Alex Neibart CS225 12/15/2022

Diabetes Classification

Introduction:

There are a number of diagnostic criteria that may indicate the presence of diabetes. Although these criteria can already be interpreted by doctors, it may still be beneficial to create a model capable of identifying individuals who are likely to have diabetes, to ensure patients in need of a diagnosis know to seek out proper medical care.

To that end, this project will determine the effectiveness of various classification algorithms in detecting diabetic patients.

Methodology:

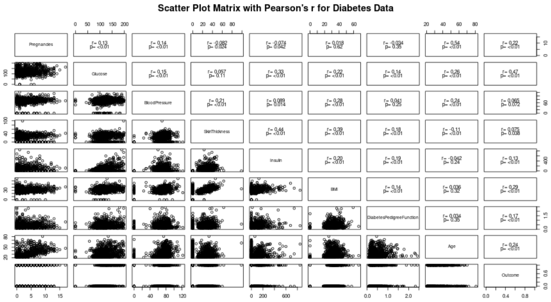
The analysis of classification algorithms was performed according to the CRISP-DM framework, which consists of five phases: business understanding, data understanding, data preparation, evaluation, and deployment.

Business Understanding:

The goal of this project was to create models that classify diabetic and non-diabetic patients from an existing dataset.

Data Understanding:

Figure 1: Scatter Plot Matrix with Pearson's r



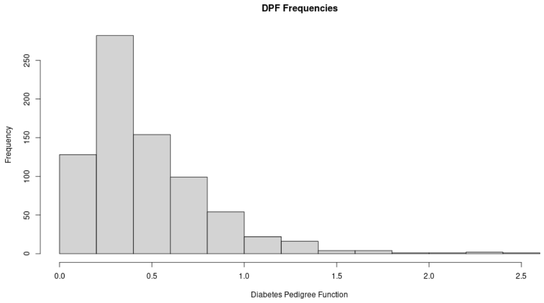
The dataset used in this project contains information from 768 women at least 21 years old or older, and of Pima Indian descent. For each of these patients, it provides nine features:

Outcome: Whether or not the patient has diabetes.

Age: The patient's age in years.

DiabetesPedigreeFunction (DPF): A score which predicts the likelihood that the patient will have diabetes based on their family history.

Figure 2: DPF Frequency Histogram



BMI: The patient's body mass index in kg/m^2.

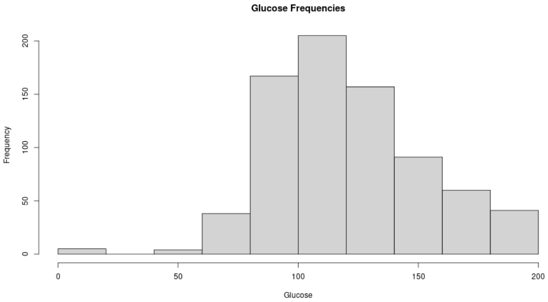
Insulin: The patient's 2-hour serum insulin in muU/ml.

SkinThickness: The patient's tricep skin fold thickness in mm.

BloodPressure: The patient's diastolic blood pressure in mmHg.

Glucose: The patient's plasma glucose concentration as determined by an oral glucose tolerance test.

Figure 3: Glucose Frequency Histogram



Pregnancies: The number of times the patient has been pregnant.

Of the 768 individuals in the dataset, 268 had diabetes and 500 did not.

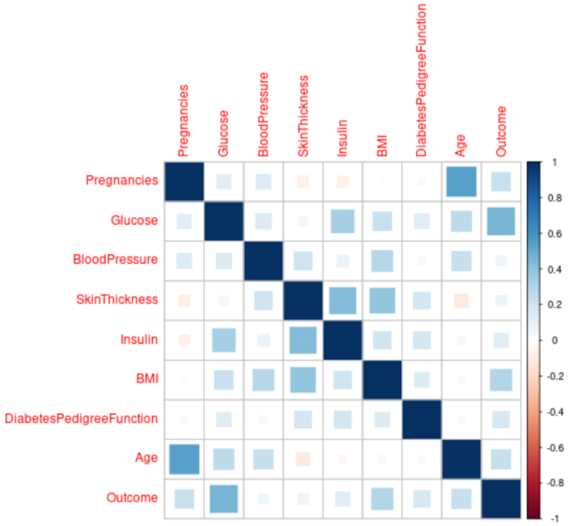
Figure 4: Summary of data

Feature	Mean	Minimum	First Quartile	Median	Third Quartile	Maximum
Age	33.24	21	24	29	41	81
DPF	0.4719	0.078	0.2437	0.3725	0.6262	2.42
BMI	31.99	0	27.3	32	36.6	67.1
Insulin	79.8	0	0	30.5	127.2	846
Skin Thickness	20.54	0	0	23	32	99
Blood Pressure	69.11	0	62	72	80	122
Glucose	120.9	0	99.0	117	140.2	199.0
Pregnancies	3.84	0	1	3	3	17

Although the overall quality of this data is good, several features have minimum values of 0 even where that would not make logical sense. These measurements are assumed to be substitutes for missing data.

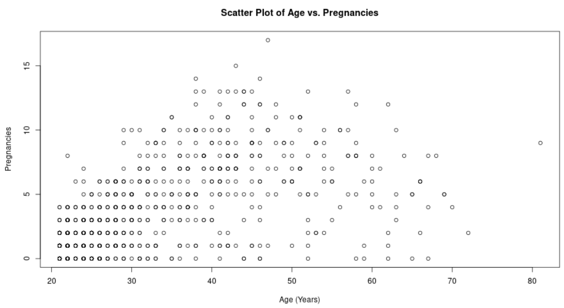
In spite of the sensitive nature of medical data, this dataset lacks identifying characteristics and should not pose any privacy concerns.

Figure 5: Correlation Plot



As is visible in figure 2, the most significant correlation between any two features is a positive correlation between a patient's age and the number of times they have been pregnant.

Figure 6: Scatter Plot of Age vs. Pregnancies:



Additionally, glucose has a positive correlation with the presence of diabetes.

Data Preparation:

Feature Derivation:

To account for missing values, glucose, blood pressure, skin thickness, insulin, and BMI were each replaced with equivalent categorical features. Patients who previously had a value of 0 for one of those features received a new value of "N/A."

Other patients were sorted into equal-frequency bins based on which quartiles they belonged to. To avoid distortion created by the missing values, the calculations which determined the ranges of these bins excluded data points with a value of 0.

Feature Selection:

Several feature selection processes were performed on the revised version of the dataset. These processes selected the following as features of interest:

Correlation and entropy search: Age, glucose, and BMI.

Forward greedy search: Glucose and skin thickness.

Backward greedy search: Pregnancies, DPF, age, glucose, blood pressure, skin thickness, and BMI.

Hill Climb Search: DPF, glucose, blood pressure, skin thickness, and insulin.

Exhaustive Search: DPF, age, glucose, blood pressure, skin thickness, and BMI.

Although not all feature selection processes selected all features, the features that were not selected were inconsistent. Therefore, the project proceeded using all features available in the dataset.

The features most often selected was glucose, whereas the insulin was selected by only hill climb search.

Evaluation:

Because this project entails identifying individuals who may have diabetes, it may be more advantageous to evaluate models based on how well they identify diabetic individuals, rather than their overall accuracy. As such, sensetivity was prioritized when determining which models were of the most interest.

Figure 7: Model Evaluation Metrics

Model	**Accuracy**	**Sensitivity**	**Specificity**
OneR	0.7239	0.51	0.90
Naive Bayes	0.78	0.69	0.82
Naive Bayes (with class proportions maintained)	0.72	0.62	0.77
Decision Tree	0.7068	0.78	0.56
Rule Set	0.6971	0.77	0.55
Logistic Regression	0.755	0.61	0.83
SVM	0.765	0.55	0.88
Neural Network	0.667	0.17	0.93
KNN	0.73	0.55	0.83

OneR:

OneR selected glucose as the most significant feature. It achieved an accuracy of 0.7239, a sensetivity of 0.51, and a specificity of 0.90.

Naive Bayes:

The first of two Naive Bayes models used a training set that did not maintain class proportions. It had an accuracy of 0.72, a sensitivity of 0.69, and a specificity of 0.82.

Naive Bayes (with class proportions maintained):

The second of two Naive Bayes models used a training set where class proportions were maintained. It had an accuracy of 0.72, a sensitivity of 0.62, and a specificity of 0.77, making it uniformly worse than the first Naive Bayes model.

Decision Tree:

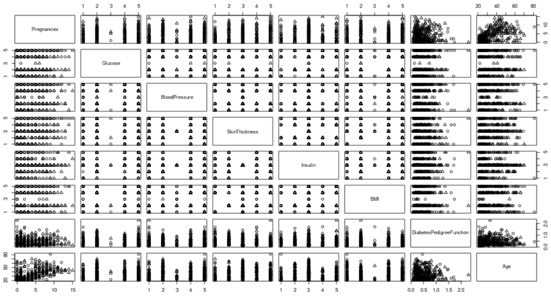
The decision tree model had an accuracy of 0.7068, a sensitivity of 0.78, and a specificity of 0.56.

Rule Set:

RIPPER produced a rule set with an accuracy of 0.6971, a sensitivity of 0.77, and a specificity of 0.55.

Logistic Regression:

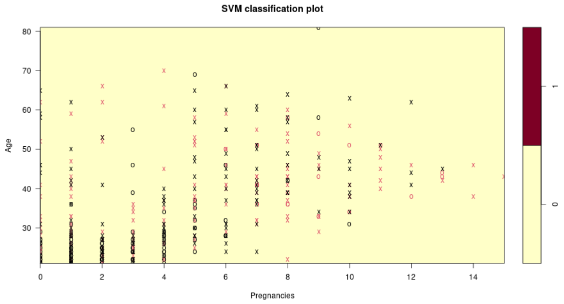
Figure 8: Logistic Regression Plot



The logistic regression model had an accuracy of 0.755, a sensitivity of 0.61, and a specificity of 0.83.

SVM:

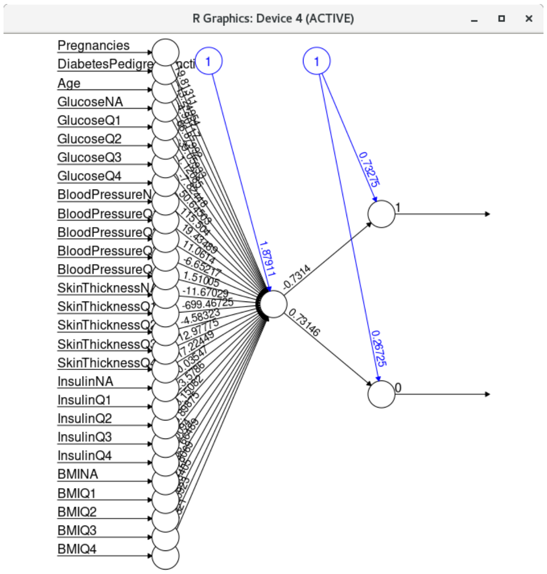
Figure 9: SVM Classification Plot of Pegnancies vs. Age



The support vector machine had an accuracy of 0.765, a sensitivity of 0.55, and a specificity of 0.88.

Neural Network:

Figure 10: Neural Network Plot



The neural network model had an accuracy of 0.667, a sensitivity of 0.17, and a specificity of 0.93. It predicted that patients were not diabetic in the vast majority of cases.

KNN:

KNN was most effective with a k of 10, and had an accuracy of 0.73, a sensitivity of 0.55, and a specificity of 0.83.

Deployment:

Of all the classifiers, the decision tree model was the most effective. Although it was less accurate than the OneR, Naive Bayes, logistic regression, SVM, and KNN models, it had much higher specificity than any of them, meaning that it was much better at identifying diabetic patients.

The next best model was the rule set created using RIPPER, which had evaluation metrics that were all slightly worse than those of the decision tree.

Of the error-based classifiers, the most effective was logistic regression, which had the highest sensitivity and accuracy only slightly worse than that of the support vector machine.

Conclusions:

If the results of these models hold true for patients beyond those included in the data set, the decision tree model will correctly identify diabetic patients approximately 78% of the time.

Although it is not a foolproof means of detecting diabetes, its predictions are still meaningful. It may be useful as a means of identifying patients who may benefit from seeking a more comprehensive medical evaluation.

Future Work:

Although the decision tree model is effective, there are likely to be many ways in which it could be improved. At present, the most salient issue is its low specificity. Although some innaccuracy is acceptable, especially in the direction of excessive caution, a specificity of 0.56 would result in an excessive number of false positives were the model actually to be deployed.

Avenues by which to further develop the decision tree model might include implementing pre- or post-pruning, compensating for missing values through different means of feature derivation, or experimention with models that use different combinations of features.