DTSA 5509 - Introduction to Machine Learning Final Project: Supervised Learning with Alzheimer's Disease Datasets

Israel Johnson

June 25, 2023

University of Colorado **Boulder**

PROBLEM STATEMENT

- There has been ongoing research in the area of developing artificial intelligence (AI) and machine learning (ML) applications for Alzheimer's disease and dementia. These applications range from detecting early stages of Alzheimer's disease (and dementia) to measuring the impact of the disease in the presence of other diseases.

- Researchers are still exploring different ML model approaches for analyzing datasets related to Alzheimer's disease and dementia, seeking which are the best models to apply in given scenarios related to the disease.

PROJECT OBJECTIVE

- The goal for this project is to apply model analysis for two different scenarios, where both scenarios relate to datasets about Alzheimer's disease and dementia
- The two scenarios are as follows:
    - Determine if Alzheimer's disease is an effective feature with data related to the COVID19 pandemic, when compared to other disease features that are present. A multilinear regression model is used for model analysis in this scenario.
    - Determine if whether or not there is a strong correlation between age and gender features among patients who have Alzheimer's disease or dementia. Also, in either case of correlation conditions, find ways to improve model accuracy and performance, if necessary. For the model approach for this case, a logistic regression model is applied to the corresponding data.

DATA SOURCES

1.) U.S Department of Health and Human Services, Centers for Disease Control and Prevention (2021). Monthly Counts of Deaths By Select Causes (2020-2021) [Data set]. Retrieved from https://catalog.data.gov/dataset/monthly-counts-of-deaths-by-select-causes-2020-2021-2785a

2.) U.S Department of Health and Human Services, Centers for Disease Control and Prevention (2021). Conditions Contributing to Deaths Involving Coronavirus Disease 2019 (COVID-19) by Age Group [Data set]. Retrieved from https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group

3.) Open Access Series of Imaging Studies (OASIS) Brains Project. OASIS: Cross-sectional MRI Data in Young, Middle-Aged, Nondemented, and Demented Older Adults (2007) [Data set]. Retrieved from https://www.oasis-brains.org/#oasis1

4.) Open Access Series of Imaging Studies (OASIS) Brains Project. OASIS: Longitudinal MRI Data in Nondemented and Demented Older Adults (2010) [Data set]. Retrieved from https://www.oasis-brains.org/#oasis2

5.) Dinner, Baris (2022). Alzheimer Features (2022) [Data set]. Retrieved from https://www.kaggle.com/datasets/brsdincer/alzheimer-features

University of Colorado **Boulder**

GITHUB REPOSITORY SOURCE

- **Source:** https://github.com/IsraelsLibrary/DTSA_5509_Intro_to_Machine_Learning

DATA CLEANING PROCESS

- For data in scenario one, various column features were found to contain null values. A four percent null value threshold was established to limit the number of null values that were permitted. Any column features that exceeded the given threshold were removed from the dataset.

- In preparation for scenario two, two ‘helper’ functions were created to drop unrelevant columns from associated data as well as remove any row features that contained null values. Unlike scenario one where data cleaning was done for one data source, scenario two involved four different sources, which required more effort when it came to the data cleaning process. The data cleaning process for scenario two occurs after further data transformations are made (extracting key features from the different data sources and combining them into one dataset).

DATA CLEANING PROCESS

```
threshold = int(0.04 * len(monthly_provisional_data))

features_to_impute = [col for col in monthly_provisional_data.columns if
                      monthly_provisional_data[col].isnull().sum() != 0 and
                      monthly_provisional_data[col].isnull().sum() <= threshold]
features_to_throw = [col for col in monthly_provisional_data.columns if
                     monthly_provisional_data[col].isnull().sum() > threshold]

old_monthly_provisional_data = monthly_provisional_data

monthly_provisional_data.drop(columns=features_to_throw, axis=1, inplace=True)
```

University of Colorado **Boulder**

DATA CLEANING PROCESS

```python
# 'Helper' functions that perform data cleaning for the datasets in scenarios one and two.

def drop_columns(columns_to_drop, dataset):
    for col in columns_to_drop:
        if col in dataset.columns:
            dataset = dataset.drop(columns=col, axis=1)
    return dataset

def clean_data(dataset):
    dataset = dataset.dropna()
    if dataset.isnull().sum().sum() > 0:
        clean_data(dataset)
```

University of Colorado **Boulder**

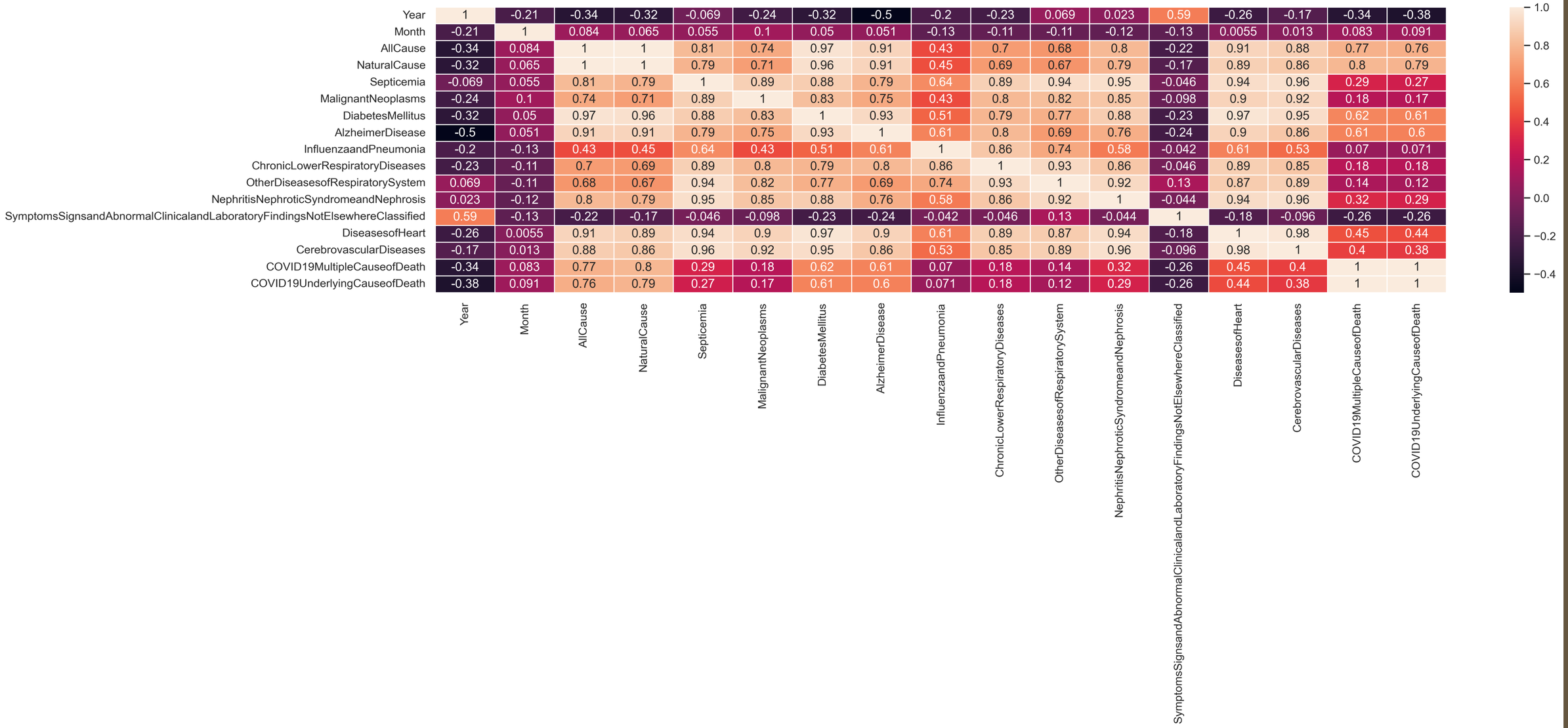
DATA CLEANING PROCESS

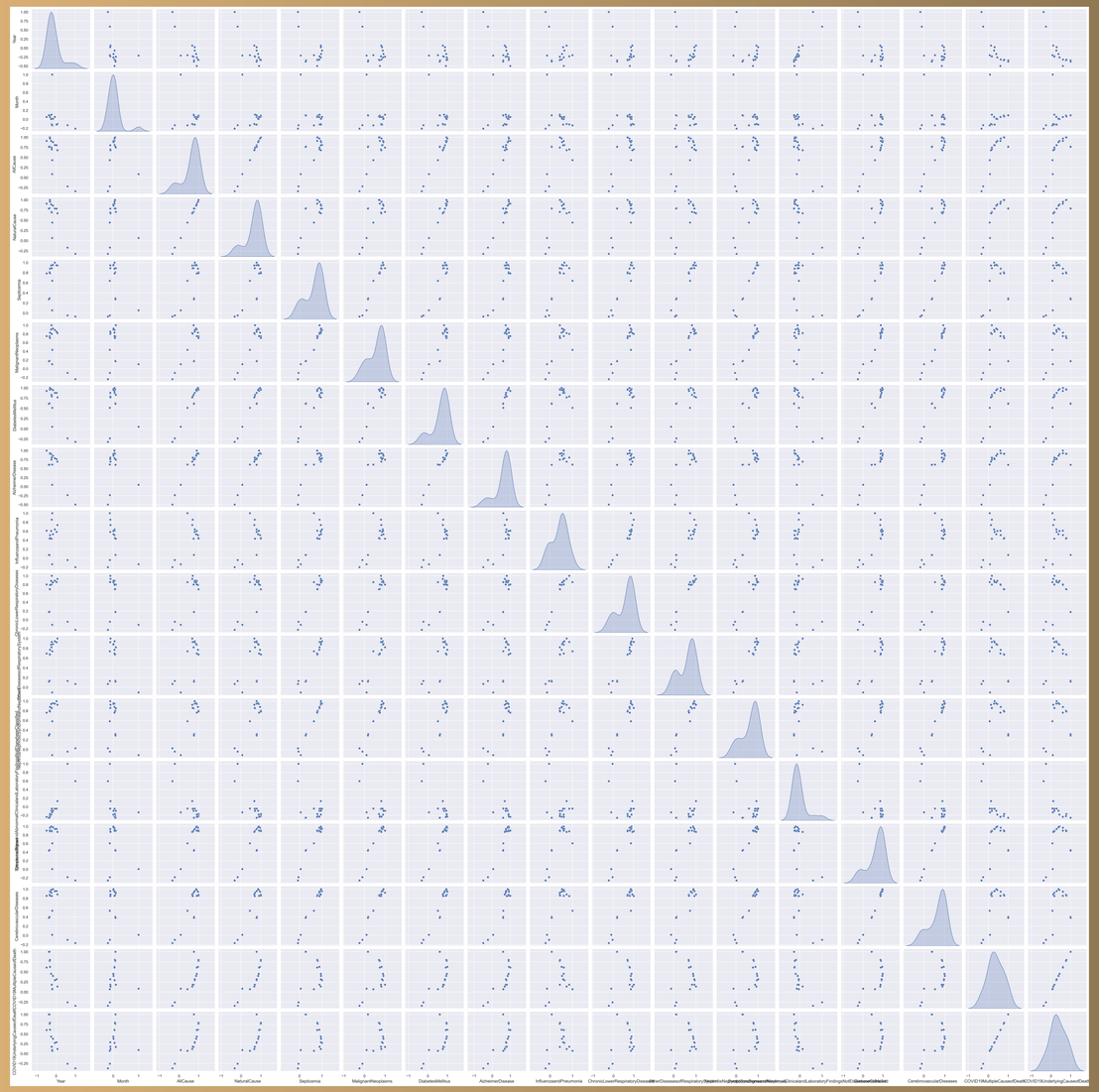
```python
# 'Helper' function that forms the final version of the dataset for scenario 2.
def transform_dataset2(datasets, limit):
    new_dataset = {}

    new_dataset['age'] = []
    new_dataset['sex'] = []

    for dataset in datasets.values():
        new_dataset['age'].extend(dataset[col].tolist()[:limit] for col in dataset if col in 'Age')
        new_dataset['sex'].extend(dataset[col].tolist()[:limit] for col in dataset if col in 'M/F')

    all_ages = []
    genders = []
    group = []

    for age_,sex_ in zip(new_dataset['age'], new_dataset['sex']):
        all_ages += age_
        genders += sex_

    for ind in range(len(genders)):
        if genders[ind] == 'M':
            genders[ind] = 0
        else:
            genders[ind] = 1

    new_dataset['age'] = all_ages
    new_dataset['sex'] = genders

    return pd.DataFrame.from_dict(new_dataset)

```

University of Colorado **Boulder**

EXPLORATORY DATA ANALYSIS (SCENARIO ONE)

- The first stage of exploratory data analysis (EDA) in scenario one was to generate a correlation matrix to see where strong correlation and multicollinearity can be found.
- Along with a correlation matrix, a series of pair plots were also created to explore correlation and collinearity among features


Year 1 -0.21 -0.34 -0.32 -0.069 -0.24 -0.32 -0.5 -0.2 -0.23 0.069 0.023 0.59 -0.26 -0.17 -0.34 -0.38
Month -0.21 1 0.084 0.065 0.055 0.1 0.05 0.051 -0.13 -0.11 -0.11 -0.12 -0.13 0.0055 0.013 0.083 0.091
AllCause -0.34 0.084 1 1 0.81 0.74 0.97 0.91 0.43 0.7 0.68 0.8 -0.22 0.91 0.88 0.77 0.76
NaturalCause -0.32 0.065 1 1 0.79 0.71 0.96 0.91 0.45 0.69 0.67 0.79 -0.17 0.89 0.86 0.8 0.79
Septicemia -0.069 0.055 0.81 0.79 1 0.89 0.88 0.79 0.64 0.89 0.94 0.95 -0.046 0.94 0.96 0.29 0.27
MalignantNeoplasms -0.24 0.1 0.74 0.71 0.89 1 0.83 0.75 0.43 0.8 0.82 0.85 -0.098 0.9 0.92 0.18 0.17
DiabetesMellitus -0.32 0.05 0.97 0.96 0.88 0.83 1 0.93 0.51 0.79 0.77 0.88 -0.23 0.97 0.95 0.62 0.61
AlzheimerDisease -0.5 0.051 0.91 0.91 0.79 0.75 0.93 1 0.61 0.8 0.69 0.76 -0.24 0.9 0.86 0.61 0.6
InfluenzaandPneumonia -0.2 -0.13 0.43 0.45 0.64 0.43 0.51 0.61 1 0.86 0.74 0.58 -0.042 0.61 0.53 0.07 0.071
ChronicLowerRespiratoryDiseases -0.23 -0.11 0.7 0.69 0.89 0.8 0.79 0.8 0.86 1 0.93 0.86 -0.046 0.89 0.85 0.18 0.18
OtherDiseasesofRespiratorySystem 0.069 -0.11 0.68 0.67 0.94 0.82 0.77 0.69 0.74 0.93 1 0.92 0.13 0.87 0.89 0.14 0.12
NephritisNephroticSyndromeandNephrosis 0.023 -0.12 0.8 0.79 0.95 0.85 0.88 0.76 0.58 0.86 0.92 1 -0.044 0.94 0.96 0.32 0.29
SymptomsSignsandAbnormalClinicalandLaboratoryFindingsNotElsewhereClassified 0.59 -0.13 -0.22 -0.17 -0.046 -0.098 -0.23 -0.24 -0.042 -0.046 0.13 -0.044 1 -0.18 -0.096 -0.26 -0.26
DiseasesofHeart -0.26 0.0055 0.91 0.89 0.94 0.9 0.97 0.9 0.61 0.89 0.87 0.94 -0.18 1 0.98 0.45 0.44
CerebrovascularDiseases -0.17 0.013 0.88 0.86 0.96 0.92 0.95 0.86 0.53 0.85 0.89 0.96 -0.096 0.98 1 0.4 0.38
COVID19MultipleCauseofDeath -0.34 0.083 0.77 0.8 0.29 0.18 0.62 0.61 0.07 0.18 0.14 0.32 -0.26 0.45 0.4 1 1
COVID19UnderlyingCauseofDeath -0.38 0.091 0.76 0.79 0.27 0.17 0.61 0.6 0.071 0.18 0.12 0.29 -0.26 0.44 0.38 1 1
Year
Month
AllCause
NaturalCause
Septicemia
MalignantNeoplasms
DiabetesMellitus
AlzheimerDisease
InfluenzaandPneumonia
ChronicLowerRespiratoryDiseases
OtherDiseasesofRespiratorySystem
NephritisNephroticSyndromeandNephrosis
SymptomsSignsandAbnormalClinicalandLaboratoryFindingsNotElsewhereClassified
DiseasesofHeart
CerebrovascularDiseases
COVID19MultipleCauseofDeath
COVID19UnderlyingCauseofDeath
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



University of Colorado **Boulder**

EDA SUMMARY AND RESULTS (SCENARIO ONE)

- Strong correlation and multicollinearity was found between the features related to types of causes of death and the features related to different disease types. This conclusion serves as the basis for the multilinear regression model approach and further analysis.

MULTILINEAR REGRESSION MODEL (SCENARIO ONE)

```
# Creating the first iteration of the Multilinear Regression model with 'AllCause' as the predicting label.
x = monthly_provisional_data[['MalignantNeoplasms', 'DiabetesMellitus','DiseasesofHeart','AlzheimerDisease',
                              'InfluenzaandPneumonia', 'ChronicLowerRespiratoryDiseases', 'OtherDiseasesofRespirator
                              'DiseasesofHeart']]
y = monthly_provisional_data[['AllCause']]
regr = linear_model.LinearRegression()
regr.fit(x, y)

# with statsmodels
x = sm.add_constant(x) # adding a constant

model = sm.OLS(y, x).fit()
predictions = model.predict(x)
```

EVALUATION METRICS (SCENARIO ONE)

Out[71]:

OLS Regression Results

Dep. Variable:	AllCause	**R-squared:**	0.966
Model:	OLS	**Adj. R-squared:**	0.958
Method:	Least Squares	**F-statistic:**	136.3
Date:	Sun, 25 Jun 2023	**Prob (F-statistic):**	5.63e-23
Time:	08:26:55	**Log-Likelihood:**	-439.56
No. Observations:	42	**AIC:**	895.1
Df Residuals:	34	**BIC:**	909.0
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7829.8612	1.41e+04	0.554	0.583	-2.09e+04	3.65e+04
MalignantNeoplasms	-0.7712	0.969	-0.796	0.432	-2.740	1.197
DiabetesMellitus	49.9709	9.938	5.028	0.000	29.775	70.167
DiseasesofHeart	-2.0285	1.151	-1.762	0.087	-4.368	0.311
AlzheimerDisease	7.0316	3.323	2.116	0.042	0.279	13.785
InfluenzaandPneumonia	-4.7917	3.937	-1.217	0.232	-12.792	3.208
ChronicLowerRespiratoryDiseases	0.3633	5.382	0.068	0.947	-10.574	11.301
OtherDiseasesofRespiratorySystem	16.7025	9.921	1.683	0.101	-3.460	36.865
DiseasesofHeart	-2.0285	1.151	-1.762	0.087	-4.368	0.311

Omnibus:	1.254	**Durbin-Watson:**	1.389
Prob(Omnibus):	0.534	**Jarque-Bera (JB):**	1.117
Skew:	0.382	**Prob(JB):**	0.572
Kurtosis:	2.764	**Cond. No.**	1.43e+18

Out[18]:

OLS Regression Results

Dep. Variable:	COVID19MultipleCauseofDeath	**R-squared:**	0.883
Model:	OLS	**Adj. R-squared:**	0.859
Method:	Least Squares	**F-statistic:**	36.65
Date:	Sun, 25 Jun 2023	**Prob (F-statistic):**	4.93e-14
Time:	15:27:55	**Log-Likelihood:**	-441.40
No. Observations:	42	**AIC:**	898.8
Df Residuals:	34	**BIC:**	912.7
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.583e+04	1.48e+04	1.073	0.291	-1.42e+04	4.58e+04
MalignantNeoplasms	-3.3477	1.012	-3.308	0.002	-5.405	-1.291
DiabetesMellitus	48.2608	10.383	4.648	0.000	27.160	69.362
DiseasesofHeart	-2.8382	1.203	-2.360	0.024	-5.282	-0.394
AlzheimerDisease	7.2680	3.472	2.093	0.044	0.212	14.324
InfluenzaandPneumonia	-6.3079	4.113	-1.534	0.134	-14.666	2.050
ChronicLowerRespiratoryDiseases	1.1252	5.623	0.200	0.843	-10.302	12.553
OtherDiseasesofRespiratorySystem	9.8939	10.366	0.954	0.347	-11.172	30.960
DiseasesofHeart	-2.8382	1.203	-2.360	0.024	-5.282	-0.394

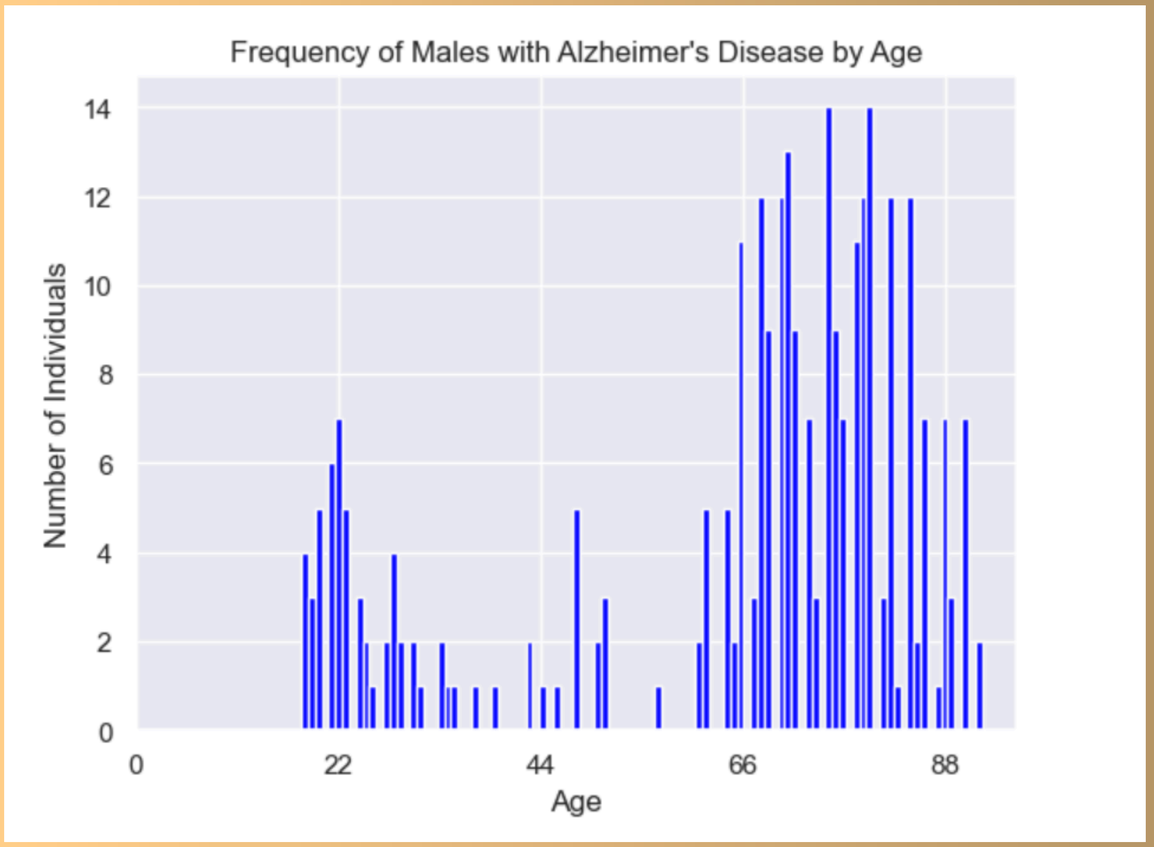
Omnibus:	1.176	**Durbin-Watson:**	1.306
Prob(Omnibus):	0.555	**Jarque-Bera (JB):**	1.190
Skew:	0.347	**Prob(JB):**	0.552
Kurtosis:	2.556	**Cond. No.**	1.43e+18

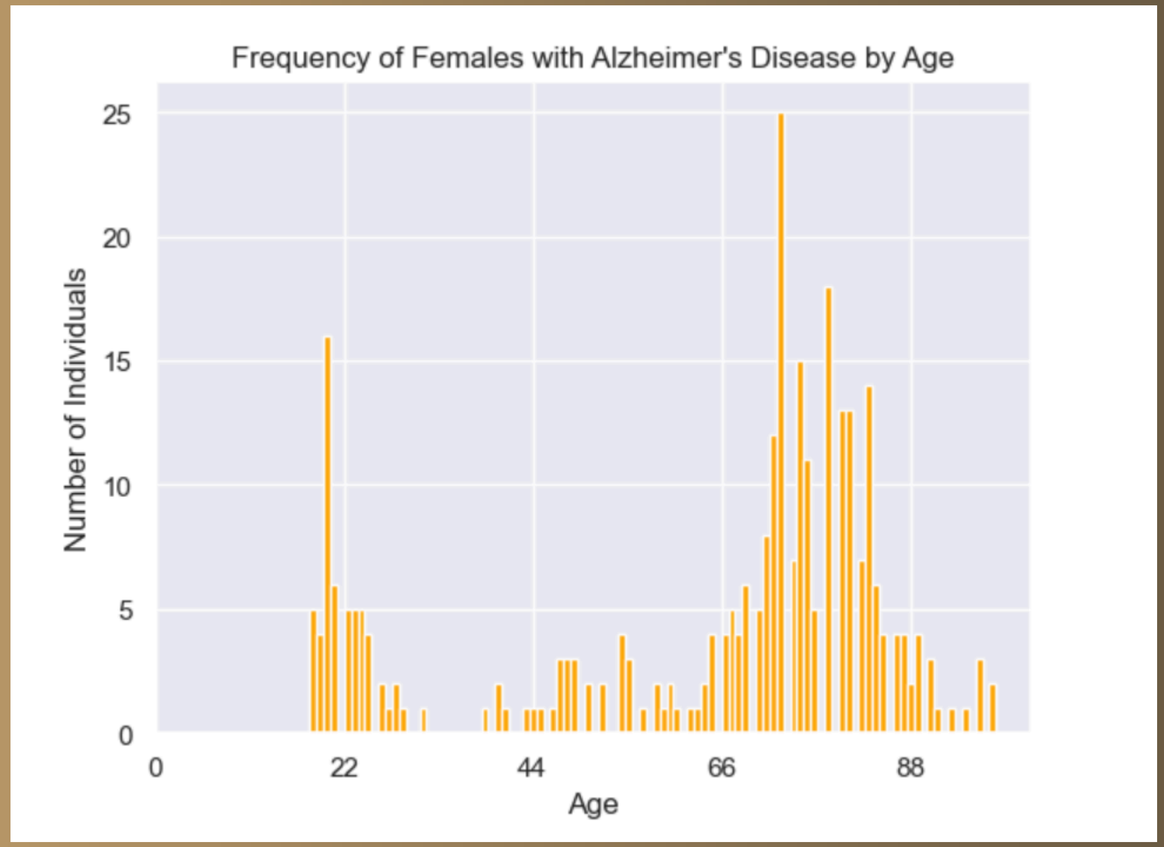
University of Colorado **Boulder**

EXPLORATORY DATA ANALYSIS (SCENARIO TWO)

- Selected features for the second scenario of the project are as follows: ‘age’ and ‘gender’.
- Generated histograms show the frequency of male and female patients with Alzheimer’s disease among different ages.

EXPLORATORY DATA ANALYSIS (SCENARIO TWO)





LOGISTIC REGRESSION MODEL (SCENARIO TWO)

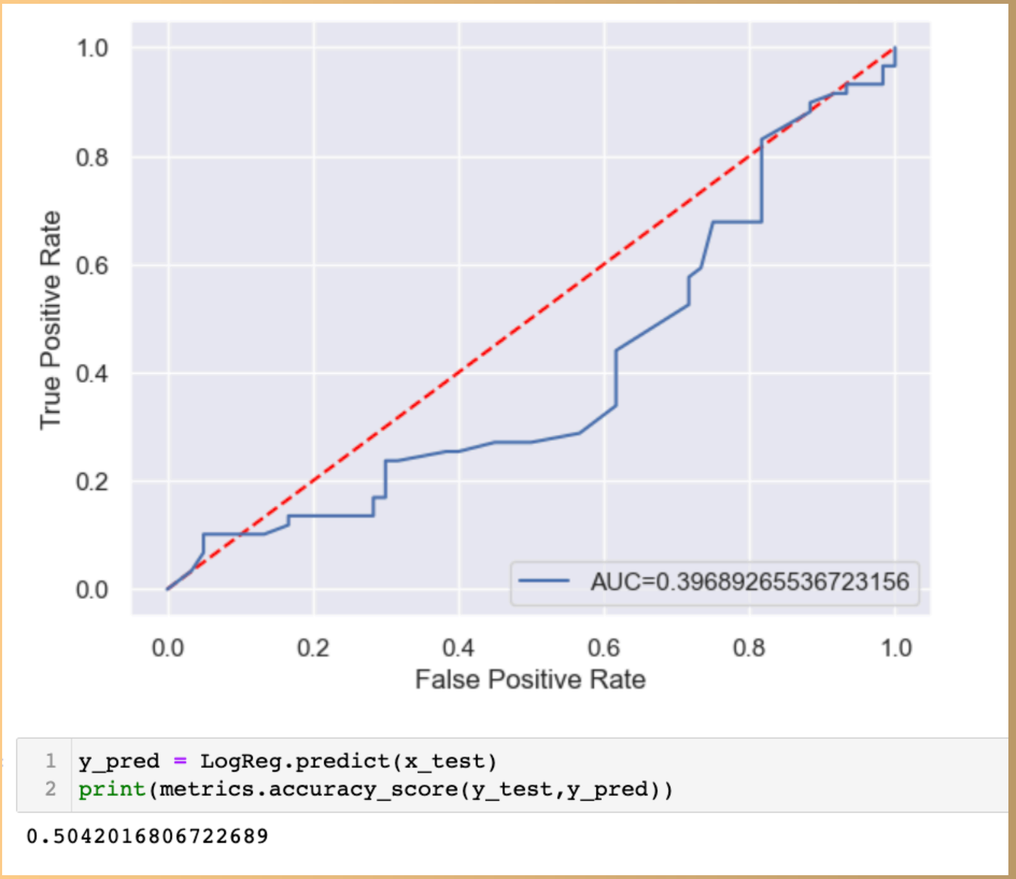
In [48]:

```
# Forms the first iteration of the logistic regression for scenario 2.

X = scenario2_data.iloc[:, :-1].values
y = scenario2_data.iloc[:,-1].values
x_train, x_test, y_train, y_test = model_selection.train_test_split(X,y,test_size=0.2, random_state=5)

LogReg = LogisticRegression(solver='liblinear')
LogReg.fit(x_train, y_train)
```

Out[48]:

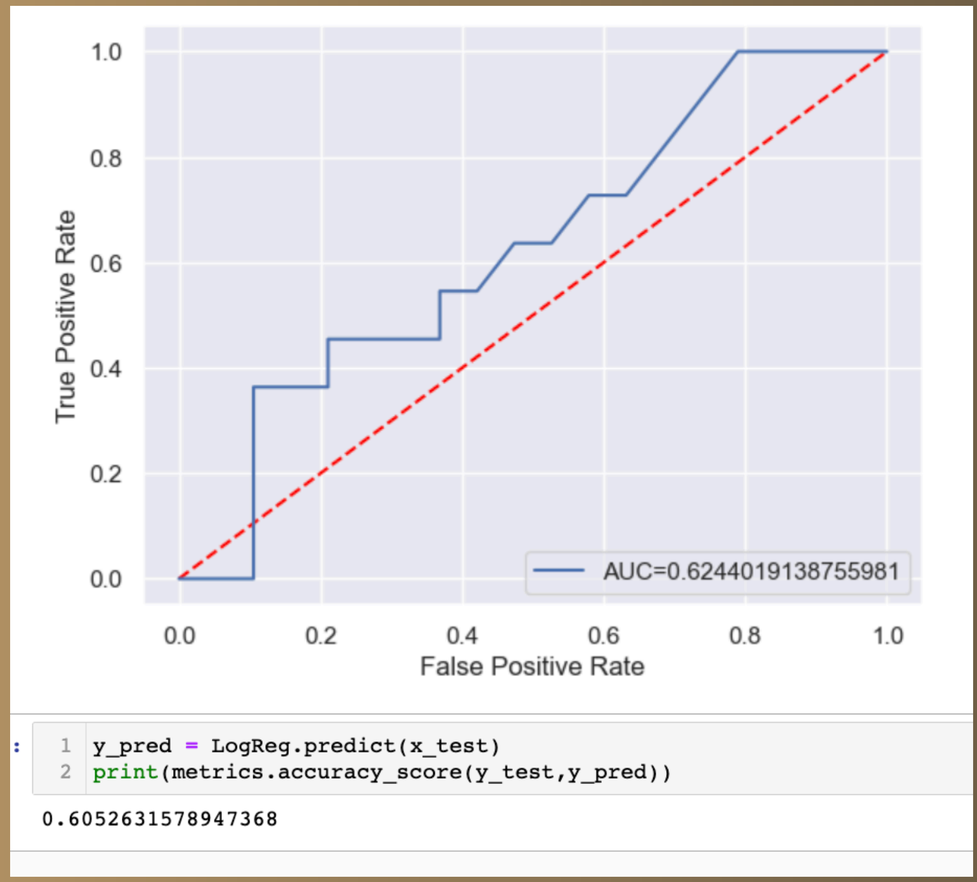
```
LogisticRegression
LogisticRegression(solver='liblinear')
```

EVALUATION METRICS (SCENARIO TWO)




```
y_pred = LogReg.predict(x_test)
print(metrics.accuracy_score(y_test,y_pred))
```

```
0.5042016806722689
```




```
y_pred = LogReg.predict(x_test)
print(metrics.accuracy_score(y_test,y_pred))
```

```
0.6052631578947368
```

University of Colorado **Boulder**

RESULTS AND DISCUSSION

- For scenario one, the following column features were selected as dependent variables for the multilinear regression model: ‘AllCause’, ‘COVID19MultipleCauseofDeath’, and ‘COVIDUnderlyingCauseofDeath’. As a column feature, Alzheimer’s disease was found to be significant for two of these variables: ‘AllCause’ (category for all causes of death) and ‘COVID19MultipleCauseofDeath’ (category for multiple causes leading to COVID19 deaths). The goal for scenario one was achievable.

- In scenario two, the selected features (‘age’ and ‘gender’) were found to be not well correlated. Despite this, model analysis revealed effective methods for necessary improvements for model performance and accuracy: dataset pruning and increasing random shuffling on input data. As a result, concluding AUC metrics and accuracy increased for this scenario. The initial goals for scenario two were achievable.

LESSONS LEARNED

- Data cleaning was revealed to be a critical aspect for the project.
- Repeating scenarios with different datasets can reveal different correlation results for analysis.
- Testing the same machine learning models with larger scales of data can determine robustness of the given models.

CONCLUSION

- In conclusion, multilinear regression models were found to be an effective model approach in identifying Alzheimer's disease as a significant feature in the presence of other diseases. This model approach can be beneficial for other similar scenarios where multiple diseases are involved.

- For scenario two, even though strong correlation was not found, dataset pruning and increased random shuffling of input data were found to be effective methods when improving accuracy of the given logistic regression model. Future tests will need to be conducted, and repeating the scenario with different features would be beneficial in finding stronger correlation and collinearity.

THANK YOU

University of Colorado **Boulder**