Alzheimer’s Disease Prediction

Tessa Clary
Applied Computer Science
University of Colorado
Boulder, CO
sccl8634@colorado.edu

Abstract:

Alzheimer's disease is a complex and devastating condition that impacts millions of people around the globe. Early detection is paramount for patient care as well as advancing our knowledge of the disease. The motivation for this project was to develop predictive models of age, gender and race that could help identify those who are at risk of cognitive decline. This was done through the usage of logistic regressions, decision tree classifiers, and naïve bayes models. All models resulted in accuracies of over 90%, establishing the validity of the results. As a result of these models, it was found that race can be a notable contributing factor to cognitive decline, but age and gender have negligble impacts.

Introduction:

As one of the most pressing healthcare challenges in modern medicince around the world, Alzheimer’s disease is difficult to predict accurately, leading to a fair amount of research and study on its potential predictive factors.

This study seeks to find and answer several different questions relating to potentially predictive demographic factors for Alzheimer's disease using the Healthy Aging dataset provided by the CDC. It contains a subset of BRFSS data. The primary focus of this study will be to use the metric of cognitive decline to accurately predict what factors among age, gender, and race may play an influential role in the early stage development of this disease. This research serves to unravel the difficult and intricate relationship between core demographics and their impact on cognitive decline in an aging population.

Related Work:

There is been a multitude of related works and studies on the predictive factors of Alzheimer's disease. Various domains of potential predictive factors have been studied, everything from genetics, to demographics, to lifestyle, and other measures of cognitive health. Ultimately, previous findings for predictive factors such as demographics have been inconclusive, as their exact mechanisms of interaction are poorly understood and are still the subject of ongoing research. Further attempts to understand these predictive factors are crucial for early dectection and preventive care for those at risk.

Literature Survey:

“Multivariate Logistic Regression Analysis of Complex Survey Data with Application to BRFSS Data”
This study uses a multivariate logistic regression analysis to study the association of risk factors for heart attacks and strokes. This closely aligns with the goals of this project albeit with a different focus. The multivariate regression used here focuses on the study of risk factors for heart attack and stroke with a particular focus on BRFSS data from 2009. [1]

"A smart Alzheimer’s patient monitoring system with IoT-assisted technology through enhanced deep learning approach"
This study utilized the BRFSS dataset and IoT devices & deep learning techniques to help predict Alzheimer's disease. With the high accuracy of the results of this study of over 98%,

the results of this study imply that reasonable predictions can be made on this dataset with good accuracy. [2]

"Population measures of subjective cognitive decline: A means of advancing public health policy to address cognitive health"
This study focuses on SCD, or subjective cognitive decline. It is an early cognitive indicator that complements this project given that it includes an early-stage assessment parameter. Given the issues with studying Alzheimer's Disease via population-based measures, this focuses on identifying a specific subset of cognitive decline. [3]

"Racial and Ethnic Differences in Subjective Cognitive Decline — United States, 2015–2020"
This report from the CDC utilizes the same BRFSS study to find the minorities with the highest reported prevalence of sudden cognitive decline. Their analysis was conducted using SAS-callable SUDANN; while the dataset and focus of this report is similar to this project, the analysis performed is not necessarily close to what this project has accomplished. [4]

Data Set:

The Behavioral Risk Factor Surveillance System (BRFSS) is a nationwide survey in the United States led by the Center for Disease Control and Prevention (CDC). It collects data on health related risk behaviors, chronic conditions, and preventive service usage among adults. It is one of the largest ongoing telephone health surveys globally since 1984. It typically gathers information on around 400,000 Americans every year. This derivative dataset in particular was published by the CDC with a focus on Alzheimer's Disease and Healthy Aging Data. This data set initially contained 39 different attributes. After initial exploratory analysis of the dataset, 29 of the attributes were found to not be relevant to the analysis of this project and were subsequently dropped. Below is the list of all of the original attributes:

RowId: Dropped
This attribute concatentates several different other attributes: Datasource, YearStart, YearEnd, LocationID, QuestionID, TopicID, StratificationCategoryID1, and StratificationCategoryID2. Given that all of these attributes are already present in the dataset, this attribute does not contribute to the study.

YearStart: Dropped
The focus of this study was to observe and relate potential demographic factors; given that this is not a study over time this attribute was dropped from the cleaned dataset.

YearEnd: Dropped
This attribute was additionally dropped from the dataset for the same reason as YearStart.

LocationAbbr: Dropped
This is an abbreviation of the location description, entirely unneeded for the scope of this study.

LocationDesc: Dropped
As this project is focusing more specifically on the predictive capability of age, gender, and race, location was dropped as an unused attribute.

Datasource: Dropped
The only value present for this attribute is "BRFSS", as it is the only source of data in this dataset.

Class:
This is the topic class for the question asked to the person being interviewed. Of note is the class category of cognitive decline. This class was reimplemented as a binary dummy variable wherein a 1 denotes that the topic class is cognitive decline, and 0 denotes any other topic class. As a result this is the dependent variable

that the models in this study will attempt to predict.

Topic:
This goes one step further than the class attribute, further drilling down the types of questions asked to the interviewee. Given the prevelance of cognitive decline in the prediction of Alzheimer's disease, there is a degree of lessened merit to use these subcategories of questions to act as the target variable. By decreasing the size of the dataset relating to the target, the models resultingly have less to work with in assessing their predictive power of the dependent variable.

Question:
This is the specific question asked of the interviewer during the interview. As this is even more specific than the usage of topic or class, this attribute was not used for predictive purposes but allowed to remain in the dataset to ensure that there is not to much loss in information.

Response: Dropped
This was an entirely blank attribute.

Data_Value_Unit: Dropped
This denoted if the attribute Data_Value was a percentage or mean, as this study uses dummy variables, all of these related attributes were subsequently dropped.

DataValueTypeID: Dropped
This served the same purpose as Data_Value_Unit, albiet abbreviated as words.

Data_Value_Type: Dropped
This served the same purpose as Data_Value_Unit, albiet unabbreviated.

Data_Value: Dropped
Given the difficulty of converting these relative percentages into a format of data that can be utilized by the models used in this study, this column was dropped due to irrelevancy despite its importance in the dataset.

Data_Value_Alt: Dropped
This was a duplicate of Data_value and as such served no purpose.

Data_Value_Footnote_Symbol: Dropped
This attribute was blank.

Data_Value_Footnote: Dropped
This attribute was blank.

Low_Confidence_Limit: Dropped
Dropped due to lack of relevance to the model study.

High_Confidence_Limit: Dropped
Dropped due to lack of relevance to the model study.

Sample_Size: Dropped
This attribute was blank.

StratificationCategory1: Dropped
This attribute only contained one value of “Age Group” and was subsequently dropped from the cleaned dataset.

Stratification1:
This attribute split between [50-64], [65+], and Overall. The dataset was split on this attribute, with rows including [50-64] and [65+] placed into a separate derivative dataset focusing only on age.

StratificationCategory2: Dropped
This attribute had two possible values, Gender & Race/Ethnicity. This was ultimately dropped due to redundancy.

Stratification2:
This attribute was used to create dummy variables for the independent variables of the study on race. Of the possible values related to the race models, the following were used: Asian/Pacific Islander, Black, non-Hispanic, Hispanic, Native Am/Alaskan Native, and White, non-Hispanic.

StratificationCategory3: Dropped
This attribute was blank.

Stratification3: Dropped
This attribute was blank.

Geolocation: Dropped
Location coordinate attributes were dropped due to irrelevancy.

ClassID: Dropped
Gives a code to represent the different question classes covered. Dropped due to redundancy.

TopicID: Dropped
Gives a code to represent the different question topics covered. Dropped due to redundancy.

QuestionID: Dropped
Gives a code to represent the different questions covered. Dropped due to redundancy.

ResponseID: Dropped
This attribute was blank.

LocationID: Dropped
Gives a number ranging from 1 to 4 digits. Dropped due to irrelevancy.

StratificationCategoryID1:
The dataset was split here, with the subsequent derivative dataset's rows splitting on [50-64] and [65+].

StratificationID1:
Breaks down the age attribute into 3 categories: [50-64], [65+], and AGE_OVERALL. The derivative age dataset drops rows containing AGE_OVERALL due to irrelevancy, but uses [50-64] and [65+] as the basis for the independent variables.

StratificationCategoryID2:
Contains data values of RACE, GENDER, and OVERALL. The dataset was split here to allow for derivative datasets based around race and gender. Rows containing OVERALL were discarded due to irrelevancy.

StratificationID2:
This attribute contains the gender and race of the interviewee. This column was used for both race and gender datasets to help label each row with the relevant demographic data.

StratificationCategoryID3: Dropped
This attribute was blank.

StratificationID3: Dropped
This attribute was blank.

Report: Dropped
This attribute was blank.

Main Techniques Applied:

This project utilized several different techniques from Python machine learning libraries to investigate the impact of different types of demographics on cognitive decline.

Data preprocessing began with the usage of Pandas to drop irrelevant attributes and rows with missing values. The original dataset was split into three different derivative datasets as not every row had the relevant demographics for all three demographic focuses. The subsequent derivative datasets then had dummy variables implemented for their respective independent variables as well as the common dependent variable, cognitive decline as a form of feature engineering.

Logistic regressions were employed for each of the main demographic variable studies. These involved encoding categorical variables into dummy variables to allow the data to fit into a logistic regression model. The subsequent coefficient results helped determine the impact of specific demographic variables on cognitive decline. Accuracy measures as well as area under the ROC curve were used to evaluate the predictive power of these models.

Decision tree classifiers were employed for each of the main demographic variable studies. These involved creating a tree like structure where

nodes indicate demographic catagories, and branches denoting patterns leading to cognitive decline. The visualization of the decision trees help aid the understanding of how the different values contribute to predictions. Accuracy measures, feature importance, and area under the ROC curve were used to evaluate the predictive power of these models.

Gaussian Naïve Bayes was also used for each of the main demographic variable studies. These used Bayes's theorem with an assumption of independence between features. It then calculated the probability of cognitive decline for each category using Gaussian distributions. Accuracy measures, precision/recall curves, and area under the ROC curve were used to evaluate the predictive power of these models.

Key Results:

Three different models were used for each demographic focus: logistic regression, decision tree classifier, and naïve bayes.

The models for age did not yield any particularly strong or notable results. The logistic regression had a 92.3% accuracy, with a feature coefficient of -0.067 for age bracket [50-64], and a coefficient of 0.034 for age bracket [65+]. The decision tree classifier showed a feature importance of 0.878 for age bracket [50-64], and 0.121 for age bracket [65+] and the same accuracy. Lastly, the naïve bayes model also had a 92.3% accuracy, with the precision dropping rapidly after the recall increases to around 0.32. With an AUC of 0.51 these models are no better than a random classifier.

The models for gender additionally did not yield any particularly strong or notable results. The logistic regression had a 91.5% accuracy, with a feature coefficient of -0.031 for male, and a coefficient of 0.031 for female. The decision tree classifier showed a feature importance of 0 for male, and 1 for female+] and the same accuracy. Lastly, the naïve bayes model also had a 91.5% accuracy, with the precision dropping rapidly after the recall increases to around 0.52. With an AUC of 0.51 these models are still no better than a random classifier.

The models for race did yield particularly notable results. The logistic regression had a 93.5% accuracy, with feature coefficients of 0.334 for Asian/Pacific Islander, -0.001 for Native American/Alaskan Native, -0.285 for Black non-Hispanic, -0.339 for Hispanic, and 0.290 for White non-Hispanic. The decision tree classifier had a 93.5% accuracy, with a feature importance of 0 for male, and 1 for female. Lastly, the naïve bayes model had a 93.5% accuracy, with the precision dropping rapidly after the recall increases to around 0.1. With an AUC of 0.57 this model does have merit compared to the earlier demographics, with the strength of the coefficients backing it up. Of immediate note from these results are the positive coefficients for Asian/Pacific Islander and White non-Hispanic races. Hispanic and Black non-Hispanic also notably have the strongest negative coefficients.

The results of the models on age as a predictive factor for cognitive decline as an early sign of Alzheimer's did not produce the expected outcomes. Historically, age has been a notable predictive factor for Alzheimers disease, so to find a complete lack of relationship in these models is rather tellling.

Additonally, the results of the models on gender as a predictive factor for cognitive decline as an early sign of Alzheimer's also did not produce the expected outcomes. Historically, gender has also been a notable predictive factor for Alzheimers disease, so to find a second major factor to have no relevance in these models brings into question the validity of the dataset.

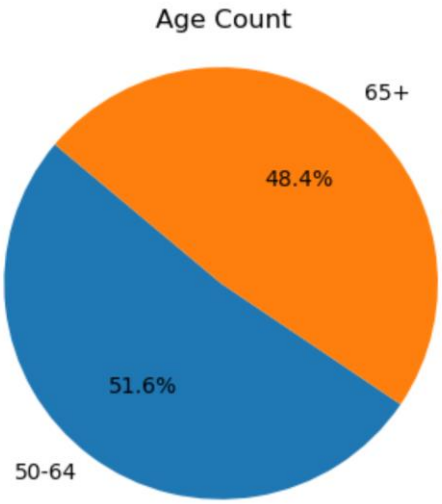
Lastly out of the three demographic focuses, race had a notable influential factor in predicting

cognitive decline. This is also a well known predictive factor for Alzheimer's, albeit the relationship is poorly understood. The fact that the models on race produced results with better prediction capability than a random classifier puts weight on the likelihood that race is an influential factor.

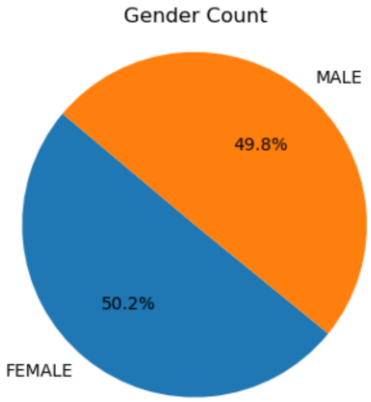
The key takeaways from these results however are not necessarily their end result predictive power, but the inability for two out of the three demographic focuses to meet the expected predictive thresholds commonly established in scientific circles. In a vacuum, if only one demographic was focused on instead, (age, gender, race), it would have been seen that there is some predictive power (given the focus was on race), but that age and gender have no relationship to cognitive decline. The validity of the results for the models on race in comparison to the lack of validity on age and gender imply that the data mining may have been performed correctly, but that the dataset may not be accurate or representative of the populations that they are trying to study.

Below are the counts and relative percentages of the independent variables being studied:

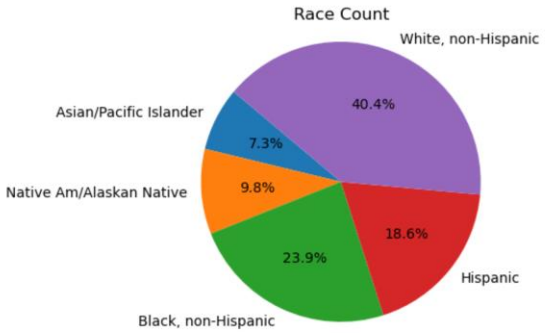
Age:



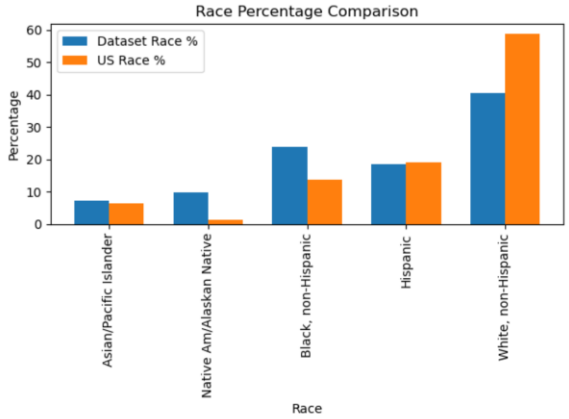
Gender:



Race:



While the percentages of age and gender are relatively equal, the percentages of the races interviewed in the dataset did not seem to fit very well. An additional examination of these counts in comparison to percentages of race in the US population [5] reveals that the dataset supplied by the CDC may not be as accurate as expected:



As evidenced above, it can be seen that white is under-represented in the dataset, with black and Native American/Alaskan Native both over-represented. Given that white is the highest in terms of feature importance and coefficients, their under-representation in the dataset may be showing issues with the data collection methods being used. Conversely, Native American/Alaskan Native, black, and Hispanic have the lowest feature importance and coefficients, but Hispanic is unlike the other two in that it is not over-represented in the dataset. The fact that the coefficient for Hispanic sits between black and Native American/Alaskan Native despite not being over-represented implies that the issues with the data collection methodology may be multifaceted and difficult to address. If all minority percentages were skewed in a similar direction, the issue would likely be much clearer. However, the fact that some minority population representations are skewed more than others on top of their expected predictive factors shows that there may be more than one major issue with the data collection.

Applications:

Race was found to have a notable influence on the prediction of cognitive decline in this study. This is already a well known predictive factor, but to see its significance in comparison to age or gender, it belies a potential genetic risk component to Alzheimer's disease. Future genetic analysis of populations may be able to identify and flag these genetic risk factors disproportionately impacting specific demographics as a result.

This study found limited evidence to suggest that either age or gender have a significant influence on cognitive decline predicition. While this may not necessarily be the actual case, it does however bring into focus the quality of the dataset being used for analysis. An inability to prove any kind of a relationship between such integral factors as cognitive decline and age/gender calls into question the validity of the dataset itself. If the dataset utilizes outdated data collection methodology, then it may not be representative of the populations that it is analysing and attempting to draw important and insightful information from. This lack of model accuracy may imply that the data collection methodology should be updated to better represent the actual populations that they are intending to represent.

The most important applications of this study is not actually the predictive capability of these models, but instead the spotlight it has drawn on the potential shortcomings of a landline based telephone survey running from 2015 to 2021. While data mining is very powerful in being able to help create accurate predictive models, it can also serve to help identify issues with the information being used to create said models. The results of this study have concluded that while race may be a predictive factor in cognitive decline due to Alzheimer's Disease, the CDC should reconsider their methodology in regards to gathering data on their aging population.

Conclusion:

Through the usage of data mining and machine learning techniques, demographic data was analysed in relation to cognitive decline in an attempt to predict early factors of Alzheimer's Disease. This study revealed a several links between race and cognitive decline in an aging population. It also however revealed potential issues with the dataset and how it is currently designed as it failed to predict other demographic factors with known influences- age and gender. While some insights may be drawn from this study in regards to the influence of race on cognitive decline, it has also underscored the need for a reexamination of the data being collected and the methodologies being used.

References:

[1] Lu, M., & Yang, W. (n.d.). *Multivariate Logistic Regression Analysis of Complex Survey Data with Application to BRFSS Data.* Journal of Data Science. Retrieved December 7, 2023, from https://jds-online.org/journal/JDS/article/1213/info

[2] Arunachalam, R., Sunitha, G., Shukla, S., Nath Pandey, S., Urooj, S., & Rawat, S. (2023, August 9). *A smart Alzheimer's patient monitoring system with IoT-assisted technology through enhanced deep learning approach.* Springer Link. Retrieved December 7, 2023, from https://link.springer.com/article/10.1007/s10115-023-01890-x

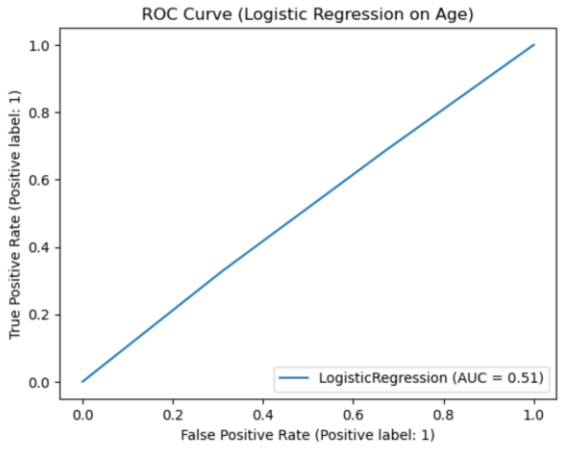
[3] Olivari, B., Baumgart, M., Taylor, C., & McGuire, L. (2021, February 18). *Population measures of subjective cognitive decline: A means of advancing public health policy to address cognitive health.* Alzheimer's Association. Retrieved December 7, 2023, from https://alz-journals.onlinelibrary.wiley.com/doi/full/10.1002/trc2.12142

[4] CDC. (2023, March 10). *Racial and Ethnic Differences in Subjective Cognitive Decline — United States*, 2015–2020. CDC.gov. Retrieved December 7, 2023, from https://www.cdc.gov/mmwr/volumes/72/wr/mm7210a1.htm

[5] U.S. Census Bureau. *U.S. Census Bureau QuickFacts: United States.* Census.gov. Retrieved December 7, 2023, from https://www.census.gov/quickfacts/fact/table/US/PST045222

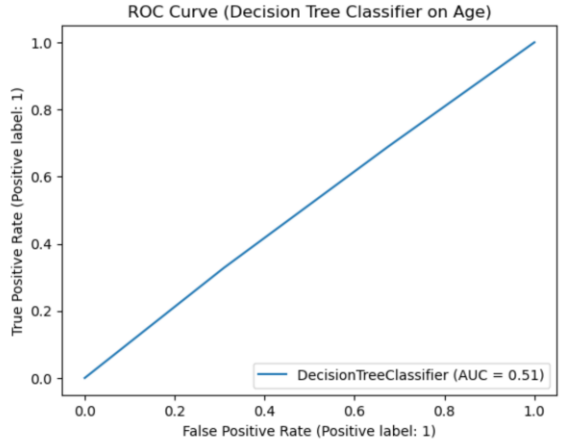
Age Models: Logistic Regression:

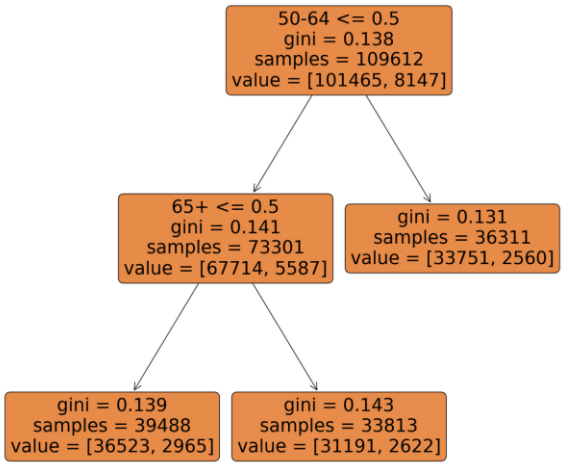
```
Accuracy: 0.9239891986571304
  Feature  Coefficient
0   50-64    -0.067897
1     65+     0.034872
```



Age Models: Decision Tree Classifier:

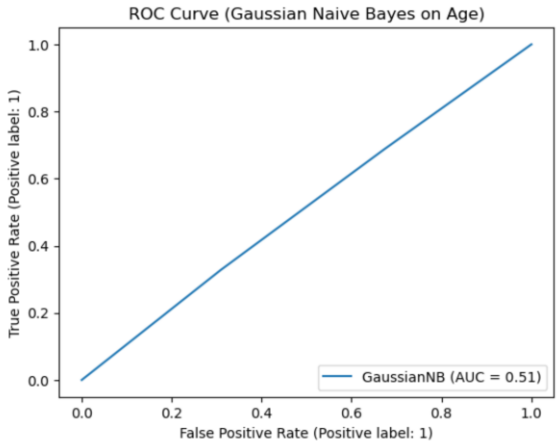
```
Accuracy: 0.9239891986571304
  Feature  Importance
0   50-64     0.87825
1     65+     0.12175
```

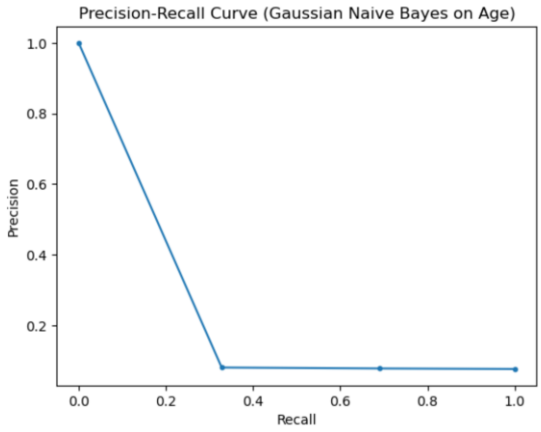




Age Models: Naïve Bayes:

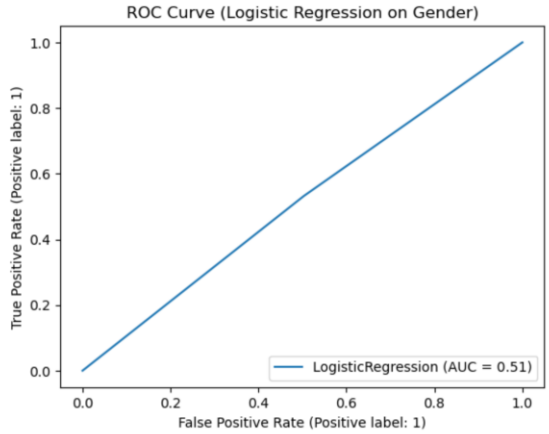
```
Accuracy: 0.9239891986571304
```





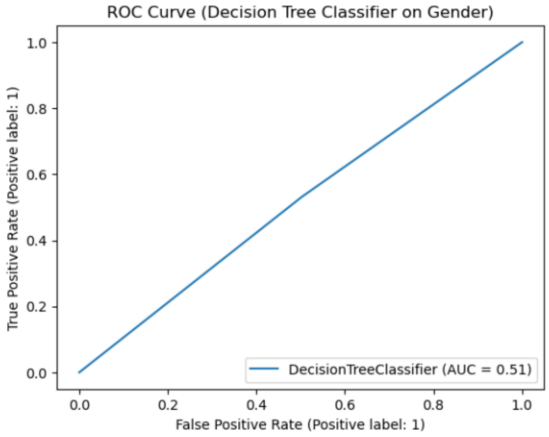
Gender Models: Logistic Regression:

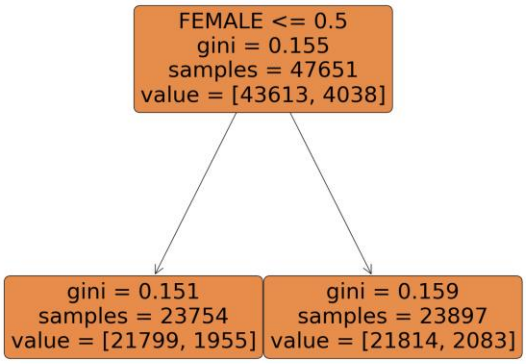
```
Accuracy: 0.9158062620666498
  Feature  Coefficient
0    MALE    -0.031350
1  FEMALE     0.031347
```



Gender Models: Decision Tree Classifier:

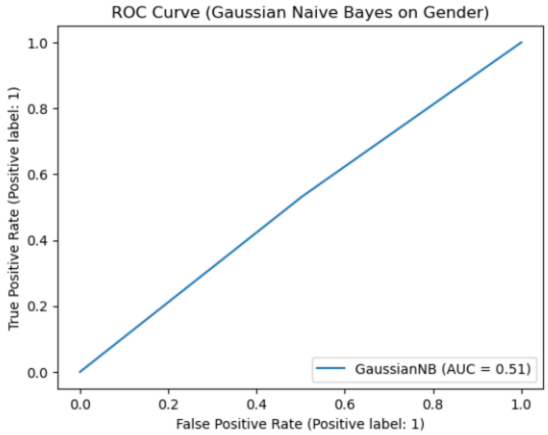
```
Accuracy: 0.9158062620666498
  Feature  Importance
0    MALE         0.0
1  FEMALE         1.0
```

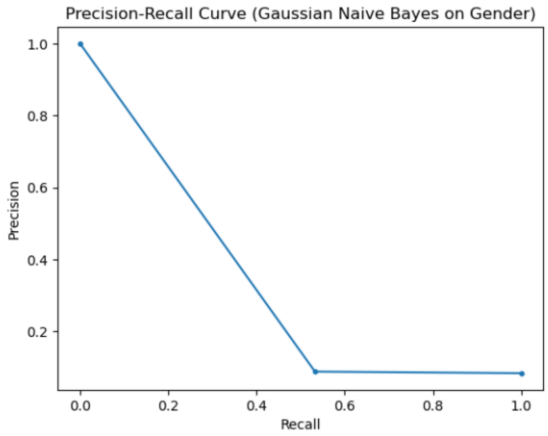




Gender Models: Naïve Bayes:

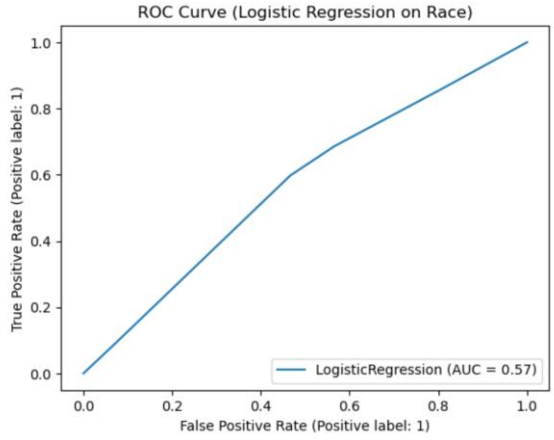
```
Accuracy: 0.9158062620666498
```





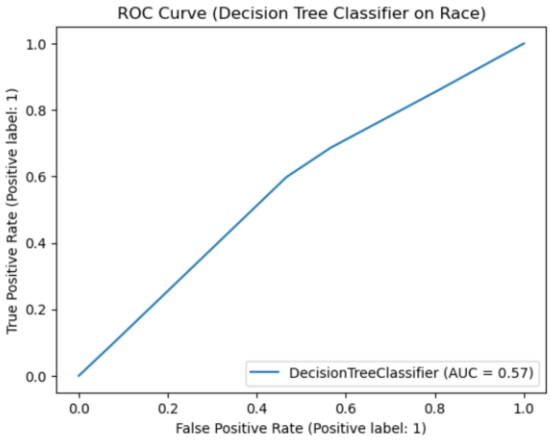
Race Models: Logistic Regression:

```
Accuracy: 0.9351881737783229
                      Feature  Coefficient
0    Asian/Pacific Islander     0.334463
1  Native Am/Alaskan Native    -0.001227
2       Black, non-Hispanic    -0.284679
3                  Hispanic    -0.338887
4       White, non-Hispanic     0.290339
```



Race Models: Decision Tree Classifier:

```
Accuracy: 0.9351881737783229
                      Feature  Importance
0    Asian/Pacific Islander    0.283279
1  Native Am/Alaskan Native    0.068637
2       Black, non-Hispanic    0.000000
3                  Hispanic    0.002119
4       White, non-Hispanic    0.645965
```



Race Models: Naïve Bayes:

```
Accuracy: 0.9351881737783229
```