**Machine Learning Models Can Save Lives**

**Problem description**

Heart related diseases have been the leading cause of death worldwide that has resulted in the loss of millions in each every year to this common type of disease. There are three sets of datasets with different sets of attributes that are able to be used as indicators that a machine learning model can utilize to predict a possible heart disease in a patient. The need for creating machine learning models to predict possible heart diseases in patients with the highest possible accuracy so medical care facilities have the ability of early detection of heart disease in a patient to be able to begin managing and caring for the patient’s disease as early as possible in order to increase the odds of their survival.

The three datasets we utilized in order to create machine learning models to learn and predict heart related issues through different sets of similar attributes related to determining heart issues.

**Dataset 1:** [Heart Failure Prediction Dataset | Kaggle](https://www.kaggle.com/fedesoriano/heart-failure-prediction)

11 Attributes

Total: 1190 observations

Unique: 918 Observations

Duplicated: 272 observations

**Dataset 2:** [Heart Disease UCI | Kaggle](https://www.kaggle.com/ronitf/heart-disease-uci)

13 Attributes

Total: 303 observations

**Dataset 3:** [Heart Failure Prediction | Kaggle](https://www.kaggle.com/andrewmvd/heart-failure-clinical-data)

12 Attributes

Total: 299

**Algorithms implemented**

The two types of validation methods we utilized for all the models used in all 3 datasets were k-fold cross validation and the usual train and test split. The reason we chose to implement and try out k-fold cross validation was due mainly to the fact that the size of our dataset would not have us experiencing the disadvantage of k-fold cross validations which is waiting for the time consuming training of every model to be completed.

We trained and tested every model just to be able to understand how they all compared to each other with the use of both types of validations and testing various data augmented versions of the dataset. We were only able to utilize keras dense model due to the dataset used is neither an image or a body of text that can be utilized by CNN to employ any form of deep learning. The sklearn models we used in every dataset were for regression we used 'K-Nearest Neighbors regressor, Decision Tree Regressor, Random Forest Regressor, Bagging Regressor, and Linear Regression. The sklearn classifiers we utilized in every dataset were K-Nearest NeighborsClassifier, GaussianNB, BernoulliNB, Decision Tree Classifier, Random Forest Classifier, Logistic Regression, and SVC.

The data augmentation we utilized were the products of attributes to see if the creation of more attributes can help the models be able to learn and be able to further increase accuracy or lower MSE based on a the newly created attributes based on the relationships they have with themselves and with other attributes within the dataset to see if anything can be learned from the the result of the dual combinations. We also created a Normalized and Standardized version of each dataset to help alleviate the influence of having some attributes seen as more important than other attributes due to their values in measurements.

**Experimental Results** (including accuracies or mean squared errors and running times for each algorithm and parameter choice. You may want to use tables and or plots to illustrate this)

**Dataset 1:**

**Regression:**

Training all the models with this dataset variant: Original dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.2204

Elapsed time = 0.0044 secs

Evaluating Decision Tree Regressor

MSE = 0.2935

Elapsed time = 0.0034 secs

Evaluating Random Forest Regressor

MSE = 0.1534

Elapsed time = 0.2516 secs

Evaluating Bagging Regressor

MSE = 0.1676

Elapsed time = 0.0320 secs

Evaluating Linear Regression

MSE = 0.1527

Elapsed time = 0.0035 secs

By utilizing this dataset variant: Original dataset

The best regression model that performed under this dataset is Linear Regression

Lowest MSE acheived = 0.1527

Training all the models with this dataset variant: Normalized Original Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.1504

Elapsed time = 0.0038 secs

Evaluating Decision Tree Regressor

MSE = 0.2826

Elapsed time = 0.0029 secs

Evaluating Random Forest Regressor

MSE = 0.1500

Elapsed time = 0.2629 secs

Evaluating Bagging Regressor

MSE = 0.1582

Elapsed time = 0.0331 secs

Evaluating Linear Regression

MSE = 0.1527

Elapsed time = 0.0017 secs

By utilizing this dataset variant: Normalized Original Dataset

The best regression model that performed under this dataset is Random Forest Regressor

Lowest MSE acheived = 0.1500

Training all the models with this dataset variant: Standardized Original Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.1524

Elapsed time = 0.0044 secs

Evaluating Decision Tree Regressor

MSE = 0.2717

Elapsed time = 0.0036 secs

Evaluating Random Forest Regressor

MSE = 0.1514

Elapsed time = 0.2583 secs

Evaluating Bagging Regressor

MSE = 0.1722

Elapsed time = 0.0311 secs

Evaluating Linear Regression

MSE = 0.1527

Elapsed time = 0.0030 secs

By utilizing this dataset variant: Standardized Original Dataset

The best regression model that performed under this dataset is Random Forest Regressor

Lowest MSE acheived = 0.1514

Training all the models with this dataset variant: Products of Attributes Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.2348

Elapsed time = 0.0096 secs

Evaluating Decision Tree Regressor

MSE = 0.2500

Elapsed time = 0.0209 secs

Evaluating Random Forest Regressor

MSE = 0.1319

Elapsed time = 0.9695 secs

Evaluating Bagging Regressor

MSE = 0.1570

Elapsed time = 0.0991 secs

Evaluating Linear Regression

MSE = 0.1363

Elapsed time = 0.0080 secs

By utilizing this dataset variant: Products of Attributes Dataset

The best regression model that performed under this dataset is Random Forest Regressor

Lowest MSE acheived = 0.1319

Training all the models with this dataset variant: Normalized Products of Attributes Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.1504

Elapsed time = 0.0057 secs

Evaluating Decision Tree Regressor

MSE = 0.2554

Elapsed time = 0.0199 secs

Evaluating Random Forest Regressor

MSE = 0.1332

Elapsed time = 0.9816 secs

Evaluating Bagging Regressor

MSE = 0.1428

Elapsed time = 0.1069 secs

Evaluating Linear Regression

MSE = 0.1379

Elapsed time = 0.0203 secs

By utilizing this dataset variant: Normalized Products of Attributes Dataset

The best regression model that performed under this dataset is Random Forest Regressor

Lowest MSE acheived = 0.1332

Training all the models with this dataset variant: Standarized Products of Attributes Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.1443

Elapsed time = 0.0074 secs

Evaluating Decision Tree Regressor

MSE = 0.2717

Elapsed time = 0.0199 secs

Evaluating Random Forest Regressor

MSE = 0.1364

Elapsed time = 0.9663 secs

Evaluating Bagging Regressor

MSE = 0.1436

Elapsed time = 0.1025 secs

Evaluating Linear Regression

MSE = 0.1383

Elapsed time = 0.0171 secs

By utilizing this dataset variant: Standarized Products of Attributes Dataset

The best regression model that performed under this dataset is Random Forest Regressor

Lowest MSE acheived = 0.1364

**Classifiers:**

Using Test and Train Split

Training all the models with this dataset variant: Original dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.7011

Elapsed time = 0.0096 secs

Evaluating GaussianNB

Accuracy = 0.8207

Elapsed time = 0.0015 secs

Evaluating BernoulliNB

Accuracy = 0.8261

Elapsed time = 0.0052 secs

Evaluating Decision Tree classifer

Accuracy = 0.7174

Elapsed time = 0.0033 secs

Evaluating Random Forest classifer

Accuracy = 0.7880

Elapsed time = 0.2020 secs

By utilizing this dataset variant: Original dataset

The best classification model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8261

Training all the models with this dataset variant: Normalized Original Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.7609

Elapsed time = 0.0141 secs

Evaluating GaussianNB

Accuracy = 0.8207

Elapsed time = 0.0034 secs

Evaluating BernoulliNB

Accuracy = 0.8315

Elapsed time = 0.0029 secs

Evaluating Decision Tree classifer

Accuracy = 0.7174

Elapsed time = 0.0040 secs

Evaluating Random Forest classifer

Accuracy = 0.7989

Elapsed time = 0.2104 secs

By utilizing this dataset variant: Normalized Original Dataset

The best classification model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8315

Training all the models with this dataset variant: Standardized Original Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.7880

Elapsed time = 0.0159 secs

Evaluating GaussianNB

Accuracy = 0.8207

Elapsed time = 0.0043 secs

Evaluating BernoulliNB

Accuracy = 0.7826

Elapsed time = 0.0030 secs

Evaluating Decision Tree classifer

Accuracy = 0.7011

Elapsed time = 0.0042 secs

Evaluating Random Forest classifer

Accuracy = 0.7826

Elapsed time = 0.2037 secs

By utilizing this dataset variant: Standardized Original Dataset

The best classification model that performed under this dataset is GaussianNB

Highest Accuracy acheived = 0.8207

Training all the models with this dataset variant: Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.6685

Elapsed time = 0.0188 secs

Evaluating GaussianNB

Accuracy = 0.8207

Elapsed time = 0.0039 secs

Evaluating BernoulliNB

Accuracy = 0.8315

Elapsed time = 0.0047 secs

Evaluating Decision Tree classifer

Accuracy = 0.7391

Elapsed time = 0.0233 secs

Evaluating Random Forest classifer

Accuracy = 0.8098

Elapsed time = 0.3078 secs

By utilizing this dataset variant: Products of Attributes Dataset

The best classification model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8315

Training all the models with this dataset variant: Normalized Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.7717

Elapsed time = 0.0136 secs

Evaluating GaussianNB

Accuracy = 0.7609

Elapsed time = 0.0026 secs

Evaluating BernoulliNB

Accuracy = 0.7826

Elapsed time = 0.0033 secs

Evaluating Decision Tree classifer

Accuracy = 0.7337

Elapsed time = 0.0221 secs

Evaluating Random Forest classifer

Accuracy = 0.8261

Elapsed time = 0.3120 secs

By utilizing this dataset variant: Normalized Products of Attributes Dataset

The best classification model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 0.8261

Training all the models with this dataset variant: Standarized Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.7935

Elapsed time = 0.0146 secs

Evaluating GaussianNB

Accuracy = 0.7609

Elapsed time = 0.0029 secs

Evaluating BernoulliNB

Accuracy = 0.7772

Elapsed time = 0.0036 secs

Evaluating Decision Tree classifer

Accuracy = 0.7283

Elapsed time = 0.0223 secs

Evaluating Random Forest classifer

Accuracy = 0.7989

Elapsed time = 0.3081 secs

By utilizing this dataset variant: Standarized Products of Attributes Dataset

The best classification model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 0.7989

K-Fold Cross Validation:

Training all the models with this dataset variant: Original dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 69.0643

Elapsed time = 0.0270 secs

Evaluating GaussianNB

Accuracy = 76.6082

Elapsed time = 0.0151 secs

Evaluating BernoulliNB

Accuracy = 83.2164

Elapsed time = 0.0183 secs

Evaluating Decision Tree classifer

Accuracy = 73.3626

Elapsed time = 0.0188 secs

Evaluating Random Forest classifer

Accuracy = 82.0760

Elapsed time = 1.7161 secs

By utilizing this dataset variant: Original dataset

The best regression model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 83.2164

Training all the models with this dataset variant: Normalized Original Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 79.8538

Elapsed time = 0.0296 secs

Evaluating GaussianNB

Accuracy = 76.6082

Elapsed time = 0.0164 secs

Evaluating BernoulliNB

Accuracy = 82.6023

Elapsed time = 0.0183 secs

Evaluating Decision Tree classifer

Accuracy = 76.0526

Elapsed time = 0.0212 secs

Evaluating Random Forest classifer

Accuracy = 83.7427

Elapsed time = 1.6783 secs

By utilizing this dataset variant: Normalized Original Dataset

The best regression model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 83.7427

Training all the models with this dataset variant: Standardized Original Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 80.3801

Elapsed time = 0.0292 secs

Evaluating GaussianNB

Accuracy = 76.6082

Elapsed time = 0.0172 secs

Evaluating BernoulliNB

Accuracy = 79.3275

Elapsed time = 0.0227 secs

Evaluating Decision Tree classifer

Accuracy = 74.4444

Elapsed time = 0.0219 secs

Evaluating Random Forest classifer

Accuracy = 81.5497

Elapsed time = 1.6920 secs

By utilizing this dataset variant: Standardized Original Dataset

The best regression model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 81.5497

Training all the models with this dataset variant: Products of Attributes Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 65.2047

Elapsed time = 0.0274 secs

Evaluating GaussianNB

Accuracy = 78.7719

Elapsed time = 0.0180 secs

Evaluating BernoulliNB

Accuracy = 80.4386

Elapsed time = 0.0218 secs

Evaluating Decision Tree classifer

Accuracy = 78.3041

Elapsed time = 0.0522 secs

Evaluating Random Forest classifer

Accuracy = 82.5731

Elapsed time = 1.8667 secs

By utilizing this dataset variant: Products of Attributes Dataset

The best regression model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 82.5731

Training all the models with this dataset variant: Normalized Products of Attributes Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 81.4620

Elapsed time = 0.0247 secs

Evaluating GaussianNB

Accuracy = 75.5848

Elapsed time = 0.0179 secs

Evaluating BernoulliNB

Accuracy = 84.2690

Elapsed time = 0.0222 secs

Evaluating Decision Tree classifer

Accuracy = 77.2222

Elapsed time = 0.0522 secs

Evaluating Random Forest classifer

Accuracy = 84.7661

Elapsed time = 1.8633 secs

By utilizing this dataset variant: Normalized Products of Attributes Dataset

The best regression model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 84.7661

Training all the models with this dataset variant: Standarized Products of Attributes Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 81.4912

Elapsed time = 0.0285 secs

Evaluating GaussianNB

Accuracy = 75.5848

Elapsed time = 0.0182 secs

Evaluating BernoulliNB

Accuracy = 77.7485

Elapsed time = 0.0225 secs

Evaluating Decision Tree classifer

Accuracy = 75.4678

Elapsed time = 0.0523 secs

Evaluating Random Forest classifer

Accuracy = 82.0175

Elapsed time = 1.8581 secs

By utilizing this dataset variant: Standarized Products of Attributes Dataset

The best regression model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 82.0175

**Logistic Regression:**

**Using Train and Test Split:**

Training Linear Regression model with this dataset variant: Original dataset

Accuracy = 0.7935

Elapsed time = 0.6203 secs

Training Linear Regression model with this dataset variant: Normalized Original Dataset.

Accuracy = 0.7772

Elapsed time = 0.2836 secs

Training Linear Regression model with this dataset variant: Standardized Original Dataset

Accuracy = 0.7989

Elapsed time = 0.2830 secs

Training Linear Regression model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.7935

Elapsed time = 0.3158 secs

Training Linear Regression model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.8207

Elapsed time = 0.0435 secs

Training Linear Regression model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.8315

Elapsed time = 0.0580 secs

By utilizing the dataset variant on Linear Regression: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 0.8315

**Using cross validation:**

Training Linear Regression model with this dataset variant: Original dataset

Accuracy = 77.6608

Elapsed time = 0.3850 secs

Training Linear Regression model with this dataset variant: Normalized Original Dataset

Accuracy = 76.6082

Elapsed time = 0.1545 secs

Training Linear Regression model with this dataset variant: Standardized Original Dataset

Accuracy = 74.9415

Elapsed time = 0.1092 secs

Training Linear Regression model with this dataset variant: Products of Attributes Dataset

Accuracy = 77.1637

Elapsed time = 0.4331 secs

Training Linear Regression model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 78.2456

Elapsed time = 0.2802 secs

Training Linear Regression model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 78.8304

Elapsed time = 0.3412 secs

By utilizing the dataset variant on Linear Regression: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 78.8304

**Support Vector Machine (SVM):**

Train and Test Split:

Training SVC model with this dataset variant: Original dataset

Accuracy = 0.6848

Elapsed time = 0.0339 secs

Training SVC model with this dataset variant: Normalized Original Dataset

Accuracy = 0.7717

Elapsed time = 0.0221 secs

Training SVC model with this dataset variant: Standardized Original Dataset

Accuracy = 0.7880

Elapsed time = 0.0228 secs

Training SVC model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.6848

Elapsed time = 0.0436 secs

Training SVC model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.7717

Elapsed time = 0.0300 secs

Training SVC model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.7880

Elapsed time = 0.0307 secs

By utilizing the dataset variant on SVC: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 0.7880

K-Fold Cross Validation:

Training SVC model with this dataset variant: Original dataset

Accuracy = 72.8363

Elapsed time = 0.0652 secs

Training SVC model with this dataset variant: Normalized Original Dataset

Accuracy = 79.8538

Elapsed time = 0.0495 secs

Training SVC model with this dataset variant: Standardized Original Dataset

Accuracy = 81.5205

Elapsed time = 0.0505 secs

Training SVC model with this dataset variant: Products of Attributes Dataset

Accuracy = 73.3041

Elapsed time = 0.0821 secs

Training SVC model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 82.0760

Elapsed time = 0.0633 secs

Training SVC model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 80.9942

Elapsed time = 0.0675 secs

By utilizing the dataset variant on SVC: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 82.0760

**Dense Network:**

Normal Dataset

Max Accuracy score for all folds:

Accuracy: 78.2608687877655

Loss: 0.5494163334369659

Products of attributes Datasets

Max Accuracy score for all folds:

Accuracy: 84.78260636329651

Loss: 0.5402981609106063

Normalized Dataset

Max Accuracy score for all folds:

Accuracy: 88.04348111152649

Loss: 0.4308886617422104

Normalized Products of Attributes

Max Accuracy score for all folds:

Accuracy: 84.78260636329651

Loss: 0.4816463261842728

Standardized Dataset

Max Accuracy score for all folds:

Accuracy: 89.01098966598511

Loss: 0.6685860067605972

Standardized Products of Attributes Dataset

Max Accuracy score for all folds:

Accuracy: 85.86956262588501

Loss: 0.8326970994472503

**Dataset 2:**

**Regression**:

Train and Test Split:

Training all the models with this dataset variant: Original dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.2052

Elapsed time = 0.0056 secs

Evaluating Decision Tree Regressor

MSE = 0.2459

Elapsed time = 0.0024 secs

Evaluating Random Forest Regressor

MSE = 0.1129

Elapsed time = 0.1620 secs

Evaluating Bagging Regressor

MSE = 0.1292

Elapsed time = 0.0245 secs

Evaluating Linear Regression

MSE = 0.1198

Elapsed time = 0.0025 secs

By utilizing this dataset variant: Original dataset

The best regression model that performed under this dataset is Random Forest Regressor

Lowest MSE acheived = 0.1129

Training all the models with this dataset variant: Normalized Original Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.2052

Elapsed time = 0.0032 secs

Evaluating Decision Tree Regressor

MSE = 0.2295

Elapsed time = 0.0015 secs

Evaluating Random Forest Regressor

MSE = 0.1218

Elapsed time = 0.1708 secs

Evaluating Bagging Regressor

MSE = 0.1307

Elapsed time = 0.0217 secs

Evaluating Linear Regression

MSE = 0.1198

Elapsed time = 0.0027 secs

By utilizing this dataset variant: Normalized Original Dataset

The best regression model that performed under this dataset is Linear Regression

Lowest MSE acheived = 0.1198

Training all the models with this dataset variant: Standardized Original Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.2052

Elapsed time = 0.0034 secs

Evaluating Decision Tree Regressor

MSE = 0.2295

Elapsed time = 0.0025 secs

Evaluating Random Forest Regressor

MSE = 0.1150

Elapsed time = 0.1751 secs

Evaluating Bagging Regressor

MSE = 0.1185

Elapsed time = 0.0215 secs

Evaluating Linear Regression

MSE = 0.1198

Elapsed time = 0.0023 secs

By utilizing this dataset variant: Standardized Original Dataset

The best regression model that performed under this dataset is Random Forest Regressor

Lowest MSE acheived = 0.1150

Training all the models with this dataset variant: Products of Attributes Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.2052

Elapsed time = 0.0039 secs

Evaluating Decision Tree Regressor

MSE = 0.2295

Elapsed time = 0.0036 secs

Evaluating Random Forest Regressor

MSE = 0.1144

Elapsed time = 0.1669 secs

Evaluating Bagging Regressor

MSE = 0.1197

Elapsed time = 0.0249 secs

Evaluating Linear Regression

MSE = 0.1198

Elapsed time = 0.0026 secs

By utilizing this dataset variant: Products of Attributes Dataset

The best regression model that performed under this dataset is Random Forest Regressor

Lowest MSE acheived = 0.1144

Training all the models with this dataset variant: Normalized Products of Attributes Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.2052

Elapsed time = 0.0031 secs

Evaluating Decision Tree Regressor

MSE = 0.1967

Elapsed time = 0.0060 secs

Evaluating Random Forest Regressor

MSE = 0.1220

Elapsed time = 0.1663 secs

Evaluating Bagging Regressor

MSE = 0.1303

Elapsed time = 0.0205 secs

Evaluating Linear Regression

MSE = 0.1198

Elapsed time = 0.0010 secs

By utilizing this dataset variant: Normalized Products of Attributes Dataset

The best regression model that performed under this dataset is Linear Regression

Lowest MSE acheived = 0.1198

Training all the models with this dataset variant: Standarized Products of Attributes Dataset

Evaluating K-Nearest Neighbors regressor

MSE = 0.2052

Elapsed time = 0.0037 secs

Evaluating Decision Tree Regressor

MSE = 0.1967

Elapsed time = 0.0039 secs

Evaluating Random Forest Regressor

MSE = 0.1172

Elapsed time = 0.1645 secs

Evaluating Bagging Regressor

MSE = 0.1157

Elapsed time = 0.0198 secs

Evaluating Linear Regression

MSE = 0.1198

Elapsed time = 0.0010 secs

By utilizing this dataset variant: Standarized Products of Attributes Dataset

The best regression model that performed under this dataset is Bagging Regressor

Lowest MSE acheived = 0.1157

**Classifiers**:

Train and Test Split:

Training all the models with this dataset variant: Original dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.6885

Elapsed time = 0.0081 secs

Evaluating GaussianNB

Accuracy = 0.7869

Elapsed time = 0.0026 secs

Evaluating BernoulliNB

Accuracy = 0.8689

Elapsed time = 0.0032 secs

Evaluating Decision Tree classifer

Accuracy = 0.7541

Elapsed time = 0.0029 secs

Evaluating Random Forest classifer

Accuracy = 0.8689

Elapsed time = 0.1773 secs

By utilizing this dataset variant: Original dataset

The best classification model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8689

Training all the models with this dataset variant: Normalized Original Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.8197

Elapsed time = 0.0096 secs

Evaluating GaussianNB

Accuracy = 0.7869

Elapsed time = 0.0038 secs

Evaluating BernoulliNB

Accuracy = 0.8689

Elapsed time = 0.0040 secs

Evaluating Decision Tree classifer

Accuracy = 0.7705

Elapsed time = 0.0023 secs

Evaluating Random Forest classifer

Accuracy = 0.8689

Elapsed time = 0.1668 secs

By utilizing this dataset variant: Normalized Original Dataset

The best classification model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8689

Training all the models with this dataset variant: Standardized Original Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.8525

Elapsed time = 0.0076 secs

Evaluating GaussianNB

Accuracy = 0.7869

Elapsed time = 0.0036 secs

Evaluating BernoulliNB

Accuracy = 0.8852

Elapsed time = 0.0047 secs

Evaluating Decision Tree classifer

Accuracy = 0.7869

Elapsed time = 0.0024 secs

Evaluating Random Forest classifer

Accuracy = 0.8525

Elapsed time = 0.1693 secs

By utilizing this dataset variant: Standardized Original Dataset

The best classification model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8852

Training all the models with this dataset variant: Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.6557

Elapsed time = 0.0111 secs

Evaluating GaussianNB

Accuracy = 0.8197

Elapsed time = 0.0040 secs

Evaluating BernoulliNB

Accuracy = 0.8361

Elapsed time = 0.0045 secs

Evaluating Decision Tree classifer

Accuracy = 0.7541

Elapsed time = 0.0104 secs

Evaluating Random Forest classifer

Accuracy = 0.8852

Elapsed time = 0.2298 secs

By utilizing this dataset variant: Products of Attributes Dataset

The best classification model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 0.8852

Training all the models with this dataset variant: Normalized Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.7705

Elapsed time = 0.0075 secs

Evaluating GaussianNB

Accuracy = 0.8033

Elapsed time = 0.0040 secs

Evaluating BernoulliNB

Accuracy = 0.8525

Elapsed time = 0.0034 secs

Evaluating Decision Tree classifer

Accuracy = 0.8197

Elapsed time = 0.0097 secs

Evaluating Random Forest classifer

Accuracy = 0.8689

Elapsed time = 0.2210 secs

By utilizing this dataset variant: Normalized Products of Attributes Dataset

The best classification model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 0.8689

Training all the models with this dataset variant: Standarized Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.8361

Elapsed time = 0.0074 secs

Evaluating GaussianNB

Accuracy = 0.8033

Elapsed time = 0.0036 secs

Evaluating BernoulliNB

Accuracy = 0.9016

Elapsed time = 0.0042 secs

Evaluating Decision Tree classifer

Accuracy = 0.7869

Elapsed time = 0.0101 secs

Evaluating Random Forest classifer

Accuracy = 0.8689

Elapsed time = 0.2354 secs

By utilizing this dataset variant: Standarized Products of Attributes Dataset

The best classification model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.9016

**K-Fold Validation:**

Training all the models with this dataset variant: Original dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.6071

Elapsed time = 0.0224 secs

Evaluating GaussianNB

Accuracy = 0.8357

Elapsed time = 0.0149 secs

Evaluating BernoulliNB

Accuracy = 0.8381

Elapsed time = 0.0200 secs

Evaluating Decision Tree classifer

Accuracy = 0.7214

Elapsed time = 0.0161 secs

Evaluating Random Forest classifer

Accuracy = 0.8024

Elapsed time = 1.5935 secs

By utilizing this dataset variant: Original dataset

The best regression model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8381

Training all the models with this dataset variant: Normalized Original Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.8024

Elapsed time = 0.0244 secs

Evaluating GaussianNB

Accuracy = 0.8357

Elapsed time = 0.0167 secs

Evaluating BernoulliNB

Accuracy = 0.8381

Elapsed time = 0.0193 secs

Evaluating Decision Tree classifer

Accuracy = 0.7381

Elapsed time = 0.0161 secs

Evaluating Random Forest classifer

Accuracy = 0.8024

Elapsed time = 1.6302 secs

By utilizing this dataset variant: Normalized Original Dataset

The best regression model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8381

Training all the models with this dataset variant: Standardized Original Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.8357

Elapsed time = 0.0273 secs

Evaluating GaussianNB

Accuracy = 0.8357

Elapsed time = 0.0168 secs

Evaluating BernoulliNB

Accuracy = 0.8357

Elapsed time = 0.0216 secs

Evaluating Decision Tree classifer

Accuracy = 0.7548

Elapsed time = 0.0175 secs

Evaluating Random Forest classifer

Accuracy = 0.7857

Elapsed time = 1.5933 secs

By utilizing this dataset variant: Standardized Original Dataset

The best regression model that performed under this dataset is K-Nearest Neighbors Classifier

Highest Accuracy acheived = 0.8357

Training all the models with this dataset variant: Products of Attributes Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.6095

Elapsed time = 0.0222 secs

Evaluating GaussianNB

Accuracy = 0.8357

Elapsed time = 0.0164 secs

Evaluating BernoulliNB

Accuracy = 0.8381

Elapsed time = 0.0203 secs

Evaluating Decision Tree classifer

Accuracy = 0.7333

Elapsed time = 0.0249 secs

Evaluating Random Forest classifer

Accuracy = 0.8024

Elapsed time = 1.6432 secs

By utilizing this dataset variant: Products of Attributes Dataset

The best regression model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8381

Training all the models with this dataset variant: Normalized Products of Attributes Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.8357

Elapsed time = 0.0293 secs

Evaluating GaussianNB

Accuracy = 0.8381

Elapsed time = 0.0153 secs

Evaluating BernoulliNB

Accuracy = 0.8381

Elapsed time = 0.0191 secs

Evaluating Decision Tree classifer

Accuracy = 0.7167

Elapsed time = 0.0262 secs

Evaluating Random Forest classifer

Accuracy = 0.7857

Elapsed time = 1.6281 secs

By utilizing this dataset variant: Normalized Products of Attributes Dataset

The best regression model that performed under this dataset is GaussianNB

Highest Accuracy acheived = 0.8381

Training all the models with this dataset variant: Standarized Products of Attributes Dataset

using cross validation

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.8357

Elapsed time = 0.0261 secs

Evaluating GaussianNB

Accuracy = 0.8381

Elapsed time = 0.0180 secs

Evaluating BernoulliNB

Accuracy = 0.8690

Elapsed time = 0.0226 secs

Evaluating Decision Tree classifer

Accuracy = 0.7833

Elapsed time = 0.0259 secs

Evaluating Random Forest classifer

Accuracy = 0.7857

Elapsed time = 1.6356 secs

By utilizing this dataset variant: Standarized Products of Attributes Dataset

The best regression model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.8690

**Logistic Regression:**

Train and Test Split:

Training Linear Regression model with this dataset variant: Original dataset

Accuracy = 0.8525

Elapsed time = 0.6524 secs

Training Linear Regression model with this dataset variant: Normalized Original Dataset

Accuracy = 0.8361

Elapsed time = 0.2838 secs

Training Linear Regression model with this dataset variant: Standardized Original Dataset

Accuracy = 0.8361

Elapsed time = 0.2857 secs

Training Linear Regression model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.8525

Elapsed time = 0.3213 secs

Training Linear Regression model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.8689

Elapsed time = 0.0215 secs

Training Linear Regression model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.8361

Elapsed time = 0.0262 secs

By utilizing the dataset variant on Linear Regression: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 0.8689

K-Fold Cross Validation

Using cross validation

Training Linear Regression model with this dataset variant: Original dataset

Accuracy = 0.8190

Elapsed time = 0.3084 secs

Training Linear Regression model with this dataset variant: Normalized Original Dataset

Accuracy = 0.8690

Elapsed time = 0.1320 secs

Training Linear Regression model with this dataset variant: Standardized Original Dataset

Accuracy = 0.8190

Elapsed time = 0.1295 secs

Training Linear Regression model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.7548

Elapsed time = 0.3545 secs

Training Linear Regression model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.8524

Elapsed time = 0.1880 secs

Training Linear Regression model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.8690

Elapsed time = 0.2218 secs

By utilizing the dataset variant on Linear Regression: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 0.8690

**Support Vector Machine (SVM):**

Train and Test Split:

Training SVC model with this dataset variant: Original dataset

Accuracy = 0.7049

Elapsed time = 0.0087 secs

Training SVC model with this dataset variant: Normalized Original Dataset

Accuracy = 0.8525

Elapsed time = 0.0060 secs

Training SVC model with this dataset variant: Standardized Original Dataset

Accuracy = 0.8361

Elapsed time = 0.0053 secs

Training SVC model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.7049

Elapsed time = 0.0087 secs

Training SVC model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.8525

Elapsed time = 0.0068 secs

Training SVC model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.8197

Elapsed time = 0.0087 secs

By utilizing the dataset variant on SVC: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 0.8525

K-Fold Cross Validation:

Training SVC model with this dataset variant: Original dataset

Accuracy = 0.6524

Elapsed time = 0.0244 secs

Training SVC model with this dataset variant: Normalized Original Dataset

Accuracy = 0.8857

Elapsed time = 0.0227 secs

Training SVC model with this dataset variant: Standardized Original Dataset

Accuracy = 0.8024

Elapsed time = 0.0224 secs

Training SVC model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.6524

Elapsed time = 0.0298 secs

Training SVC model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.8690

Elapsed time = 0.0274 secs

Training SVC model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.8524

Elapsed time = 0.0275 secs

By utilizing the dataset variant on SVC: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 0.8857

**Dense Network:**

Normal Datasets

Average scores for all folds:

Accuracy: 68.99999976158142 (+- 7.858596681051262)

Loss: 0.6013271987438202

Products of Attributes Datasets:

Average scores for all folds:

Accuracy: 64.32258069515228 (+- 12.950968805650003)

Loss: 29.64432077407837

Normalized Dataset:

Average scores for all folds:

Accuracy: 83.50537538528442 (+- 7.239393890382495)

Loss: 0.4158506393432617

Normalized Products of Attributes:

Average scores for all folds:

Accuracy: 83.80645096302032 (+- 5.6383879236024494)

Loss: 0.44195764511823654

Standardized Dataset:

Average scores for all folds:

Accuracy: 83.18279504776001 (+- 4.690242951041032)

Loss: 0.4981158494949341

Standardized Products of Attributes Dataset:

Average scores for all folds:

Accuracy: 80.54838597774506 (+- 6.120243956804952)

Loss: 0.5591489434242248

**Dataset 3:**

**Classifiers:**

**Original Dataset:**

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.5500

Evaluating GaussianNB

Accuracy = 0.6167

Evaluating BernoulliNB

Accuracy = 0.5833

Evaluating Decision Tree classifer

Accuracy = 0.7333

Evaluating Random Forest classifer

Accuracy = 0.6667

By utilizing this dataset variant: Original dataset

The best classification model that performed under this dataset is Decision Tree classifer

Highest Accuracy acheived = 0.7333

Training all the models with this dataset variant: Normalized Original Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.5833

Elapsed time = 0.0086 secs

Evaluating GaussianNB

Accuracy = 0.6333

Elapsed time = 0.0015 secs

Evaluating BernoulliNB

Accuracy = 0.5833

Elapsed time = 0.0032 secs

Evaluating Decision Tree classifer

Accuracy = 0.6167

Elapsed time = 0.0034 secs

Evaluating Random Forest classifer

Accuracy = 0.6667

Elapsed time = 0.1724 secs

By utilizing this dataset variant: Normalized Original Dataset

The best classification model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 0.6667

Training all the models with this dataset variant: Standardized Original Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.6667

Elapsed time = 0.0099 secs

Evaluating GaussianNB

Accuracy = 0.6333

Elapsed time = 0.0030 secs

Evaluating BernoulliNB

Accuracy = 0.7000

Elapsed time = 0.0029 secs

Evaluating Decision Tree classifer

Accuracy = 0.6333

Elapsed time = 0.0028 secs

Evaluating Random Forest classifer

Accuracy = 0.6833

Elapsed time = 0.1616 secs

By utilizing this dataset variant: Standardized Original Dataset

The best classification model that performed under this dataset is BernoulliNB

Highest Accuracy acheived = 0.7000

Training all the models with this dataset variant: Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.5500

Elapsed time = 0.0131 secs

Evaluating GaussianNB

Accuracy = 0.6000

Elapsed time = 0.0043 secs

Evaluating BernoulliNB

Accuracy = 0.5000

Elapsed time = 0.0026 secs

Evaluating Decision Tree classifer

Accuracy = 0.6667

Elapsed time = 0.0118 secs

Evaluating Random Forest classifer

Accuracy = 0.6667

Elapsed time = 0.2290 secs

By utilizing this dataset variant: Products of Attributes Dataset

The best classification model that performed under this dataset is Decision Tree classifer

Highest Accuracy acheived = 0.6667

Training all the models with this dataset variant: Normalized Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.5833

Elapsed time = 0.0121 secs

Evaluating GaussianNB

Accuracy = 0.5833

Elapsed time = 0.0038 secs

Evaluating BernoulliNB

Accuracy = 0.5000

Elapsed time = 0.0039 secs

Evaluating Decision Tree classifer

Accuracy = 0.7000

Elapsed time = 0.0091 secs

Evaluating Random Forest classifer

Accuracy = 0.6667

Elapsed time = 0.2330 secs

By utilizing this dataset variant: Normalized Products of Attributes Dataset

The best classification model that performed under this dataset is Decision Tree classifer

Highest Accuracy acheived = 0.7000

Training all the models with this dataset variant: Standarized Products of Attributes Dataset

Evaluating K-Nearest Neighbors Classifier

Accuracy = 0.6333

Evaluating GaussianNB

Accuracy = 0.5833

Evaluating BernoulliNB

Accuracy = 0.6667

Evaluating Decision Tree classifer

Accuracy = 0.6500

Evaluating Random Forest classifer

Accuracy = 0.6833

By utilizing this dataset variant: Standarized Products of Attributes Dataset

The best classification model that performed under this dataset is Random Forest classifer

Highest Accuracy acheived = 0.6833

**Logistic Regression:**

**Default Logistic Regression:**

Logistic Regression with default parameters

Accuracy 0.6833

Precision 0.875

Recall 0.28

F1 0.4242

[Parallel(n\_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.

[Parallel(n\_jobs=-1)]: Done 1 out of 1 | elapsed: 0.0s finished

**Cross Validation Logistic Regression:**

Using cross validation

Training Linear Regression model with this dataset variant: Original dataset

Accuracy = 0.6333

Training Linear Regression model with this dataset variant: Normalized Original Dataset

Accuracy = 0.6500

Training Linear Regression model with this dataset variant: Standardized Original Dataset

Accuracy = 0.6500

Training Linear Regression model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.6000

Training Linear Regression model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.6333

Training Linear Regression model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.6167

By utilizing the dataset variant on Linear Regression: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 0.6500

**Support Vector Machine (SVM):**

**Default SVC:**

Training SVC model with this dataset variant: Original dataset

Accuracy = 0.5833

Elapsed time = 0.0078 secs

Training SVC model with this dataset variant: Normalized Original Dataset

Accuracy = 0.6000

Elapsed time = 0.0049 secs

Training SVC model with this dataset variant: Standardized Original Dataset

Accuracy = 0.7000

Elapsed time = 0.0042 secs

Training SVC model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.5833

Elapsed time = 0.0084 secs

Training SVC model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.6000

Elapsed time = 0.0081 secs

Training SVC model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.6833

Elapsed time = 0.0081 secs

By utilizing the dataset variant on SVC: Standarized Products of Attributes Dataset

Highest Accuracy acheived = 0.7000

**SVC Cross Validation:**

using cross validation

Training SVC model with this dataset variant: Original dataset

Accuracy = 0.5833

Elapsed time = 0.0242 secs

Training SVC model with this dataset variant: Normalized Original Dataset

Accuracy = 0.6500

Elapsed time = 0.0211 secs

Training SVC model with this dataset variant: Standardized Original Dataset

Accuracy = 0.6333

Elapsed time = 0.0208 secs

Training SVC model with this dataset variant: Products of Attributes Dataset

Accuracy = 0.5667

Elapsed time = 0.0269 secs

Training SVC model with this dataset variant: Normalized Products of Attributes Dataset

Accuracy = 0.6333

Elapsed time = 0.0289 secs

Training SVC model with this dataset variant: Standarized Products of Attributes Dataset

Accuracy = 0.6500

Elapsed time = 0.0274 secs

By utilizing the dataset variant on SVC: Standarized Products of Attributes Dataset

Highest Accuracy achieved = 0.6500

**SVC Best Parameters:**

Original Data:

SVC Model with kernel = sigmoid had a MAX\_Accuracy = 0.65

Total Size = 360, Half size= 180.0, Quarter size= 90.0

Index of the Max Accuracy is in = 351

MAX\_Accuracy = 0.65

Max Accuracy with no data augmentation = 0.65

Kernel linear had a max accuracy of 0.55

Kernel rbf had a max accuracy of 0.583

Kernel sigmoid had a max accuracy of 0.65

Normalized Data:

SVC Model with kernel = sigmoid had a MAX\_Accuracy = 0.6667

Total Size = 360, Half size= 180.0, Quarter size= 90.0

Index of the Max Accuracy is in = 195

MAX\_Accuracy = 0.6667

Max Accuracy with Normalization on data = 0.667

Kernel linear had a max accuracy of 0.633

Kernel rbf had a max accuracy of 0.667

Kernel sigmoid had a max accuracy of 0.667

Standardized Data:

SVC Model with kernel = sigmoid had a MAX\_Accuracy = 0.6833

Total Size = 360, Half size= 180.0, Quarter size= 90.0

Index of the Max Accuracy is in = 180

MAX\_Accuracy = 0.7

Max Accuracy with Standardization on data = 0.7

Kernel linear had a max accuracy of 0.65

Kernel rbf had a max accuracy of 0.7

Kernel sigmoid had a max accuracy of 0.683

**Dense Network:**

Original Dense Model:

Average scores for all folds:

> Accuracy: 65.58620810508728 (+- 10.570363744521982)

> Loss: 0.9037590384483337

Original Dense Model Normalized:

> Fold 10 - Loss: 0.6088472604751587 - Accuracy: 68.96551847457886%

------------------------------------------------------------------------

Average scores for all folds:

> Accuracy: 69.89655196666718 (+- 5.1732950011033045)

> Loss: 0.5697469174861908

Dense Model Normalized Products:

> Fold 10 - Loss: 0.5286528468132019 - Accuracy: 75.82417726516724%

------------------------------------------------------------------------

Average scores for all folds:

> Accuracy: 80.83253622055054 (+- 4.413672188978272)

> Loss: 0.47131111919879914

Original Dense Model Standardized:

> Fold 10 - Loss: 0.7224110960960388 - Accuracy: 75.86206793785095%

------------------------------------------------------------------------

Average scores for all folds:

> Accuracy: 71.25287413597107 (+- 8.671793672867889)

> Loss: 0.6662678569555283

Dense Model Standardized Products::

> Fold 10 - Loss: 1.1460087299346924 - Accuracy: 72.52747416496277%

------------------------------------------------------------------------

Average scores for all folds:

> Accuracy: 78.65504026412964 (+- 6.47940541376248)

> Loss: 0.8351862132549286

**Discussion of results**

The best accuracy results achieved on Dataset 1 was an accuracy of 89% with dense models using standardized dataset and using kfold.

The best accuracy results achieved on Dataset 2 was an accuracy of 88.5% with standardized products of attributes with SVC models.

The Best accuracy results achieved on Dataset 3 was an accuracy of 80 % with normalized attributes with SVC models.

**Conclusions**

What greatly improved all accuracies and mse in all three datasets were the utilization of k-fold cross validation, using products of attributes, and standardizing the entire benefit to help combat the influence of attribute’s varying values. By being able to experience k-folds' time consuming learning and validation process in action, we were able to learn why we always utilized train and test set validation over k-fold with the type of datasets we were utilizing throughout the semester. K-fold was an acceptable validation option in all 3 cases of the datasets due to their size.