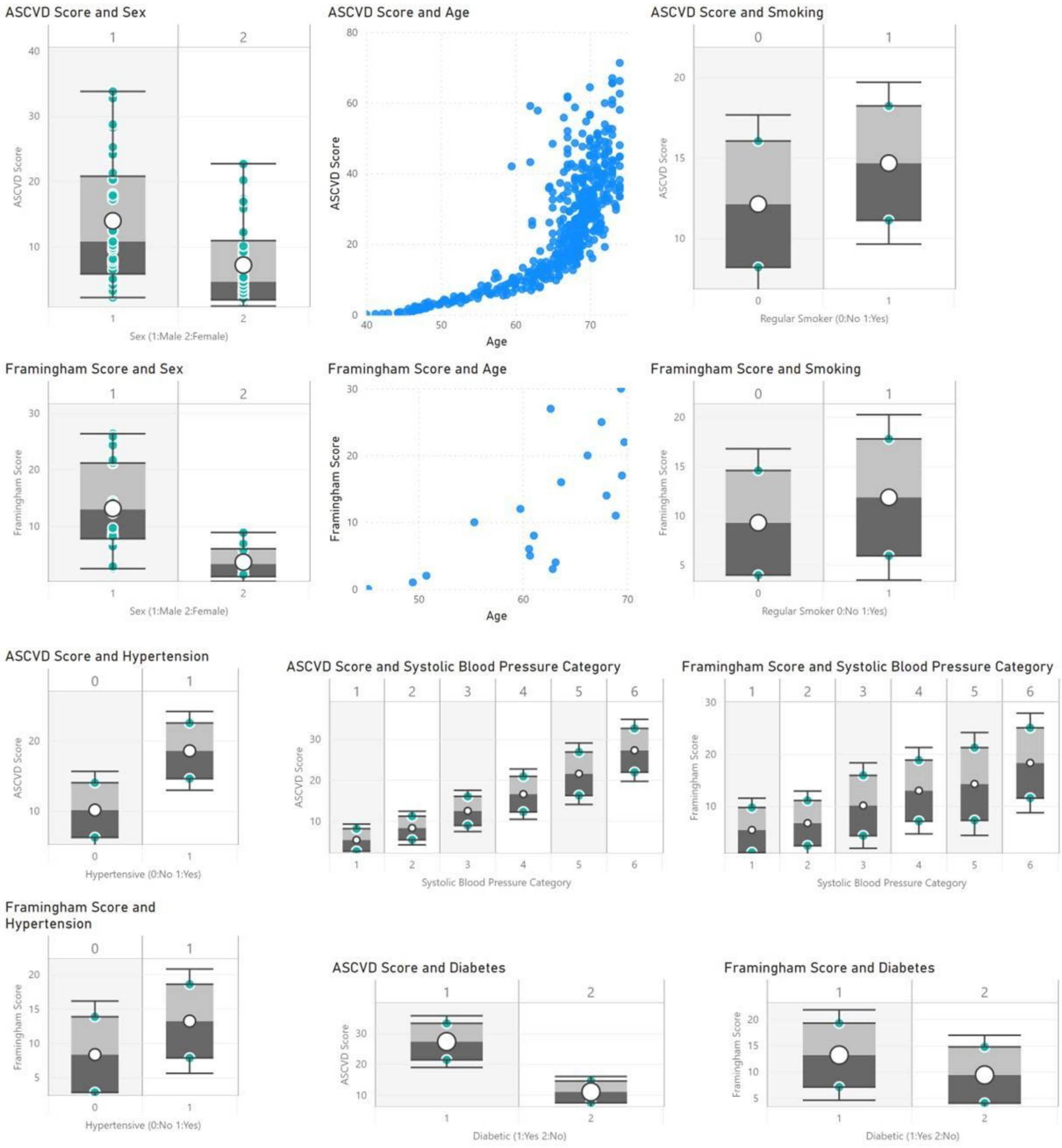
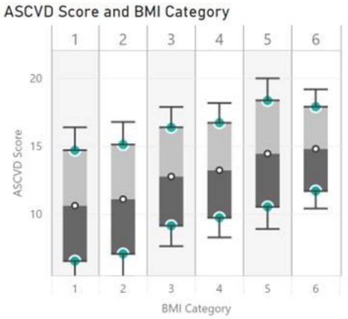
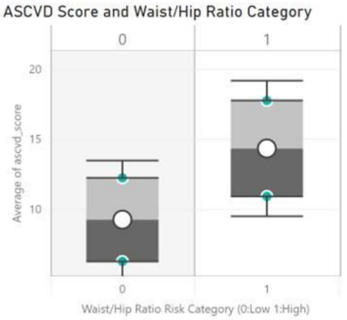
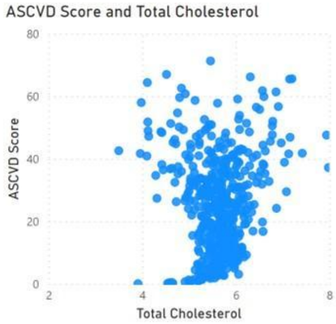
Part A - Data summarization, data preprocessing and feature selections

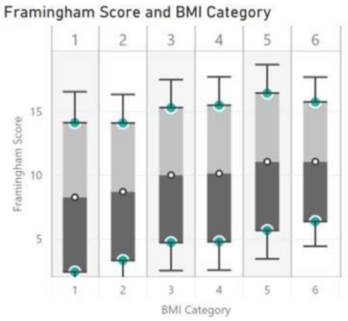
We explored the relationship between attributes in several dimensions and the cardiac risk factor scores (ASCVD Score and Framingham Score). The below scatter and box plots were prepared in Power BI using our PostgreSQL database to summarize the relationships:

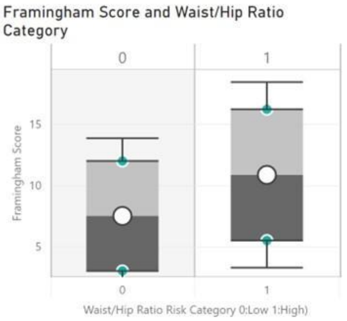


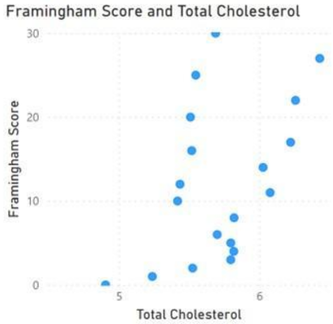

ASCVD Score and BMI Category

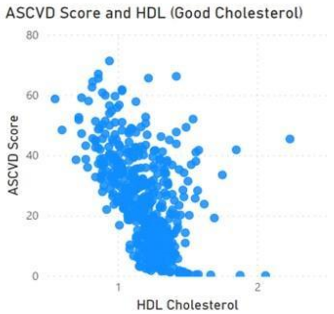

ASCVD Score and Waist/Hip Ratio Category

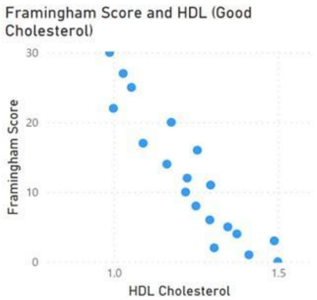

ASCVD Score and Total Cholesterol


Framingham Score and BMI Category


Framingham Score and Waist/Hip Ratio Category


Framingham Score and Total Cholesterol


ASCVD Score and HDL (Good Cholesterol)


Framingham Score and HDL (Good Cholesterol)

From the above figures, it is apparent that both cardiac risk scores can be predicted using sex, age, smoking status, hypertensive status, systolic blood pressure category, diabetes status, BMI (Body Mass Index) category, waist/hip ratio category, total blood cholesterol, and blood HDL cholesterol.

We decided to use sex, age, smoking, hypertensive status, systolic blood pressure category, diabetes status, BMI category, and waist/hip ratio category as features for predicting cardiac risk scores. From these features, sex, age, hypertensive status, diabetes status, and smoking status were used for calculating ASCVD Score when creating the fact table as described in the previous deliverable report. Similarly, for Framingham Scores, sex, age, and regular smoking status were used for calculating the scores. Thus, diabetes status, BMI category, waist/hip ratio category, systolic blood pressure category, and hypertensive status are new features (not originally used for calculation) that can predict the Framingham score. For ASCVD score, BMI category, waist/hip ratio category, and systolic blood pressure category are new features that can predict the score. Total cholesterol and HDL level which were originally used in calculating both risk scores were not chosen as features. This type of modeling can be useful in the clinical setting as it allows physicians to predict cardiac risk scores based on patient history and physical examination without the need for invasive blood tests (measuring total cholesterol and HDL).

When calculating cardiac risk scores, all necessary patient attributes (risk factors) should be present. Otherwise, the score cannot be calculated (is not valid). When creating the fact table, ASCVD or Framingham scores were coded as 999 when one of the necessary patient attributes was missing or unknown. Therefore, when querying the database, rows with risk scores equal to 999 were excluded since a cardiac risk score produced through imputation will not be valid. However, patients with Framingham and ASCVD scores could have missing/unknown values for regular smoker status (column "regsmok," missing values encoded as 9 in the database), systolic blood pressure category (column "syscat," missing values encoded as 9 in the database), diabetes status (column "diabet," missing/unknown values encoded as 7 or 9 in the database), and BMI or waist/hip ratio categories (columns "bmicat" and "whrcat," missing/unknown values encoded as 7 or 9 in the database). In these cases, when creating the data frame from the rows of the SQL join query, missing attributes (with codes 7 or 9 in the database) were encoded as NaN. Next, the columns (features) of the data frame were transformed using Min Max scaling (MinMaxScaler class of the preprocessing library), and missing (NaN) values were imputed using the KNNImputer class of the impute library.

BMI category, waist/hip ratio category, systolic blood pressure category, smoking status, sex, and diabetes status were already encoded using one-hot or one-cold encodings in the database as follows:

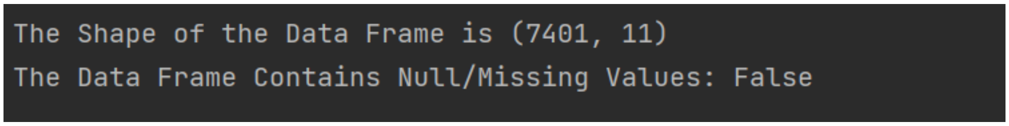
- Sex is encoded as 1 for male and 2 for female.
- Hypertensive status is encoded as 0 for non-hypertensive and 1 for hypertensive.
- Systolic blood pressure category is encoded from 1 to 6.
- BMI category is encoded from 1 to 6.
- Waist/hip ratio category is encoded as 0 for low-risk and 1 for high risk.
- Regular smoker status is encoded as 0 for non-smokers and 1 for regular smokers.
- Diabetes status is encoded as 1 for diabetic and 2 for non-diabetic.

Part B - Classification (Supervised Learning)

Cardiac risk scores were originally calculated and recorded as numeric values in the database. Therefore, for classification, ASCVD and Framingham risk scores were classified based on medical guidelines as follows:

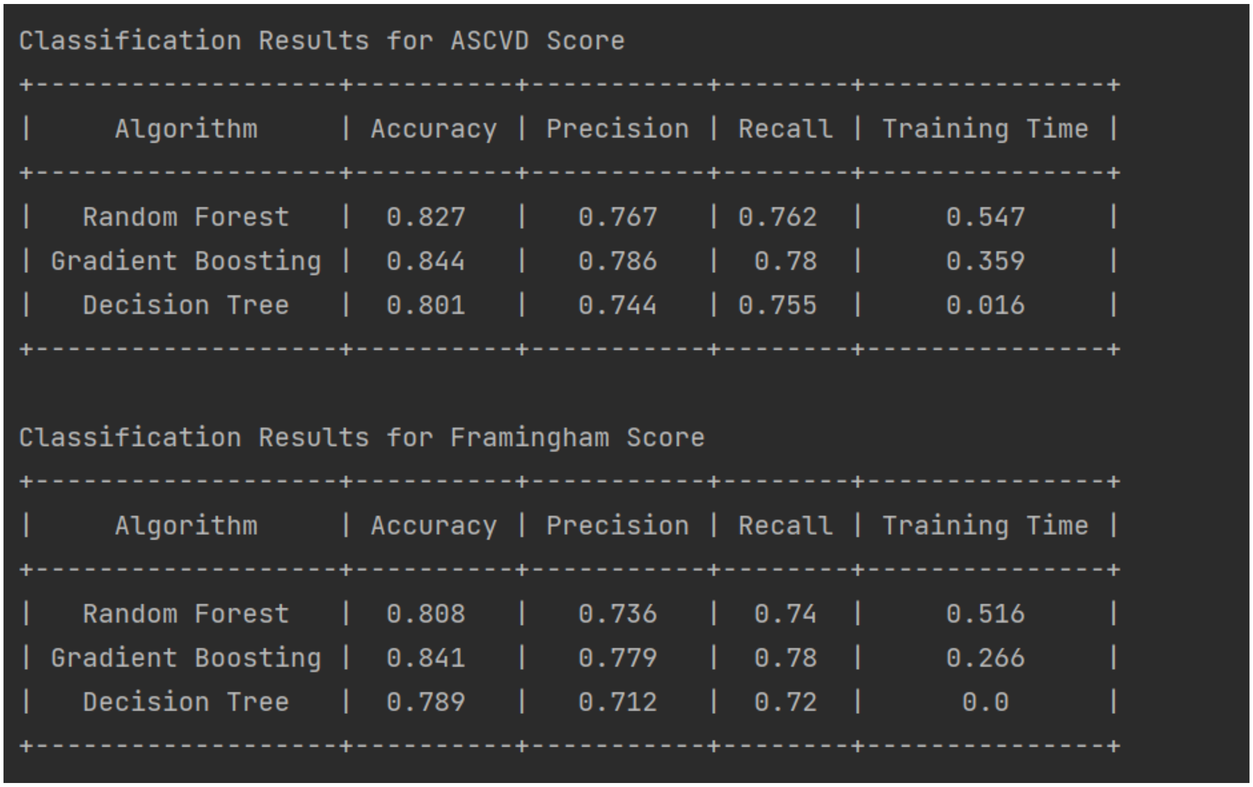
- ASCVD Score: <5: low risk, 5-7.5: borderline, 7.5-9: intermediate, >20: high
- Framingham Score: <10: low, 10-20: intermediate, >20: high
- For the purpose of detecting outliers (using the one-class SVM algorithm), the ASCVD score was classified as 0 (score under 20, low, borderline, and intermediate risks) and 1(score above 20, high risk).

The resulting data frame (which was normalized an imputed as described in part A and was used for training and testing of the learning algorithms) had 7401 rows with no missing or null values:



```
The Shape of the Data Frame is (7401, 11)
The Data Frame Contains Null/Missing Values: False
```

Decision Tree, Gradient Boosting and Random Forest classifier algorithms were trained to predict ASCVD and Framingham score classes. For each algorithm, the data frame was split into train and test sets using train_test_split(). The algorithms were trained and then tested in terms of accuracy, precision, recall, and training time. The tables below summarize the findings for each score and algorithm:



```
Classification Results for ASCVD Score
+-------------------+----------+-----------+--------+---------------+
|     Algorithm     | Accuracy | Precision | Recall | Training Time |
+-------------------+----------+-----------+--------+---------------+
|   Random Forest   |  0.827   |   0.767   | 0.762  |     0.547     |
| Gradient Boosting |  0.844   |   0.786   |  0.78  |     0.359     |
|   Decision Tree   |  0.801   |   0.744   | 0.755  |     0.016     |
+-------------------+----------+-----------+--------+---------------+

Classification Results for Framingham Score
+-------------------+----------+-----------+--------+---------------+
|     Algorithm     | Accuracy | Precision | Recall | Training Time |
+-------------------+----------+-----------+--------+---------------+
|   Random Forest   |  0.808   |   0.736   |  0.74  |     0.516     |
| Gradient Boosting |  0.841   |   0.779   |  0.78  |     0.266     |
|   Decision Tree   |  0.789   |   0.712   |  0.72  |      0.0      |
+-------------------+----------+-----------+--------+---------------+
```

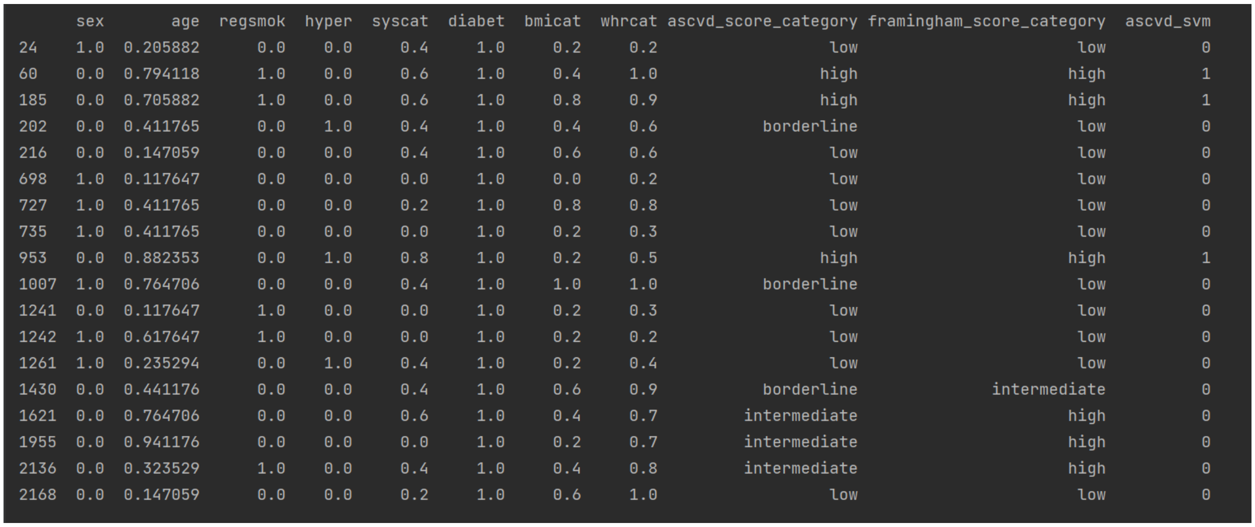
From the above results, it is evident that a patient’s ASCVD risk score category can be predicted with fairly high accuracy based on sex, age, smoking, hypertensive status, systolic blood pressure category, diabetes status, BMI category, and waist/hip ratio category. Precision and recall are slightly lower than accuracy but still above 0.75. This type of modeling can be useful in the clinical setting since it allows clinicians to predict a patient’s risk score category based on physical examination and patient history, without the need for laboratory tests. It is noteworthy that I also tried using Decision Tree, Gradient Boosting and Random Forest regression models for predicting (the actual numeric) ASCVD scores and the results were more accurate and precise (around 0.95-0.96). However, since the deliverable asks for classification tasks, the classification code and results are submitted.

For the Framingham score, accuracy, precision, and recall are lower than the ASCVD score but still acceptable (around 0.75-0.8). I also tried regression models to predict the numeric value of Framingham scores and the results were significantly more accurate and precise (around 0.92). However, the results for classification tasks are submitted according to instructions.

For both scores, random forest training time is the longest. Gradient boosting has shorter training time compared to random forest with better results in terms of accuracy, precision, and recall. Decision tree training time is significantly shorter than the other two models, but accuracy, precision and recall are also lower.

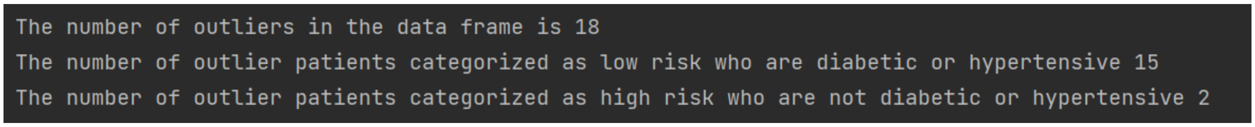
Part C - Detecting Outliers

For this task, the ASCVD score was classified as 0 (score under 20, low, borderline, and intermediate risks) and 1 (score above 20, high risk). The one-class SVM algorithm was used to detect outliers in the data frame. Out of the 7401 rows, 18 ASCVD scores were detected as outliers. The screenshot below shows the identified rows:



```
      sex       age  regsmok  hyper  syscat  diabet  bmicat  whrcat ascvd_score_category framingham_score_category  ascvd_svm
24    1.0  0.205882      0.0    0.0     0.4     1.0     0.2     0.2                  low                       low          0
60    0.0  0.794118      1.0    0.0     0.6     1.0     0.4     1.0                 high                      high          1
185   0.0  0.705882      1.0    0.0     0.6     1.0     0.8     0.9                 high                      high          1
202   0.0  0.411765      0.0    1.0     0.4     1.0     0.4     0.6           borderline                       low          0
216   0.0  0.147059      0.0    0.0     0.4     1.0     0.6     0.6                  low                       low          0
698   1.0  0.117647      0.0    0.0     0.0     1.0     0.0     0.2                  low                       low          0
727   1.0  0.411765      0.0    0.0     0.2     1.0     0.8     0.8                  low                       low          0
735   1.0  0.411765      0.0    0.0     0.0     1.0     0.2     0.3                  low                       low          0
953   0.0  0.882353      0.0    1.0     0.8     1.0     0.2     0.5                 high                      high          1
1007  1.0  0.764706      0.0    0.0     0.4     1.0     1.0     1.0           borderline                       low          0
1241  0.0  0.117647      1.0    0.0     0.0     1.0     0.2     0.3                  low                       low          0
1242  1.0  0.617647      1.0    0.0     0.0     1.0     0.2     0.2                  low                       low          0
1261  1.0  0.235294      0.0    1.0     0.4     1.0     0.2     0.4                  low                       low          0
1430  0.0  0.441176      0.0    0.0     0.4     1.0     0.6     0.9           borderline              intermediate          0
1621  0.0  0.764706      0.0    0.0     0.6     1.0     0.4     0.7         intermediate                      high          0
1955  0.0  0.941176      0.0    0.0     0.0     1.0     0.2     0.7         intermediate                      high          0
2136  0.0  0.323529      1.0    0.0     0.4     1.0     0.4     0.8         intermediate                      high          0
2168  0.0  0.147059      0.0    0.0     0.2     1.0     0.6     1.0                  low                       low          0
```

Browsing the outlier rows shows that 15 rows are diabetic or hypertensive patients categorized as low, borderline, or intermediate risk (ascvd_svm =0). Furthermore, 2 rows are non-diabetic, non-hypertensive patients categorized as high risk (ascvd_svm=1):



```
The number of outliers in the data frame is 18
The number of outlier patients categorized as low risk who are diabetic or hypertensive 15
The number of outlier patients categorized as high risk who are not diabetic or hypertensive 2
```

This can be because diabetic or hypertensive patients who were categorized as low risk (score below 20) had low cholesterol and high HDL (good cholesterol) and were non-smokers or were physically fit (low BMI category and waist/hip ratio). Similarly, non-diabetic and non-hypertensive patients who were categorized as high risk (ASCVD above 20) could have had high Cholesterol or low HDL (some hereditary diseases can result in abnormal lipid levels in otherwise healthy people) or had high BMIs and waist/hip ratios. Overall, this indicates that the model can be fine-tuned and improved by adding more features based on physical exam and family history.