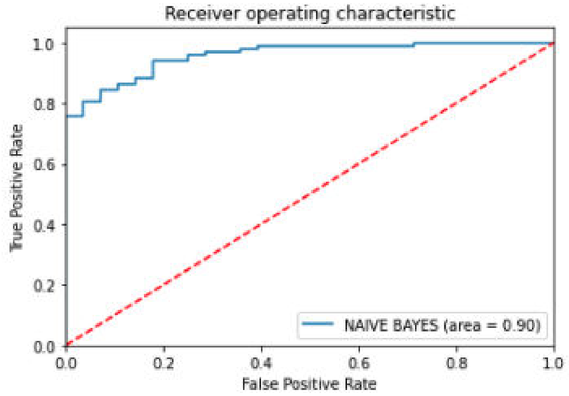
```
Classification Accuracy =  0.9461538461538461
```

- The optimal machine learning model for predicting the change in MMSE score was based on Elastic Net with an accuracy of 43%



```
Classification accuracy:

0.4312751285071022
```



Q2: Project Design

- First, I applied to the Alzheimer’s Disease Neuroimaging Initiative to receive the data needed
- Then, I preprocessed the data and used several machine learning models to classify the patient status and regression models to predict the change in MMSE scores (an indicator of dementia).
- Finally, I created a script that takes the information that a user inputs on a google form and sends an automated email with the classification and predicted score.

Q4: Conclusions

- The model classifies the condition of a patient with a much higher accuracy than my benchmark goal of 80%. However, the prediction of exact decrease in MMSE scores had a classification accuracy of only 0.43.
- This is to be expected, as it's trying to predict the exact change in the test test score. If we give the model a margin of error of one point its accuracy goes to over 80%.
- For the future, I’d like to convert this into an app rather than a google form and get more data to improve the model.

Introduction

Research Problem:

- Develop a model to accurately diagnose a patient with Alzheimer's Disease without using high costing data such as structural imaging with magnetic resonance imaging (MRI) or computed tomography (CT) and develop a model to predict the future cognitive state with easily accessible information and have an accuracy above 80%

Constraints:

- The constraints were that I couldn't use expensive data like MRI or PET to train my model and the limited amount of data that I had from ADNI was a constraint.

Background:

- The current research that has been done for classifying Alzheimer's Disease is primarily focused on using MRI imaging and Deep Learning (Convolutional neural network). While this research has brought fruitful results, it also fundamentally doesn't solve a problem that it's tasked with solving. There are many models that boast high accuracy (above 97%) in classifying and predicting AD, but they are developed and require using magnetic resonance imaging, positron emission tomography, and molecular biomarkers

- Now what's the problem? Well, the major problem is that those tests require that patients visit hospitals and they cost thousands of dollars. Millions of people simply can't afford to pay thousands of dollars for testing and even though some can, the problem of hospitals not being accessible also exists.

- Thus, it's imperative that models be created to classify if a patient has Alzheimer's Disease with data that can be easily obtained.

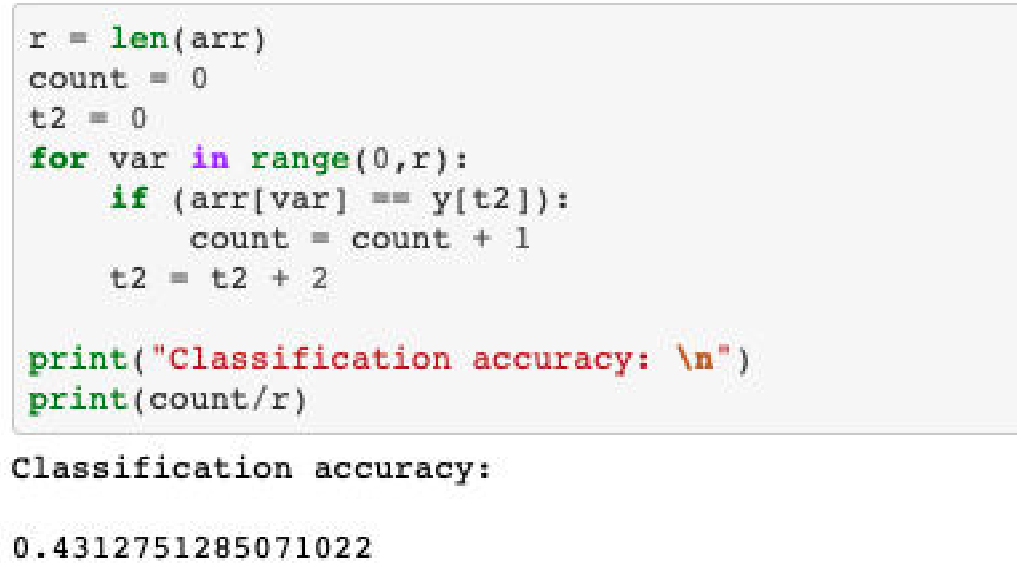
- The work that I have done is unique as it can classify if a patient has Alzheimer's Disease with an accuracy of around 95% and an AUC of 0.9. There exists other models, but all of them require MRI imaging and expensive data, which makes my model stand out due to its practicality. Additionally, my model can predict the change in Mini-Mental State Examination scores (an indicator of dementia) with an accuracy of over 80% — assuming Margin of Error of 1. Additionally, I made my model accessible to anyone. To get their information, people only need to fill out a google form and the script will send an automated email with their results.

Methods

- What did I do?
  - The process for developing my models was quite strenuous and lengthy but was also rewarding.
  - First, I joined the Alzheimer's Disease Neuroimaging Initiative (ADNI), and spent a few hours scouring through the database and downloaded the data that was the most applicable to my goals.
  - Then, I needed to preprocess the data. I went through my data in Jupyter notebook and deleted the unnecessary data such as Participant ID, Examination Date, and more. After that, I implemented One Hot Encoder to convert the categorical variables into binary to make the implementation of machine learning models simpler.
  - After that, I trained several machine learning models (to predict the change in Mini-Mental State Examination first) on the data. Including Linear Regression, Decision Trees & Random Forests, Neural Networks & K Nearest Neighbors, and Elastic Net to name a few.
  - The overall data was split, so that a majority would be used to train, and there would be another portion left for testing after fitting the model. Using that portion, I tested the results of each model individually using MAE and other indicators.
  - From those, I ended up choosing Elastic Net as the final model as it produced the lowest MAE (Mean absolute error). Since, regression is a continuous output, I converted all the outputs to integers and set a limit of ±5 for the change in MMSE scores.
  - Then, I created a python script that would take information from a google form and send the person who submitted the form an email with their predicted change in MMSE score.
  - After this, I started working on the classification algorithm. The data was different than the MMSE prediction's data, so I had to preprocess it. I removed the classifier MCI and instead changed the output to either "AD" or "NOT AD" for the purpose of implementing binary classification rather than Multi-class. Additionally, this time I removed all the outliers in the data to improve the model.
  - Just like the MMSE program, I created several models for testing. Including Naïve Bayes, Logistic Regression, K-Nearest Neighbors, Decision Trees, and SVM (using all 4 kernels). The model implementing Naïve Bayes was the most accurate as it had the highest Classification Accuracy and AUC.
  - Finally, I created another python script that takes information from a google form and sends an automated email with classification.

Results — MMSE Prediction

The Elastic Net Algorithm was able to predict the change in MMSE test scores with an accuracy of 0.43. While this number is relatively small, it's the accuracy of predicting the change perfectly without any margin of error.



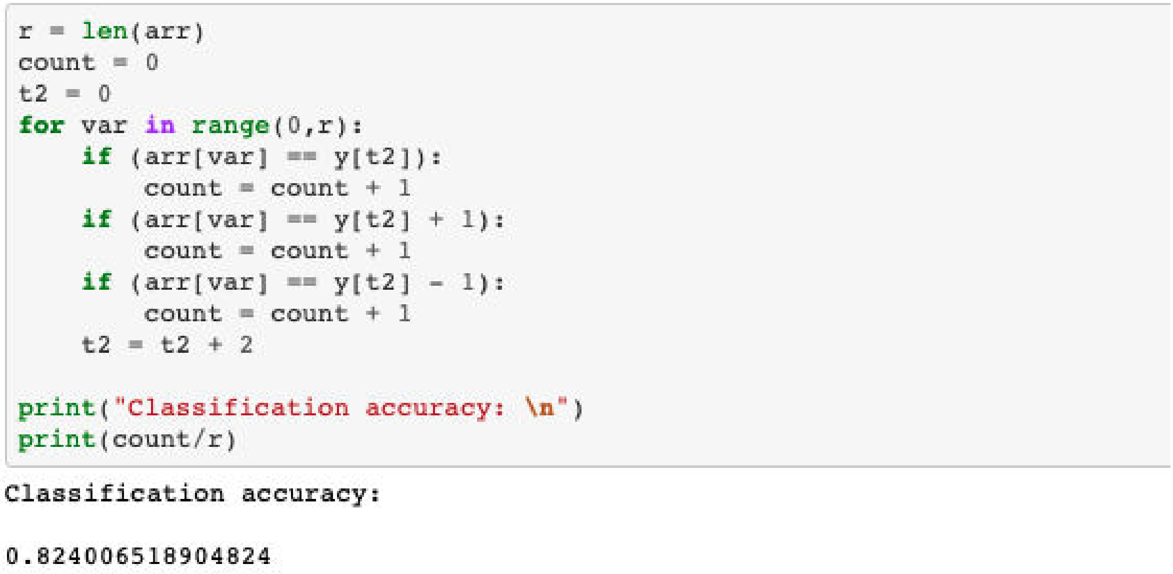
```
r = len(arr)
count = 0
t2 = 0
for var in range(0,r):
    if (arr[var] == y[t2]):
        count = count + 1
    t2 = t2 + 2

print("Classification accuracy: \n")
print(count/r)
```

```
Classification accuracy:

0.4312751285071022
```

The Elastic Net Algorithm was able to predict the change in MMSE test scores with an accuracy of 0.824 when a margin of error of 1 is considered.



```
r = len(arr)
count = 0
t2 = 0
for var in range(0,r):
    if (arr[var] == y[t2]):
        count = count + 1
    if (arr[var] == y[t2] + 1):
        count = count + 1
    if (arr[var] == y[t2] - 1):
        count = count + 1
    t2 = t2 + 2

print("Classification accuracy: \n")
print(count/r)
```

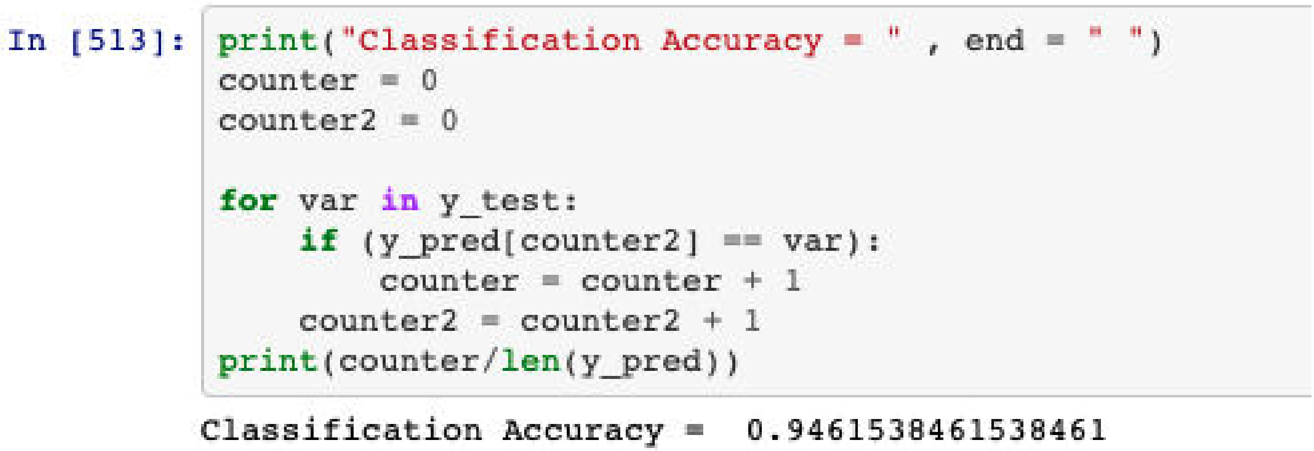
```
Classification accuracy:

0.824006518904824
```

This prototype meets my engineering goal as it is accurately predicting the change in MMSE scores. While the accuracy for predicting it on the dot is quite average, the accuracy with a margin of error of 1 (meaning that it's a success if the prediction is 1 away from the actual result) is 0.824. Having an accurate model that can predict the change perfectly is a tough task and is likely impossible but predicting the change in MMSE with a margin of error of 1 accurately is a good benchmark.

Results — Classification

The Naive Bayes model was able to accurately predict if a person has Alzheimer's disease or not with a classification accuracy of 0.946 and an AUC score of 0.9, which is an outstanding result.



```
In [513]: print("Classification Accuracy = " , end = " ")
          counter = 0
          counter2 = 0

          for var in y_test:
              if (y_pred[counter2] == var):
                  counter = counter + 1
              counter2 = counter2 + 1
          print(counter/len(y_pred))

          Classification Accuracy =  0.9461538461538461
```

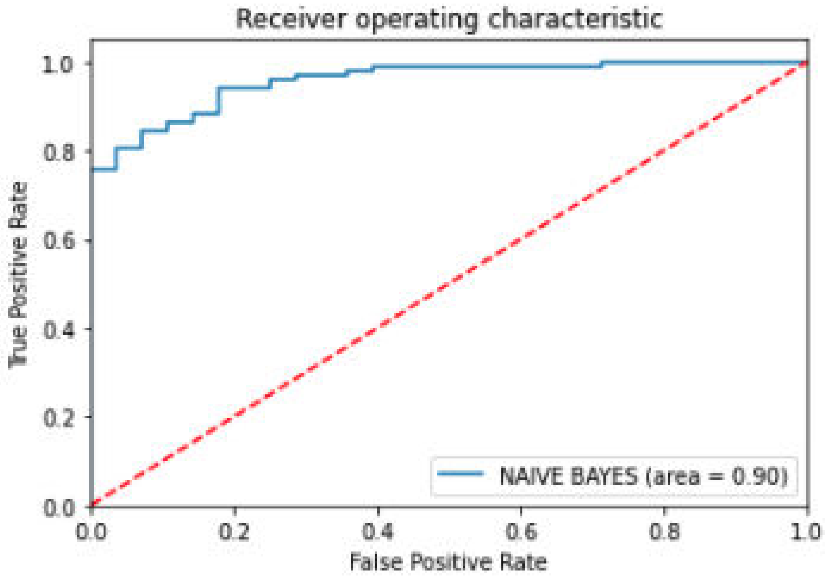
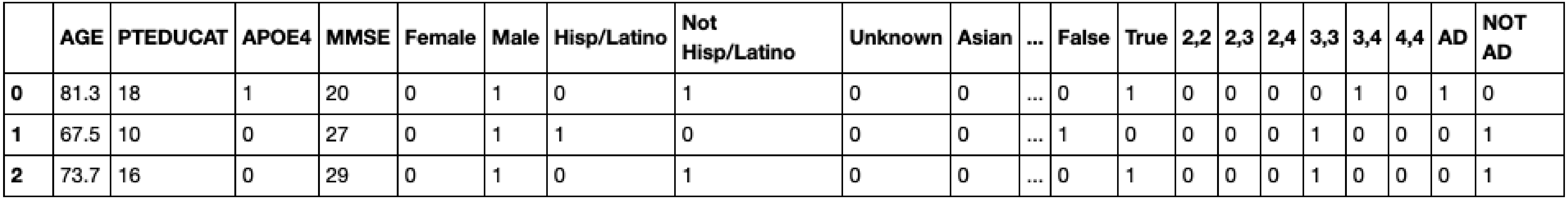


Image of the data:



	AGE	PTEDUCAT	APOE4	MMSE	Female	Male	Hisp/Latino	Not Hisp/Latino	Unknown	Asian	...	False	True	2,2	2,3	2,4	3,3	3,4	4,4	AD	NOT AD
0	81.3	18	1	20	0	1	0	1	0	0	...	0	1	0	0	0	0	1	0	1	0
1	67.5	10	0	27	0	1	1	0	0	0	...	1	0	0	0	0	1	0	0	0	1
2	73.7	16	0	29	0	1	0	1	0	0	...	0	1	0	0	0	1	0	0	0	1

This prototype goes beyond the original engineering goal of accuracy above 80%. The model had a classification accuracy of around 95%, which is far above the original goal. Additionally, the model had an AUC of 0.9. Area Under the Curve (AUC) is a metric that considers true positives and negatives and measures how well predictions are ranked, which makes it better determiner of the accuracy of a model.

Discussion

What do these results mean?

- These results mean a lot in the study of Alzheimer's Disease.
- Alzheimer's Disease is a disease that has no cure, but in 30-50 years that may no longer be the case. The only way to accurately develop a cure is with clinical trials and taking patients into the trials. A problem is that many people are classified with AD too late.
- These results show that it's possible to classify AD in patients and predict the future cognitive state (using MMSE scores) without using expensive data. Most of the models that have been developed can predict if a patient will develop AD (and some can even show the hypothetical track), but all the current models require data such as Magnetic resonance imaging (MRI), Computerized tomography (CT), and Positron emission tomography (PET).
- My model can classify if a person has AD with an accuracy of 94%. This is extremely useful as although it may not be 95-100%, it doesn't require any advanced technology. All the data that you input is easily accessible, and even the ones that you must pay for (APEO4 gene from 23andMe) can be ignored and the model will still be accurate.

Application and Challenges?

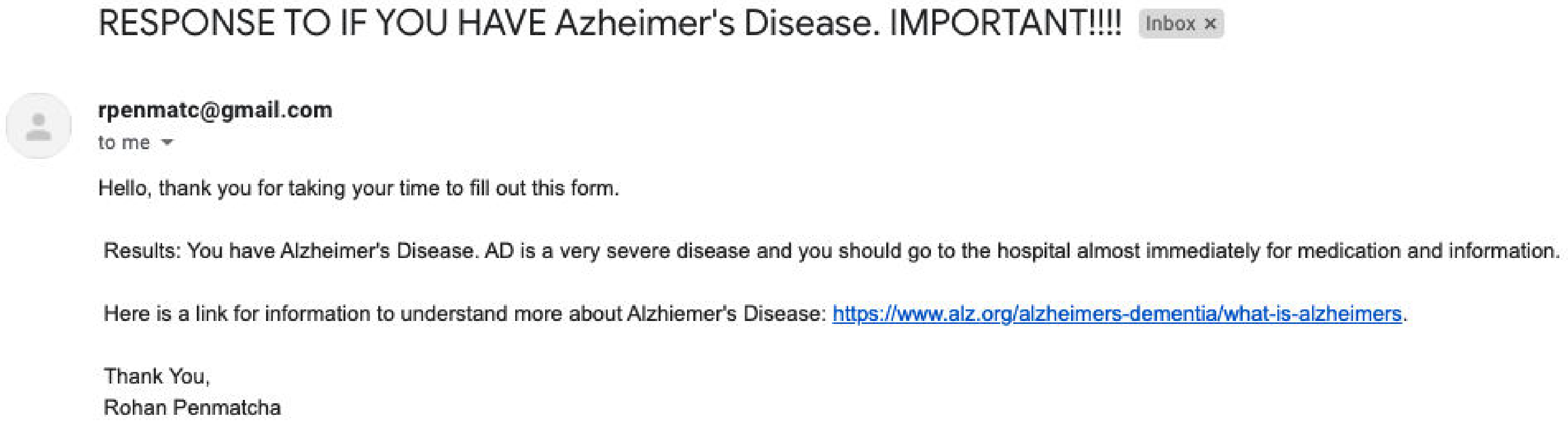
- My work has several applications that can help countless people and several challenges that can hinder that ability. One application that I wish to pursue is converting this script into an app, so that anyone can enter their data on their phone that they use everyday to see if they have AD and see their predicted decrease in MMSE scores.
- Additionally, this can specifically be applied in rural communities where hospitals are rare. A possible option includes going and offering free tests to collect data from people and run their data in the program to possibly save more lives.
- However, there also exists several challenges. This data is not easily accessible, and it will require countless more testing to determine if this accuracy is solely for this data or if it can be translated to practical use.
- Also, a problem with people submitting their private information is that they'll be worried about others using it for nefarious purposes.

Conclusions

Summary and Takeaways:

- When I envisioned this project in my mind, I hoped to reach an accuracy of around 80%, but I was pleasantly surprised when the model passed that benchmark for classification and prediction: 95% and 82%. This model is a success for those reasons, and it’s better than existing ones as it’s more accessible and requires less expensive data for its use.
- The next step is to improve the models by testing with more data. Either from collecting it myself or downloading more data and preprocessing it from the Alzheimer’s Disease Neuroimaging Initiative (ADNI).
- Specifically, I will need to investigate improving the MMSE model as an accuracy of 0.82 (with a margin of error of 1) can be improved.
- The real-world applications include developing this into a user-friendly app and making it possible for low-income and rural communities to get these tests.

- Sample Automated Email Response:



RESPONSE TO IF YOU HAVE Azheimer's Disease. IMPORTANT!!!!

rpenmatc@gmail.com

Hello, thank you for taking your time to fill out this form.

Results: You have Alzheimer's Disease. AD is a very severe disease and you should go to the hospital almost immediately for medication and information.

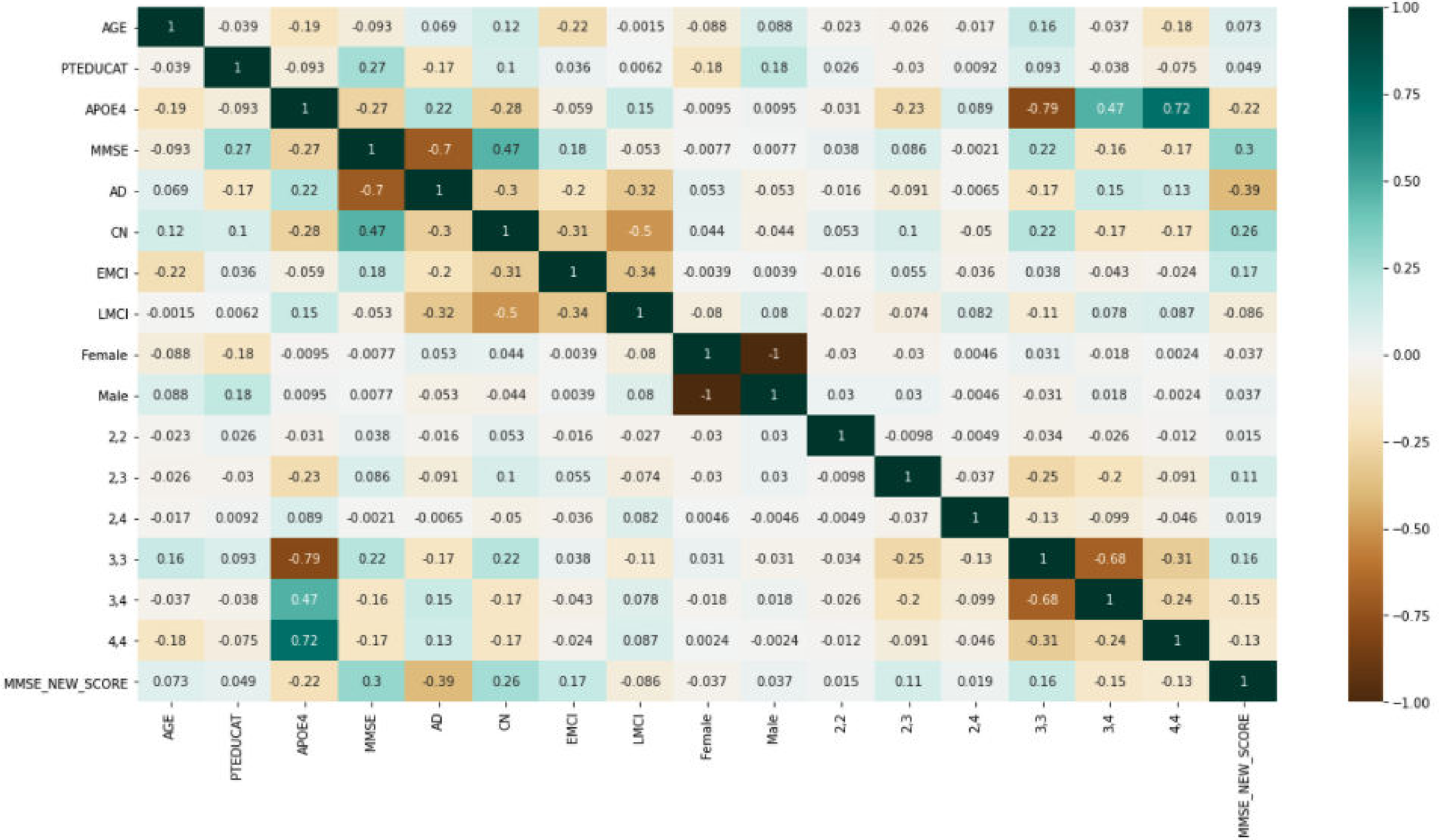
Here is a link for information to understand more about Alzhiemer's Disease: https://www.alz.org/alzheimers-dementia/what-is-alzheimers.

Thank You,
Rohan Penmatcha

References

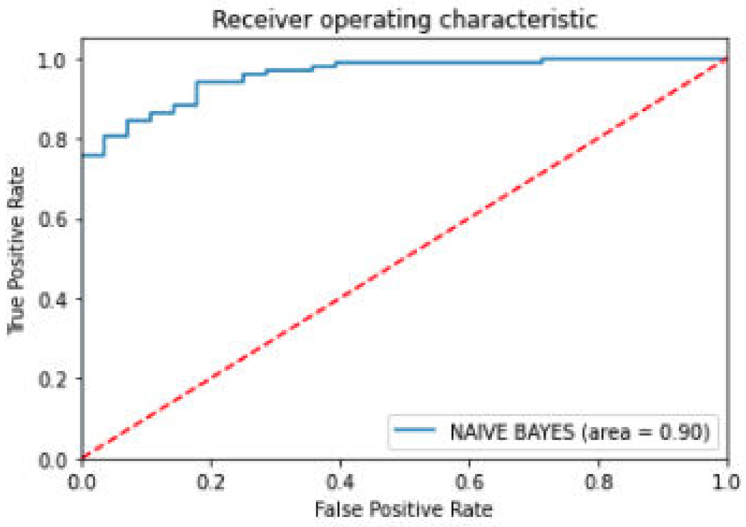
- M. A. Deture and D. W. Dickson, “The neuropathological diagnosis of Alzheimer’s disease,” Molecular Neurodegeneration, vol. 14, no. 1, 2019.
- “2020 Alzheimer’s disease facts and figures,” Alzheimer’s & Dementia, vol. 16, no. 3, pp. 391–460, 2020.
- A. Payan and G. Montana, “Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks,” arXiv.org, 09-Feb-2015.
- S. Esmaeilzadeh, D. I. Belivanis, K. M. Pohl, and E. Adeli, “End-To-End Alzheimer’s Disease Diagnosis and Biomarker Identification,” Machine Learning in Medical Imaging Lecture Notes in Computer Science, pp. 337–345, 2018.
- “The Need for Early Detection and Treatment in Alzheimer’s Disease,” EBioMedicine, vol. 9, pp. 1–2, 2016.

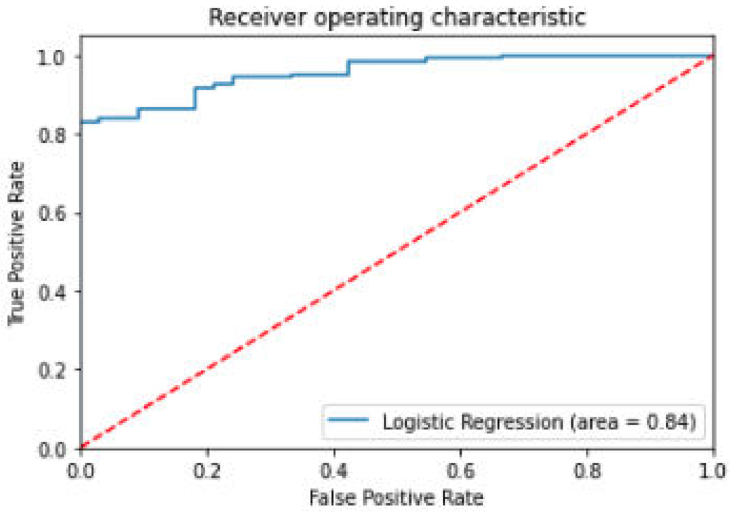
Virtual Lab Notebook Excerpt 1

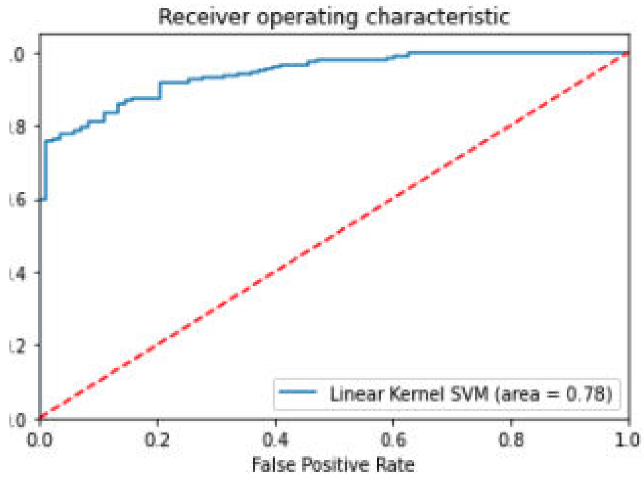


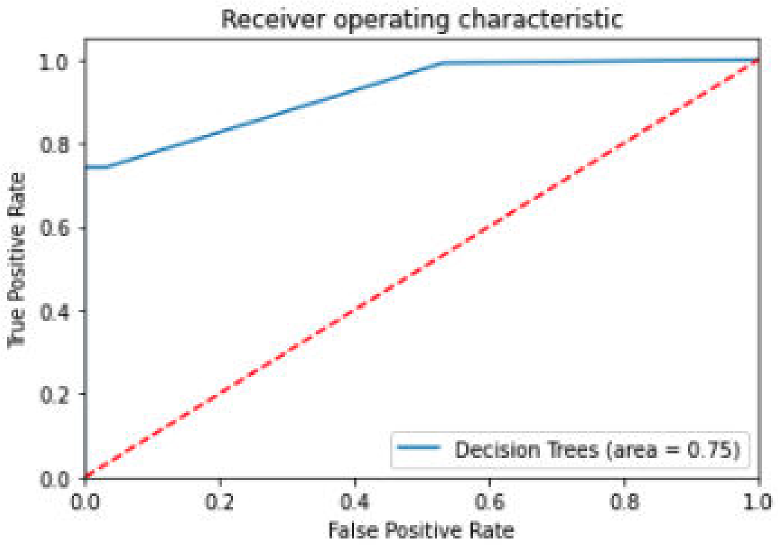
	AGE	PTEDUCAT	APOE4	MMSE	AD	CN	EMCI	LMCI	Female	Male	2,2	2,3	2,4	3,3	3,4	4,4	MMSE_NEW_SCORE
AGE	1	-0.039	-0.19	-0.093	0.069	0.12	-0.22	-0.0015	-0.088	0.088	-0.023	-0.026	-0.017	0.16	-0.037	-0.18	0.073
PTEDUCAT	-0.039	1	-0.093	0.27	-0.17	0.1	0.036	0.0062	-0.18	0.18	0.026	-0.03	0.0092	0.093	-0.038	-0.075	0.049
APOE4	-0.19	-0.093	1	-0.27	0.22	-0.28	-0.059	0.15	-0.0095	0.0095	-0.031	-0.23	0.089	-0.79	0.47	0.72	-0.22
MMSE	-0.093	0.27	-0.27	1	-0.7	0.47	0.18	-0.053	-0.0077	0.0077	0.038	0.086	-0.0021	0.22	-0.16	-0.17	0.3
AD	0.069	-0.17	0.22	-0.7	1	-0.3	-0.2	-0.32	0.053	-0.053	-0.016	-0.091	-0.0065	-0.17	0.15	0.13	-0.39
CN	0.12	0.1	-0.28	0.47	-0.3	1	-0.31	-0.5	0.044	-0.044	0.053	0.1	-0.05	0.22	-0.17	-0.17	0.26
EMCI	-0.22	0.036	-0.059	0.18	-0.2	-0.31	1	-0.34	-0.0039	0.0039	-0.016	0.055	-0.036	0.038	-0.043	-0.024	0.17
LMCI	-0.0015	0.0062	0.15	-0.053	-0.32	-0.5	-0.34	1	-0.08	0.08	-0.027	-0.074	0.082	-0.11	0.078	0.087	-0.086
Female	-0.088	-0.18	-0.0095	-0.0077	0.053	0.044	-0.0039	-0.08	1	-1	-0.03	-0.03	0.0046	0.031	-0.018	0.0024	-0.037
Male	0.088	0.18	0.0095	0.0077	-0.053	-0.044	0.0039	0.08	-1	1	0.03	0.03	-0.0046	-0.031	0.018	-0.0024	0.037
2,2	-0.023	0.026	-0.031	0.038	-0.016	0.053	-0.016	-0.027	-0.03	0.03	1	-0.0098	-0.0049	-0.034	-0.026	-0.012	0.015
2,3	-0.026	-0.03	-0.23	0.086	-0.091	0.1	0.055	-0.074	-0.03	0.03	-0.0098	1	-0.037	-0.25	-0.2	-0.091	0.11
2,4	-0.017	0.0092	0.089	-0.0021	-0.0065	-0.05	-0.036	0.082	0.0046	-0.0046	-0.0049	-0.037	1	-0.13	-0.099	-0.046	0.019
3,3	0.16	0.093	-0.79	0.22	-0.17	0.22	0.038	-0.11	0.031	-0.031	-0.034	-0.25	-0.13	1	-0.68	-0.31	0.16
3,4	-0.037	-0.038	0.47	-0.16	0.15	-0.17	-0.043	0.078	-0.018	0.018	-0.026	-0.2	-0.099	-0.68	1	-0.24	-0.15
4,4	-0.18	-0.075	0.72	-0.17	0.13	-0.17	-0.024	0.087	0.0024	-0.0024	-0.012	-0.091	-0.046	-0.31	-0.24	1	-0.13
MMSE_NEW_SCORE	0.073	0.049	-0.22	0.3	-0.39	0.26	0.17	-0.086	-0.037	0.037	0.015	0.11	0.019	0.16	-0.15	-0.13	1

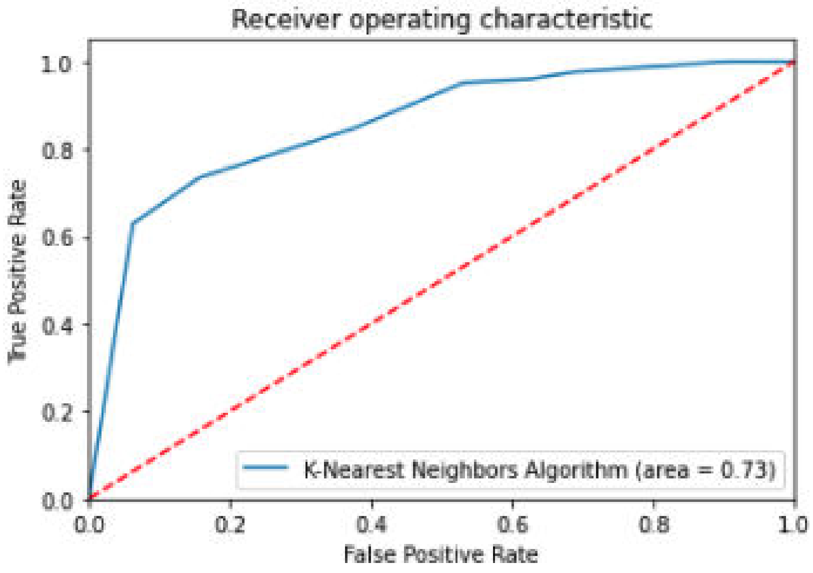
Virtual Lab Notebook Excerpt 2





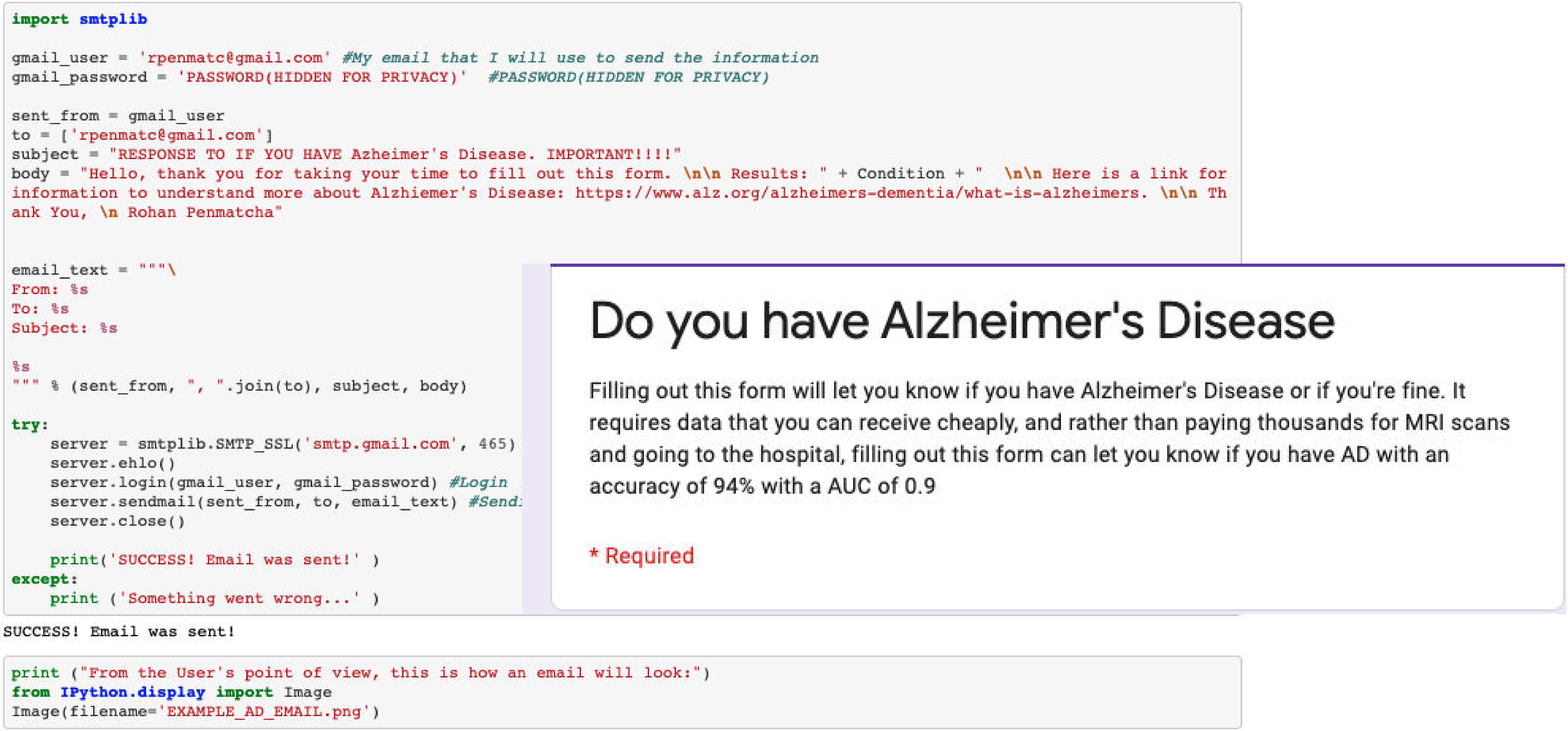






Naïve Bayes: 0.9
Logistic Regression: 0.84
Linear Kernel SVM: 0.78
Decision Teres: 0.75
K-Nearest Neighbors: 0.73

Virtual Lab Notebook Excerpt 3



```
import smtplib

gmail_user = 'rpenmatc@gmail.com' #My email that I will use to send the information
gmail_password = 'PASSWORD(HIDDEN FOR PRIVACY)'  #PASSWORD(HIDDEN FOR PRIVACY)

sent_from = gmail_user
to = ['rpenmatc@gmail.com']
subject = "RESPONSE TO IF YOU HAVE Azheimer's Disease. IMPORTANT!!!!"
body = "Hello, thank you for taking your time to fill out this form. \n\n Results: " + Condition + "  \n\n Here is a link for
information to understand more about Alzhiemer's Disease: https://www.alz.org/alzheimers-dementia/what-is-alzheimers. \n\n Th
ank You, \n Rohan Penmatcha"


email_text = """\
From: %s
To: %s
Subject: %s

%s
""" % (sent_from, ", ".join(to), subject, body)

try:
    server = smtplib.SMTP_SSL('smtp.gmail.com', 465)
    server.ehlo()
    server.login(gmail_user, gmail_password) #Login
    server.sendmail(sent_from, to, email_text) #Send
    server.close()

    print('SUCCESS! Email was sent!' )
except:
    print ('Something went wrong...' )
```

```
SUCCESS! Email was sent!
```

```
print ("From the User's point of view, this is how an email will look:")
from IPython.display import Image
Image(filename='EXAMPLE_AD_EMAIL.png')
```

Do you have Alzheimer's Disease

Filling out this form will let you know if you have Alzheimer's Disease or if you're fine. It requires data that you can receive cheaply, and rather than paying thousands for MRI scans and going to the hospital, filling out this form can let you know if you have AD with an accuracy of 94% with a AUC of 0.9

* Required