Prediction of Diabetes Using Classification Methods

By Victor Josifovski

Project Description:

Diabetes remains a costly and harmful issue to American society and healthcare, as the chronic disease affects millions of patients in the U.S. each year. As mentioned in a 2020 CDC report, nearly thirty four million Americans suffer from diabetes, with another eighty eight million having prediabetes yet to form. This evidences the fact that diabetes is not only widespread in the United States but continuing to be a problem for future generations. Being a chronic disease, diabetes costs both its patients and the healthcare systems that treat them, as its continuous treatment proves to be exorbitant. Diabetes is the most expensive chronic condition in the nation, with an annual cost of two hundred and thirty seven billion dollars, and one in every four dollars spent in the U.S. healthcare system being rooted in diabetes treatment. Current trends indicate that soon nearly one in three Americans will suffer from diabetes in their lifetime, and as mentioned earlier, nearly eighty eight million Americans are expected to develop the disease in the future. Diabetes is costly, widespread, chronic, and growing, but early detection can serve as a mitigating medium. Early diagnosis can not only delay, but potentially prevent the progression of diabetes, and serve to aid patients while creating massive cuts in spending across healthcare systems. Thus it becomes urgent to establish early detection services and abilities, especially with the colossal amount of expected future patients, which will serve to magnify the already harmful effects of diabetes on both patients and healthcare systems if they are not effectively mitigated.

With new data being collected on diabetes patients, early detection becomes possible through machine learning models. Data sets containing predictors and measurements pertaining to diabetes diagnosis can be used to create models that can provide early diagnosis methods. While the prediction of diabetes within a patient has proven to be difficult with numerical data, a database using the correct parameters may be used to predict diabetes through a logistic regression model. Using a database publicly provided by UCI machine learning on Kaggle, and derived from the National Institute of Diabetes and Digestive and Kidney Diseases, diabetes prediction can be conducted through the aforementioned logistic regression models. Using binary classification it is established whether or not a patient has diabetes within the dataset, thus leaving the potentiality for a logistic regression model that can use the parameters within the table to create a predictive algorithm of diabetes within patients. The data set contains seven hundred and thirty eight cases and diagnostic measurements of

Pima Indian women, with 500 cases being negative for diabetes and 238 being positive. Further, the dataset consists of eight different independent variables, in the form of diagnostic measurements, that pertain to the dependent and dichotomous data of a positive or negative diagnosis on diabetes. Thus, the data can be fit to a model, and further used for predictive capabilities in other groups of patients. Specifically, with a diagnosis being a binary data type, a logistic regression model becomes the best model for prediction. Logistic regression is used to model the relationship between multiple independent variables in contact with a binary or dichotomous dependent variable, and thus provide a probability measurement of a certain event occurring or not occuring. This means a logistic regression model can be applied to create a model of the testing data of the Pima data set in order to train the equation to create future predictive outcomes.

With the extreme numbers of diabetes patients within the United States, and especially with the large numbers of prediabetes patients within the United States, it becomes important that a supervised learning method can be trained and applied to the structured Pima Data set. Improved prediction and early diagnosis of diabetes could provide U.S. healthcare with billions in spending mitigated, as well as to help patients from developing the severe chronic condition of diabetes itself, or inevitably moving towards type-2 diabetes. With so many potential cases that can be mitigated and even removed in the future, it should be clear that predictive prevention is requisite, and it can be provided through the coupled work of the Pima data set and a logistic regression model. Thus it becomes my desired end goal to produce a logistic regression model, and potentially some other supervised models, which can accurately predict the onset of diabetes in a patient, using the Pima data set. With the data set being calibrated toward the usage of supervised learning models for prediction, I believe I can create accurate and useful prediction capabilities using a logistic regression model, which pairs with the quantitative diagnostic measurements and the binary outcomes. Therefore, through a logistic regression model, I plan on using the Pima data set to fulfill the need for a predictive ability for diabetes diagnosis. This would make my desired end goal to be able to predict the onset of diabetes within patients through a classification model, preferably logistic regression, in an accurate manner and through the use of the structured and provided Pima data set. Along with this, I would like to use classification methods to derive from the data set which attributes are most capable of predicting the onset of diabetes in a patient, among the eight which are measured in the table. Overall, it is obvious that diabetes is an urgent and ongoing issue within the United States, and there being a need for predictive ability surrounding the binary diagnosis

of diabetes, as well as the structured data set provided, my end goal has become about providing that through a classification model.

Link to dataset:

Kaggle link: https://www.kaggle.com/uciml/pima-indians-diabetes-database

The Pima Indians diabetes database, which I refer to as the Pima data set, consists of seven hundred thirty eight cases of Pima Indian women twenty one years or older, and their diagnostic measurements surrounding diabetes. As mentioned before it is a structured data set, measuring eight different independent variables, and outputting one dependent variable. The eight independent measurements found in the data set include measurements of Glucose, Blood pressure, Skin thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. These eight variables correspond to the dependent variable, which is a positive or negative diagnosis of diabetes, insinuated by a negative 0 or a positive 1 in binary.

Planned ML models/Expected Output Data Visualization Analytics: :

For my project, I will be using a logistic regression model, as well as a decision tree and any other supervised classification model I find to be accurate. This is because the particular data set I have chosen to use is tailored towards a supervised learning model, as it is structured and with a dichotomous dependent variable. Further, the nature of diagnosis for diabetes is binary in itself, as it is a variable that can be calculated with a yes or no answer. Therefore it becomes apparent that the best models for the prediction of a positive or negative diagnosis of diabetes involve a classification mode, and more specifically a logistic regression model. However, it is important for me to attempt and find the best possible algorithm for predictive capability. Therefore, it would be important for me to test through different classification models, such as a decision tree or random forest. I will thus apply a logistic regression model to calculate the probability of a positive or negative diagnosis, however I will also ensure that other classification models such as decision trees are not more accurate or capable themselves.

Block Diagram/System Overview

Overall, I ended up creating two different algorithms, a logistic regression algorithm, as well as a decision tree. For my logistic regression model, I began by checking my data for null values and cleaning my data as well as checking for duplicate values. I then normalized my data in order to create my variables in a way that contributed equally to the model. I then began producing my logistic regression model. The logistical regression model was to analyze the eight different variables within the data set and produce a prediction on the binary outcome of a positive or

negative diagnosis of diabetes. The model used hyperparameter tuning using the solvers Lbfgs as well as libline, as they worked best with smaller data sets, of which I was using. The model achieved a 75.97 percent accuracy from a testing set of twenty percent of the data. My second, and less accurate model implemented was the decision tree. Once again following the same data cleaning methods the model was derived from a training set of twenty percent of the original data set. The model achieved a 75.32 percent accuracy score. Overall, this leaves both of the models in a similar range regarding their accuracy.

Data Analytics

In the second portion of my files I have provided data analytics graphs for both of the models. Using a seaborn pairplot function was a method I researched and realized could be valuable for data visualization. The pairplot function demonstrates the relationship between each of the eight variables and the predicted outcome of a diagnosis. This includes the relationship between the variables with each other, and combined, as well as in comparison with the dependent variable of the outcome. It becomes apparent which variables and diagnostics share a close correlation with each other, and more importantly with the diagnosis outcome. The pair plot features an x and y axis in which the different predictors are lined up, and can thus be compared with each other, with the plotted positive and negative diagnoses on them and the key on the right. It becomes apparent, with the histograms moving diagonally across the grid, what predictors are most correlated with a positive or negative diagnosis within the data. There are expanded graphs provided on these histograms for each of the variables. For both the decision tree and logistic regression files, there is provided a histogram analyzing each of the predictors in correlation with the outcome. As can be discerned, the data visualization graphs correlate three main predictors with the most predictive capability in diagnosing diabetes. First, the graphs demonstrate a large correlation between measures of BMI and a positive or negative diagnosis. This can be seen in the histogram in which BMI and outcome are compared, in which the line of distribution matches similarly for both BMI and outcome. Next, the graphs discern another correlation between a measure of Skin thickness and a positive or negative diagnosis. Once again there seems to be a large correlation between the distribution of positive and negative diagnoses and the thickness of the subjects skin, as the graph demonstrates. Finally, in another histogram comparison provided in the file, the correlation between blood pressure and outcome is pronounced through the histogram. So, from these analytic histogram graphs, the data visualization demonstrates that there is a substantial correlation between the diagnostic measurements of BMI, Skin Thickness, and Blood Pressure in discerning a positive or negative

diagnosis of diabetes. The data analytics charts are thus provided in two steps, the larger pairplot, summarily analyzing the predictors with plotted positive and negative diagnoses, and the specific histograms which analyze each of the predictors with the overall diagnoses.

Machine Learning/Prediction Models Used

As stated before, the files contain both a logistic regression and decision tree model, both being seventy five percent accurate. First, the data was analyzed for both null values and repeated values, and then made ready for a logistic regression model. This included normalization for data fitting to occur. The data was then fitted to a logistic regression model, using a 20-80 train test split and a random state of 42. Using a logistic regression formula like the one we were taught during class, and in use of the Albery weather file, the diabetes data was fitted to a logistic regression model. Twenty different cases were analyzed and returned an accuracy score of 75.97 percent in predicting a positive or negative diagnoses of diabetes within a patient after being put through an accuracy test. Then the model was put through hyperparameter tuning, in which the solvers Lbfgs and libline were used. This was because libline is generally optimized for smaller data sets, and the Pimi Indian dataset is not the largest data set in its respects. The best accuracy provided was 76.66 percent after the hyperparameter tuning. Thus, the logistic regression model consisted of a cleansing and analysis of data, followed by normalization and preparation of data, fitting of the data into a logistic regression model, and finally an accuracy test and hyperparameter tuning. This all returned a model capable of predicting a positive or negative case a little over three fourths of the cases. The second model incorporated in my files includes a decision tree model. The decision tree model consists of the same data cleansing, however without the data normalization, as data normalization is not known to have an effect on a decision tree's performance. Thus the data was applied to a decision tree coded similarly to that of the famous Iris classification model we were exposed to in class. There was a 20-80 training and testing data split, and a random state of 42 proved to provide the most accurate results. The decision tree was 75.32 percent accurate at predicting a positive or negative diagnosis, thus making it slightly less accurate than the logistic regression model. Thus both a logistic regression and decision tree were used, with negligible predictive difference between the two classification models.

Planned Goals Vs Achieved End Results

The main desire of this project was to provide a classification model that could be fit within the data to predict the diagnosis of a diabetes patient at a rate perceivable enough to be accurate for early diagnosis of diabetes patients. Furthermore, using a classification model provided the ability to analyze which of the diagnostic measurements and parameters provided in the table gave the greatest ability to predict the onset of a positive case. Finally, it was also desired to determine which of the classification models would be more accurate in its ability. For the logistic regression model, as well as the decision tree, a predictive ability of eighty percent or greater was preferred. However, both models were similar in their ability at around seventy five percent predictive ability. Thus, in the end two models were produced, able to classify at a rate slightly below which was preferred. Secondly however, the ability to analyse the relationship between the outcomes of a diagnosis and the diagnostic measurements proved more fruitful. The models and data visualization were able to discern which of the parameters provided by the data set were most correlated with a positive diagnosis of diabetes. As mentioned earlier, these parameters were BMI, Blood Pressure, and Skin Thickness. This was a successful portion of the project, which was provided through the abilities of a classification model at examining the relation between the independent variables, or parameters, and the outcomes they correlate with. The graphs and data visualization were clearly able to discern which of the parameters contained the best predictive values and thus that portion of the model was fulfilled. Finally however, the efforts to discover which of the classification models would do better at predicting the outcome was not as successful. With both models achieving a similar rate of accuracy, there is no ability to understand which may potentially be more accurate. Thus it becomes a matter of other factors, such as efficiency as well as the progression onto other future data, which would identify the more accurate of the two models. Thus, the original planned results were to obtain a model with an eighty percent accuracy, of which there were two obtained with seventy five percent accuracy. Furthermore the similar nature of the predictive accuracy of both the block chain decision tree and the logistic regression model inhibits the ability to understand which is truly more accurate. However, observation of parameters using data visualization was successful at discerning which parameters were the most important in predicting diabetes and thus was achieved.

Summary

Overall, the predictive necessity required for the diagnosis and prevention of diabetes remains a current and growing problem. With diabetes being an expanding, expensive, chronic, and preventative disease, it is imperative for a predictive method to be devised. This is a problem

that has great potentiality for the use of machine learning methods and models. Thus, provided with a structured data set, consisting of seven hundred and thirty eight different cases of Pima Indian women and their diabetic diagnostics, a classification model was primed for design and prediction in the outcome of a positive or negative diagnosis. Further, different classification models were to be examined and parameters were to be examined for their accuracy. Considering the amateurity of my machine learning abilities, much of what I used was explored and learned from textbooks and data designs we had been exposed to before, as well as what could be researched and tailored for the specific project. Using the teaching provided for us, the data was fitted into two different models and came below producing the predictive accuracy desired for such a model to be effective in the medical field at around seventy five percent. However disappointing, the nature of the Pima data set is not very large, and being a structured data set, the classification models seemed to be the desired path to pursue, and so they were. The models were applied to their ability and can still be serviceably capable, while the parameters within the data set have been analyzed in their own right and have provided their own information as BMI, Skin Thickness, and Blood Pressure have all been highlighted as potentially strong predictors. The constraints of the data set were navigated with the classification models known, and this has served as a fitting introduction to the application of machine learning within a real world issue. In summary of the three goals, to have a high predictive capability of a diagnostic outcome, to provide information on parameters, and to discern which classification model was most accurate; I believe the application of the classification models was not able to achieve the first or last, and only the second. However, valuable information has been understood regarding the diagnosis of diabetes. If the project were to be done over again, it becomes apparent that potentially an unsupervised model including potentially clustering, could be applied to discern patterns within a diabetes data set. However, this could not be achieved within the supervised data set provided, and so I am satisfied with the performance and application of the classification models that were used. There is much potential for the use of machine learning for predictive features in any field, and there is even more potential to expand on the similar work that has been produced in this project. I would also like to thank you, professor Shiva, for helping introduce me to the beautiful tools of Python and machine learning. Thank you and all the best!

Files and Links:

- [ ] Dataset in CSV format on external site usage:
  https://www.kaggle.com/uciml/pima-indians-diabetes-database