Overview:

I have changed Outcome from a numeric variable to a factor, to better represent its place in the data, and have made categorical versions of all continuous variables in my dataset using equal-frequency binning, either as a means of accounting for missing data or as an alternative version to use in models.

Detail:

Feature Modifications and Fixes:

Because Outcome is intended to be a categorical variable where 1 represents a patient who has been diagnosed with diabetes and 0 represents a patient who has not, I have changed it from a numeric variable to a factor.

Feature Derivation:

Certain tuples have values of zero for attributes where such a measurement does not make sense, specifically Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI. These most likely represent missing data.

As such, I have created categorical variables based on each of those attributes where an original value of zero results in the new value “N/A.” The other four values are equal-frequency bins based on the quartiles of all original data points that were not equal to zero.

In practice, these quartile categories have similar but not perfectly equal frequencies. I think that this is because the quartiles themselves occur multiple times in the data. If that’s not the case, I may have programmed my binning function incorrectly.

I’ve kept the original numeric variables to see how they compare to their binned counterparts. Similarly, I’ve made binned versions of Age, Pregnancies, and DiabetesPedigreeFunction that do not consider zeros missing values, also for the sake of comparison.

Feature Selection:

When using the “fixed” attribute set (which makes use of binning only to compensate for missing data), the random forest approach selected Glucose, Age, BMI, and Pregancies as the four most significant features.

When working with the “simplified” data set, which uses binned versions of all numerical features, it instead selected Glucose, BMI, Age, and Insulin.

A correlation and entropy search with the fixed attribute set selected Age, Glucose, and BMI, whereas one with the simplified set selected Glucose, Insulin, BMI, and Age.

When using the fixed data set, FSelelector attribute selection returned the following results:

Forward greedy search: Glucose and Skin Thickness

Backward greedy search: Pregnancies, DiabetesPedigreeFunction, Age, Glucose, Blood Pressure, Skin Thickness, and BMI

Hill Climb Search: DiabetesPedigreeFunction, Glucose, Blood Pressure, Skin Thickness, and Insulin

Exhaustive Search: DiabetesPedigreeFunction, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI

When using the simplified data set, FSelelector attribute selection instead returned the following results:

Forward greedy search: Glucose

Backward greedy search: Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pegnancies, Age, and DiabetesPedigreeFunction.

Hill climb search: Glucose, Blood Pressure, BMI, and Pregnancies

Exhaustive Search: Glucose, Insulin, BMI, Age, and DiabetesPedigreeFunction