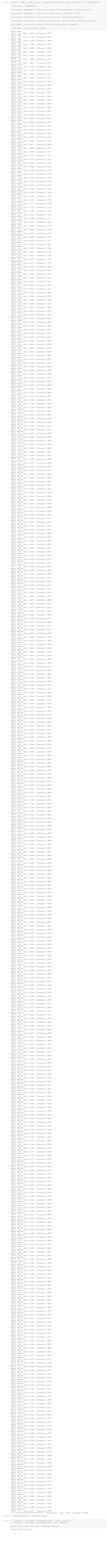
	Introduction For purposes of this CA, I was given a dataset "glass_data.csv", the dataset consists of following columns:
	 Id number: 1 to 214 RI: refractive index Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10) Mg: Magnesium Al: Aluminium Si: Silicon K: Potassium Ca: Calcium Ba: Barium Fe: Iron Type of glass: (class attribute)
	 1 building_windows_float_processed 2 building_windows_non_float_processed 3 vehicle_windows_float_processed 4 vehicle_windows_non_float_processed (none in this database) 5 containers 6 tableware 7 headlamps The objective for me is to conduct an analysis using a neural network with the following dataset is as follows: 1. To perform EDA on the dataset and discuss findings and what relevance they have on the Neural network I wish to create.
In [1]:	import pandas as pd
	<pre>import tensorflow as tf from tensorflow import keras from tensorflow.keras.models import Sequential from tensorflow.keras.layers import Dense from tensorflow import optimizers from sklearn import metrics from sklearn.metrics import accuracy_score from sklearn.model_selection import train_test_split from sklearn import preprocessing from sklearn.preprocessing import MinMaxScaler from sklearn.preprocessing import StandardScaler</pre>
	<pre>from sklearn.preprocessing import LabelEncoder from sklearn.ensemble import RandomForestClassifier import matplotlib from matplotlib import pyplot as plt %matplotlib inline import warnings warnings.filterwarnings("ignore", category=FutureWarning) warnings.filterwarnings("ignore", category=DeprecationWarning) from keras.models import Sequential from keras.layers import Dense</pre>
	import xgboost as xgb from xgboost import XGBClassifier Libraries - I imported multiple libraries for the following: • Pandas - for creating a data frame. • Keras and tesnsorflow - For creating machine learning and Neural Network Models, along with multiple parameters that I would use for my test models.
In [2]:	 Numpy - To convert the data to a numpy array. Sklearn - For pre-processing data and checking accuracy, metrics among others. Matplotlib - For creating plots. xgboost - For Xgboost test and classification Warnings - To supress warnings. 1. Exploratory Data Analysis Creating a variable to store the column names for EDA analysis column_names = ['Id', 'RI', 'Na', 'Mg', 'Al', 'Si', 'K', 'Ca', 'Ba', 'Fe', 'Glass Type'] Reading the file and creating a data frame data = pd.read_csv('glass_data.csv', names = column_names, index_col = 0, header = None)
Out[8]:	RI Na Mg Al Si K Ca Ba Fe Glass Type Id 1 1.52101 13.64 4.49 1.10 71.78 0.06 8.75 0.0 0.0 1 2 1.51761 13.89 3.60 1.36 72.73 0.48 7.83 0.0 0.0 1 3 1.51618 13.53 3.55 1.54 72.99 0.39 7.78 0.0 0.0 1 4 1.51766 13.21 3.69 1.29 72.61 0.57 8.22 0.0 0.0 1 5 1.51742 13.27 3.62 1.24 73.08 0.55 8.07 0.0 0.0 1 I used the ID index column data as index for the created data frame, this is due to the fact that the ID data serves no purpose for predicting the output target glass types in classification or neural network models that I will create further in my analysis. Check the shape of data
	data.shape (214, 10) Check null values and data types within the dataset data.info() <class 'pandas.core.frame.dataframe'=""> Int64Index: 214 entries, 1 to 214 Data columns (total 10 columns): # Column Non-Null Count Dtype</class>
In [6]:	4 Si 214 non-null float64 5 K 214 non-null float64 6 Ca 214 non-null float64 7 Ba 214 non-null float64 8 Fe 214 non-null float64 9 Glass Type 214 non-null int64 dtypes: float64(9), int64(1) memory usage: 18.4 KB Check the data for value spread and distribution of values for mean and min-max between the columns of independent variables data.describe()
Out[6]:	RI Na Mg AI Si K Ca Ba Fe Glass Type count 214,000000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,000000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,00000 214,000000 214,00000 214,00000 214,0
	Glass_types_sum
In [9]:	1 70 7 29 3 17 5 13 6 9 Name: Glass Type, dtype: int64 data['Glass Type'].value_counts().plot(kind='bar') plt.xlabel("Glass Types", labelpad=14) plt.ylabel("Count", labelpad=10) Text(0, 0.5, 'Count')
out[9].	70 - 60 - 50 - 40 - 30 - 20 -
	There are six types of glass in the dataset, it may be correct in assuming the data is imbalanced for categorical classification glass types with the most prominent type being 'type 2' and least being 'type 6'. Check Correlation of Glass dataset
<pre>In [10]: Out[10]:</pre>	plt.figure(figsize=(12, 6)) corr=data.corr() sns.heatmap(corr,cmap='viridis',annot=True) <matplotlib.axessubplots.axessubplot 0x23e9f17f850="" at=""> 2 - 1</matplotlib.axessubplots.axessubplot>
	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$
In [11]:	There is high correlation between Calcium and Refractive index showing 81%, also some low correlation among Barium, Aluminium and Potassium. The least correlating feature being Magnesium in the glass dataset excluding the target variable Glass type. plt.figure(figsize=(10, 5))
	cmap= sns.diverging_palette(150, 275, s=80, 1=55, n=9) heatmap = sns.heatmap(data.corr()[['Glass Type']].
	Al - 0.6 Ba - 0.58 - 0.50 Na - 0.5 Si - 0.15 Ca - 0.00095 K 0.01 RI 0.16 Fe 0.19 - 0.75 - 0.50 - 0.25 - 0.00 0.25 0.25 0.25 0.75
	Sodium, Calcium and Barium is showing moderate to good correlations with the target variable Glass type. Preliminary report for EDA The dataset shares float values for all independent variables within the dataset, the target variable consists of glass type values as integers and no data records for glass type 4 form the dataset.
	Due the nature of how the nature of how the integers values are assigned to glass data type I will need encode glass type column to individual categorical units and convert them in to a 1 dimensional array for my neural network. I can verify the method further by cross validating the data with other algorithms, using the target variable with no One-hot or categorical encoding. 1. Pre-processing data for the machine learning models and justification on how the chosen methods should help me. 2. Create and implement a Dense Neural Network that will output the glass type classification. 3. Test and try to improve the accuracy of the model using various methods, parameters in its accuracy. 4. Make a Classification using the Final choice of model and validate the results. 2. Preprocess the data to fit the machine learning models and evalution with algorithms i. Converting the exisisting dataframe to numpy for Machine learning model going forward
In [13]:	<pre>data_num = data.to_numpy() X = data_num[:,0:9] y = data_num[:,9] y[150:170] array([3., 3., 3., 3., 3., 3., 3., 3., 3., 3.,</pre>
In [15]:	<pre>ii. Scaling the variables scaler = MinMaxScaler() dataset1 = data.values X = dataset1[:,0:9] y = dataset1[:,9] X = scaler.fit_transform(X)</pre>
In [16]:	<pre>iii. Using XGBoost to find the feature importance model = XGBClassifier(n_estimators=50) model.fit(X, y) feature_importance = model.feature_importances_ plt.figure(figsize=(10, 6))</pre>
	<pre>plt.yscale('log', nonposy='clip') plt.bar(range(len(feature_importance)), feature_importance, align='center') plt.xticks(range(len(feature_importance)), rotation='horizontal') plt.title('Feature importance', size=15) plt.ylabel('Importance', size=12) plt.xlabel('Features', size=12) plt.show()</pre>
	C:\Users\USER\anaconda3\lib\site-packages\xgboost\sklearn.py:888: UserWarning: The use of label encod er in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2,, [num_class - 1]. warnings.warn(label_encoder_deprecation_msg, UserWarning) [16:52:22] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.3.0/src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old beha vior. Feature importance Feature importance
	mportance 10-1
	Using feature importance algorithm model for XGBoost, I see the feature 7 being the most important and feature 4 as the least important in contribution to the glass type classification. I can test the accuracy further conducting a XGBoost classifier evaluation.
	<pre>iv. Split the data 70/30 in train/test for machine learning models X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) v. XGBoost Classifier xgb = XGBClassifier(max_depth = 6, n_estimators = 100, learning_rate = 0.07, random_state = 2) xgb.fit(X_train,y_train)</pre>
In [19]:	<pre>y_pred = xgb.predict(X_test) xgb_accuracy = metrics.accuracy_score(y_pred,y_test) [16:52:23] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.3.0/src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if you'd like to restore the old beha vior. print ('The Accuracy for XGBoost classifier:',xgb_accuracy)</pre>
	The Accuracy for XGBoost classifier: 0.8307692307692308 I achieved an accuracy score of 83% approximately after choosing max_depth = 6 as default value and a learning rate of 0.07, it is an impressive accuracy percentage with the help of regularization and tree pruning methods the XGBoost method offers. It is no surprise as a prominent algorithm in the industry for data analysis.
In [20]:	<pre>vi. Check accuracy with Random Forest X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) glf=RandomForestClassifier(n_estimators=50)</pre>
	<pre>X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) clf=RandomForestClassifier(n_estimators=50) clf.fit(X_train,y_train) y_predict=clf.predict(X_test) score_R = metrics.accuracy_score(y_test, y_pred) print("Accuracy {:.2f}".format(score_R)) Accuracy 0.83</pre>
In [21]:	<pre>X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) clf=RandomForestClassifier(n_estimators=50) clf.fit(X_train,y_train) y_predict=clf.predict(X_test) score_R = metrics.accuracy_score(y_test, y_pred) print("Accuracy {:.2f}".format(score_R))</pre>
In [21]: Out[21]:	<pre>X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) clf=RandomForestClassifier(n_estimators=50) clf.fit(X_train,y_train) y_predict=clf.predict(X_test) score_R = metrics.accuracy_score(y_test, y_pred) print("Accuracy (:.2f)".format(score_R)) Accuracy 0.83 y_predict[0:5] array([6., 7., 7., 2., 1.]) IImported the Random Forest Classifier from Sklearn library, after doing so I created a gaussian classifier with n_estimator (number of trees before cut-off) being 50. I trained and tested the values to make predictions, the accuracy score of 83%. The accuracy is quite good, and I must try to match the same if not more using a Dense Neural network. 3. Dense Neural Network Read the data for creating the dense neural network models data = pd.read_csv('glass_data.csv',names = column_names,index_col = 0,header = None) num_data = data.to_numpy() dataset1 = num_data x_cols = dataset1[:,0:9].astype(float) y_cols = dataset1[:,9]</pre>
In [21]: Out[21]: In [22]:	<pre>X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) clf=RandomPoreatClassifier(n_eatimators=50) clf.fit(X_train,y_train) y_prodict=clf.prodict(X_test) score_R = metrics.accuracy_score(y_test, y_pred) print("Accuracy {:.2f}".forwat(score_R)) Accuracy 0.83 y_predict[0:5] array([6., 7., 7., 2., 1.]) IImported the Random Forest Classifier from Sklearn library, after doing so I created a gaussian classifier with n_estimator (number of trees before cut-off) being 50.1 trained and tested the values to make predictions, the accuracy score of 83%. The accuracy is quite good, and I must try to match the same if not more using a Dense Neural network. 3. Dense Neural Network Read the data for creating the dense neural network models data = pd.read_cav('glass_data.cav',names = column_names,index_col = 0,header = None) num_data = data.to_numpy() dataset1 = num_data x_cols = dataset1[i,9] preprocessing.scate(X_cols, copy=False) encoder = preprocessing.scate(X_cols, copy=False) encoder = preprocessing.labelEncoder() encoder.fit(y_cols) y_tran = cncoder.transform(y_cols).reshape(-1, 1) y_enc = np_utils.to_categorical(y_tran) y_enc_x_cols</pre>
In [21]: Out[21]: In [22]:	<pre>X_train, X_teat, y_train, y_teat - train_teat_aplit(x, y, teat_size=0.3, random_atate=1) clf=RandomForeatClassifier(n_estimators=50) cif.fit(X_train,y_truin) y_predict=clf.predict(X_teat) score_R - metrics.accuracy_score(y_teat, y_pred) print("Accuracy (:.2f)".format(score_R)) Accuracy 0.83 y_predict[0:s] array([6., 7., 7., 2., 1.]) Imported the Random Forest Classifier from Sklearn library, after doing so I created a gaussian classifier with n_estimator (number of trees before cut-off) being 50.1 trained and tested the values to make predictions, the accuracy score of 83%. The accuracy is quite good, and I must try to match the same if not more using a Dense Neural network. 3. Dense Neural Network Read the data for creating the dense neural network models data = pd.read_osv('glass_data.csv',names = column_names,index_col = 0,header = None) num_data = data.to_numpy() dataaetl = num_data X_cols = datasetl[:,0:9].astype(float) y_cols = concoder.transform(y_cols).roshape(-1, 1) y_enc = np_utils.to_categorical(y_tran)</pre>
In [21]: Out[21]: In [22]:	<pre>X_train, X_test, y_train, y_test = train_test_apilitix, y, test_size=0.5, random_atate=1: cif=Rendomeroreatulessifice(n_ostimators=50) pif,fik(X_train,y_train) y_predict-cif.predict(X_test) score_R = testrics.accuracy_score(y_test, y_pred) printf(Nomerory (1.28)**.formeh(Nomero_R)) Accuracy 0.83 y_predict(N; a) array(16, 7., 7., 2., 1.) Imported the Random Forest Classifier from Sklean Biray, after doing so created a gaussian classifier with n_estimator (number of trees before cut-off) being 50. I trained and lested the values to make predictors, the accuracy score of 83%. The accuracy is quite good, and must try to match the same if not more using a Dense Neural network. 3. Dense Neural Network Read the data for creating the dense neural network models data = pd.read_cav('glass_data.csv', names = column_names, index_col = 0, header = None) num_data = data.to_mupy() ataaset1 = num_data x_cols = dataset[[,0:3].astype(float) y_mols = data</pre>
In [21]: Out[21]: In [22]:	A train, A test, Vitain, Vitasin, Vitasin, Vitasin, Vitasin, test split(X, V, test size=0.), random state=1) cli=HendomPore+Colored [Period to Organization Section 1] strict("Colored Vivage (1, 26)", format (score X)) Amonor R = master, occurating according tests, y grand trains("Generation Colored (1, 26)", format (score X)) Amonor R = proceed occuration (Colored (1, 26)", format (score X)) Amonor R = proceded (1, 26)", format (score X)) Amonor R = proceded (1, 26)", format (score X)) Amonor R = master, occuration (1, 26)", format (score X)) Amonor Vivage (1, 26)", format (score X)) Imposted the Random Forced Chaodier from Skleam Biology, after doing so I created a gaussian classifier with no patients for the state from the same in not more using a London Brown. 3. Dense Neural Network 3. Dense Neural Network Association of the same income using a London Reversi network. 3. Dense Neural Network Association of the dense neural network models data = politication of the dense neural network models data = politication of the dense neural network models data = politication of the dense neural network models association of the same of the dense neural network models data = politication of the dense neural network models association of the dense neural network models and politic network models and bette network net
In [21]: Out[21]: In [22]:	X_notin_X_notin_X_notin_X_notin_ton_notin_ton_notin_X_N_N_ton_notin_ton_notin_X_N_N_ton_notin_ton_notin_ton_notin_X_N_N_ton_notin_ton_notin_X_N_N_ton_N_N_N_N_N_N_N_N_N_N_N_N_N_N_N_N_N_N_N
In [21]: Out[21]: In [22]:	Secretary Country Secretary Secret
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S
In [21]: Out[21]: In [22]:	Standard S



Model 2 • Sequentia	ates with higher epochs/batch size loes seem to reduce over time throot a good fit and may need to try				in errors.
 I have add Test size random_s while the r kernel_in Activation or causing Final Acti Optimizer Batch size 	I Model led 9 input layers, along with 3 hid = 0.3 a 70/30 split for training and state = 2 set as 2 to evaluate the model learns through the training sitializer = 'normal' for normal dist in = 'relu' for linear activation initial to vanishing gradient. vation = 'softmax' for classification = adam to reduce the loss and a e = 80, Epochs=1200 SE' - As Mean squared error is set	I testing the dataset at the randomness between the set. tribution of weights throughly, serves a costless means on of glass types and preadjust weights in model transfer.	e beginning. e values with various ma gh the neural layers. thod for achieving a goo edicting the probabilities aining and fast converge	d learning rate without in with last layer of logits. ence with each iteration.	mpeding the
	to obtain the mean among the err		loss over the number of	epochs by squaring the	values and

