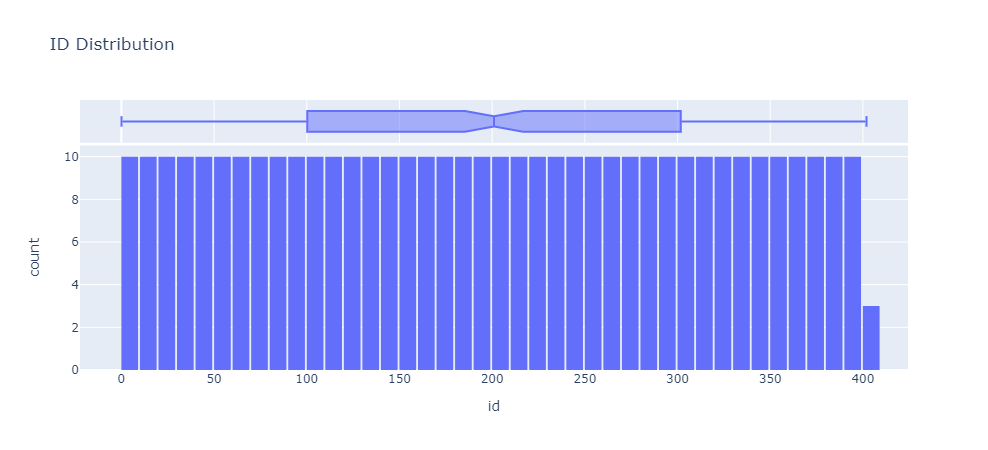
**Evaluating control variables**

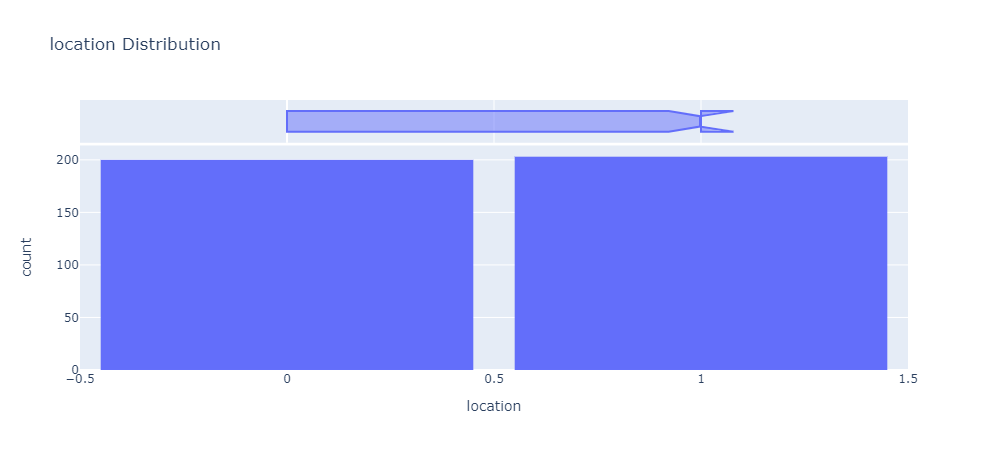
The dataset contains 19 features collected from more than 400 patients in a study that intended to investigate the prevalence of several cardiovascular diseases among them: obesity, diabetes, and heart disease. In this project, the main intend is to study the prevalence of type II diabetes (diabetes mellitus) in the patients. We will hence need to select specific variables from the dataset that can best help us create predictor classifier machine learning on the likelihood that a patient has type II diabetes. However, it is important to note that some of these cardiovascular diseases have a close relation. According to John Hong (1997), type II diabetes is strongly related with obesity. He also suggested that the hip-waist ratio is important in diagnosing diabetes and heart disease. Glycosylated hemoglobin is a sure measure of diabetes, where patients who record the value higher than 7 are diagnosed positively for diabetes. The variables that most related to diabetes from dataset were selected to be; total cholesterol, stabilized glucose, high density lipoprotein, cholesterol/hdl ratio, glycosylated hemoglobin, age, gender, height, weight, waist, hip, and systolic blood pressure measures.

**Denoising and feature engineering**

The project intended to focus on type II diabetes mellitus, hence variables that had little or no positive correlation to this thesis, were eliminated to better the performance of the classifier models. Variables such as location and customer ID were noise and were deleted; this was because they had no statistical distribution and recorded very low correlation to the target variable we intended to predict. Other variables that deleted after fitting the model included the height, systolic blood pressure and frame. Elimination of these variables slightly improved the performance of the classifier machine learning models.

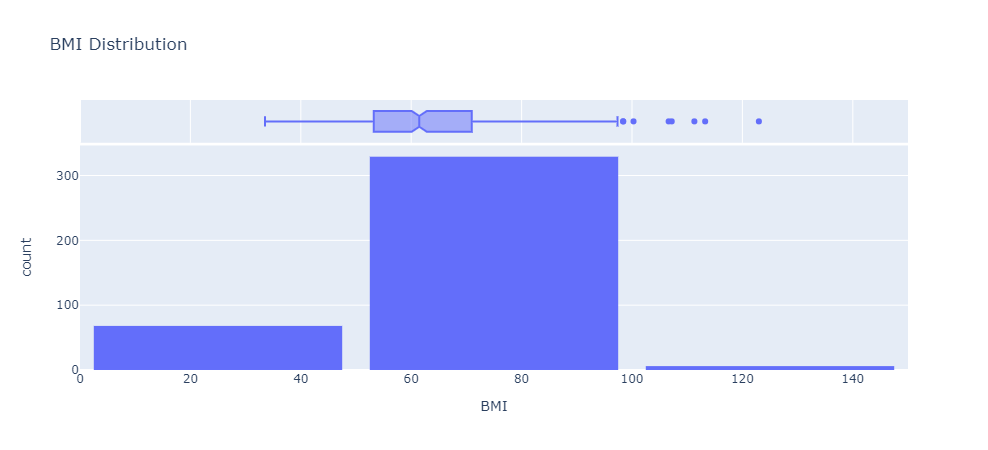


Patient ID distribution



Patient Location histogram

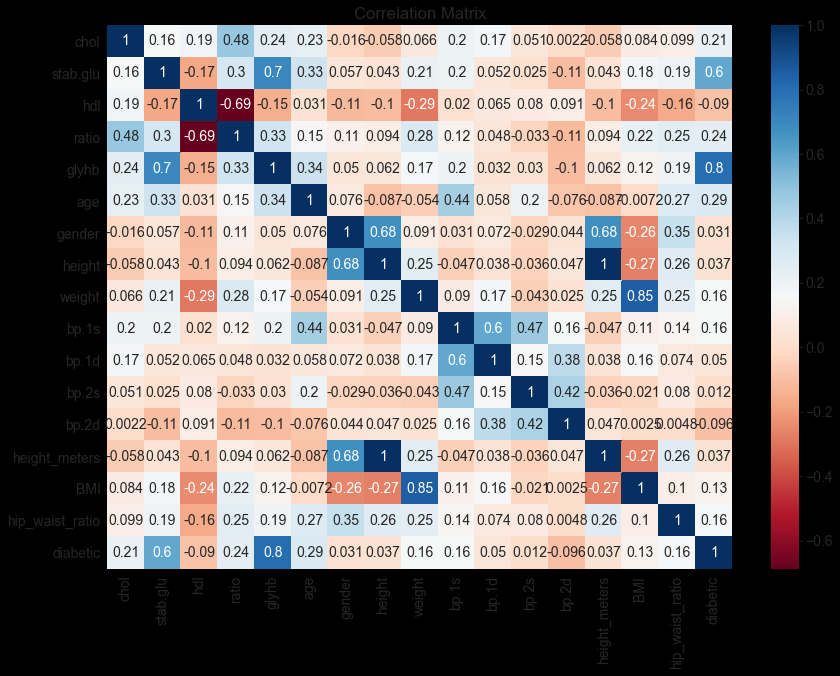
Feature engineering was carried on the provided variables to create better and more understandable features for fitting the machine learning model. These features include the BMI which was calculated from the weight and the height of the patient. The new feature had a higher correlation to the target variable than both original variables. Hip-ratio was also another contrived feature that had a better correlation. The height variable was converted to meters from inches to better calculate the BMI values. Finally, a target feature was extracted from the glycosylated hemoglobin feature to create a new feature – diabetic. It is because of this feature engineering, that some of the original variables were deleted in favor of the contrived features that had a better correlation.



A normal distribution for the engineered BMI feature

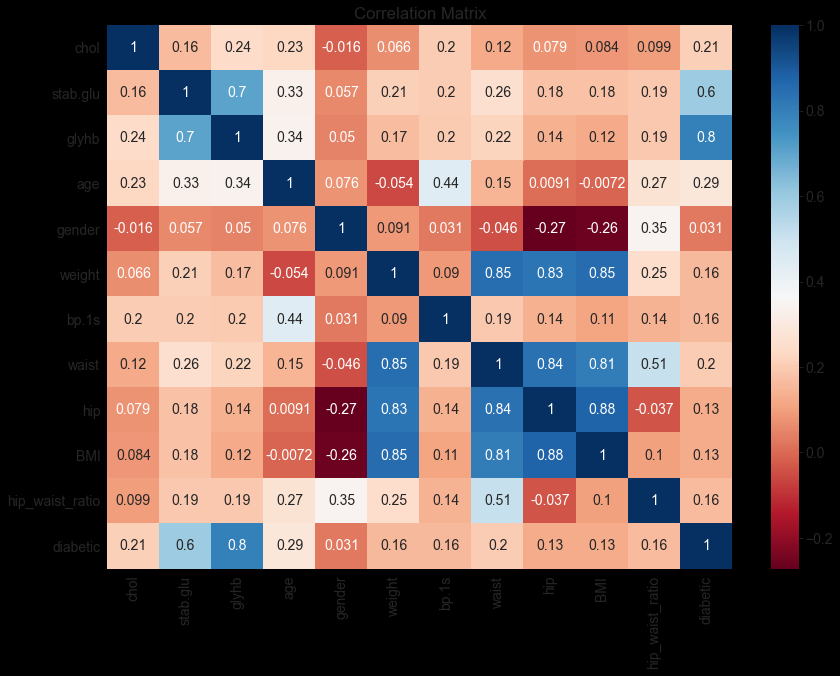
**Feature selection**

To properly asses the importance of the provided (selected) features to our modelling process, the classifier models were fitted with the entire dataset containing all the variables and with a dataset that contained on a few variables that had the best correlation with the target feature.



Correlation plot for the entire dataset

The average improvement margin in the performance of the classifier was not very huge. However, the Support Vector Machine (SVM) classifier recorded a higher improvement in performance compared to the Random Forest (RF) classifier and the Extreme Gradient Boosting (XGB) classifier. Nonetheless, the two later classifiers generally did outperform the SVM classifier. In relation the general but slight improvement, this is likely because of the models (RF and XGB) having to make lesser traverses in the data to find and rank classification patterns. As for the SVM model, fewer features enable easier evaluation of the classification hyperplane and reduced trade off between the variance and the bias of the model.



Correlation plot for the selected features.

**Train-test set ration**

To split the data into training and testing datasets, a splitting ratio of 7:3 (70% training data and 30% testing data) was used on the first cycle of training, and a ratio of 8:2 (80% training data and 20% testing data) used on the second training cycle. There was no notable significant in the general accuracy of the RF and XGB classifiers whereas the SVM classifier did score relatively better when provided with more training data. The possible reason for the insignificant change for the two ensemble classifiers can be alluded to the general size of the dataset. Since the dataset was relatively small, a 10% variation in the training and testing data is less likely to influence the outcomes of the models. A higher variation in the training and testing ratio would however most definitely affect the performance of any of the classifier model.

**Classifiers used**

For this project, three classifiers were selected to predict that a patient had type II diabetes mellitus. These are Support Vector Machine (SVM), Random Forest (RF) and Extreme Gradient Boosting algorithm (XGB) classifiers. These three classifiers are robust and the underlying architecture of each one of them makes them suitable for classification problems with both small and big data.

The Support Vector Machine classifier, suitable for both classification and regression problems, uses a hyperplane in a 2-dimensions space to separate classes based on their distance from the set hyperplane. If fitted and tuned properly, these classifiers are sure to score well many classification problems. They also are easier to fit and take a relatively short period of time on training. The regularization hyperparameter (C) for SVM classifiers, is significant in preventing overfitting of the model and maintaining a proper balance between the variance and bias of the model.

Random Forest (RF) classifiers are ensemble machine learning that optimize on the logic of utilizing multiple builder (tree classifier) models that fit and classify the data in numerous tree traverses and a vote is taken on the highest instances of a particular values for each target class and classified accordingly. This architecture makes the RF a robust model most of classification problems. The RF has numerous hyperparameters than can be adjusted depending on the kind of problem, the required (acceptable) accuracy and training time. The class weight can be adjusted to prevent bias for one class in the case of imbalanced data where random sampling is not a possible option. Other key hyperparameters for this ensemble classifier include number of estimators, maximum traversing depth, maximum features, bootstrap, maximum samples and so on.

The Extreme Gradient Boosting (XGB) classifier is an ensemble classifier that advances on the scoring logic and architecture of the Random Forest (RF) classifier. However, contrary to the RF classifiers which use a vote to determine the classification of a variable, XGB classifiers employ the use of weighted mean for each individual training instance. This mean is then used to calculate an overall weighted mean that is used to classify the class. Extreme Gradient Boosting algorithm sequentially minimize the gradient of the loss function at each training cycle; hence they perform very well when compared to most classifier algorithms.

In our project, the RF and XGB classifiers perform far better than the SVM classifier – this can generally be attributed to their ensemble architecture. The two classifiers however do not outperform each other except for the runtime. They score all similar accuracies (sometimes with very insignificant variations) on all the evaluation metrics used in the project. The Extreme Gradient Boosting classifier does however outperform the RF on training time – although insignificantly, this can be used as good determinant for the best performing model if the models are fitted on big data. Also notable, individual hyperparameter for the two best performing classifiers can also be used to narrow down on the best performing classifier. The SVM classifier, although scoring an accuracy above 90%, still lagged the ensemble classifiers. This can be attributed to its simplicity and general inability to discover more patterns and learn the training dataset much better.

**Classifier evaluation metric**

To evaluate the classifier models three metrics were used; Root Mean Squared Error (RMSE), Receiver Operating Curve – Area Under Curve (ROC-AUC) and classification report (which encompasses Precision, Recall and F1 score. The RMSE metric is generally a good evaluation for numerical models. It assesses the how the model tried to minimize loss in its prediction of unknown data. A significantly low RMSE implies the model got most of the predictions correctly. ROC-AUC is a metric used to binary probability predictors by assessing the sensitivity and specificity of the classification predictions. This is a more nuanced metric since it assesses the logical valuation of True positive over True Negative – it assigns more weight to values it knows to be true values and lesser weight to values knows are surely false. This makes it a better assessment metric than other metrics that assign equal weight for True Positives and True Negatives. This a medical science classification problem, this was seen as a good metric since it gives a better depiction of real-world classification and not a 50-50 idealistic classification that is likely to be assumed in non-medical research projects. Finally, the project assesses the models using Precision, Recall and F1 score. These metrics are as well associated with the specificity and sensitivity of the models. They take the form of ratios between the True Positives, False Positives, True Negatives and False Negatives.

**Confounding factors and threats to validity of the results**

The project did note that despite the very exemplary performance of the classification models that were trained, other factors need to be taken into consideration aside from model performance. Diabetes diagnosis is a complex and long-term process and hence requires more data collected over longer periods of times to better understanding the relationships; variances and deviations in patient data.

A patient with a glycosylated hemoglobin value of 6.99 (less than 7) will be classified by the models as non-diabetic this is even though this patient is medically pre-diabetic. In addition, the data used in the project was for patients based in only a particular geographical location; other lifestyle-based demographic factors that might predispose people in that area to cardiovascular diseases are not accounted for in the project.

From a technical vantage point, the data preparation process and training process can be changed to alter the performance of the model. For instance, used of other data encoders, imputation strategies, further feature engineering and use of different hyperparameters tuning strategies such as GridSearch and RandomSearch can impact the performance of each of the classifiers.