Taking your diabetes medication?

SENG 550 Project

Moiz Abdullah 30066638 Daniel (pin) Long 30068063

Hashir Ahmed 30070165 Rifat Haque 30074728

Preamble—

Hashir- Did data preprocessing, helped write the report. 25%

Daniel- Created ML models, did EDA, preprocessing. 25%

Moiz- EDA, data preprocessing and parameter tuning. 25%

Rifat- Created the report document. 25%

These contributions are true and signed below:

Rifat Haque: 

Daniel (Pin Long): 

Hashir Ahmed: 

Moiz Abdullah: 

The code is submitted with the report. For our original plan we were going to use flight data to predict whether a flight would be canceled on a given date, but after creating a model, we found that the data was skewed and had far too many problems to continue on. Thus, after asking the professor if we could use another data set he agreed, and as such we chose to use a diabetes dataset to predict whether someone needs diabetes medication.

Abstract— Millions of people across the world suffer from diabetes, making it a very common illness in our society. Moreover, since there are different kinds of diabetes, even those without a family history can have the condition. As such, treating this illness with care is an incredibly important thing, which is why we chose this dataset. For this dataset, we will be creating machine learning models using Spark concepts and tools. In this paper we will go over the preprocessing, design, and tuning of the dataset and in the end we will discuss the results of our work.

I. Introduction

A. *What is the problem you selected?*

The problem we selected, or more specifically the dataset we are using has been crafted over ten years (1998-2008) and has clinical patient data from over 130 hospitals from across the United States. What we will aim to tackle is to use the data provided and create a machine learning model to predict whether a person, given their symptoms and medical history as features, need diabetes medication or not. Essentially our model will, in a form, decide whether to prescribe a person diabetes medication. Note, all the samples are of people who have diabetes, but whether they are taking their medication is the goal of the model to predict.

B. Why is it an important problem?

Diabetes is one of the most common illnesses found on this planet, and it is also one that even those without hereditary history of the illness can succumb to it, in the form of type two diabetes. However, in the case of type two diabetes, it can be delayed or even flat out prevented if the right steps are taken. In order to take these steps, we first need to analyze the situation and understand where a patient stands in regards to potentially suffering this illness. All this can be achieved by seeing and understanding the patterns within an individual's life and medical history, which is why we have decided to implement machine learning, on a big data scale, to achieve this.

C. What have others done in this space?

Many have used this dataset for similar purposes using machine learning models to predict various things, such as if a person has steady insulin levels based on the data provided. Most have used this dataset to help visualize the trends in the diabetes illness as this dataset is a ripe source of information over the course of ten years.

D. What are some existing gaps that you seek to fill?

The features that we hope to tackle is performing the EDA and building meaningful machine-learning models on a big data platform. The dataset we are using has over a 100,000 samples collected, with over 50 features per sample. This, compared to our initial project idea, does have a smaller data size, but we believe that having performed this using our big data tool, like spark, we can scale this to much larger datasets. This project is meant to show that what we have built can be scaled to much larger datasets and that these models provide meaningful insights.

E. What are your data analysis questions?

Our main goal here is to look at the dataset and to find the correlation between the medical histories of patients to find out whether or not they require diabetes medication. This question is the driving force behind our project.

F. What are you proposing and what are your main findings?

As stated above, our main goal is to build machine learning models to predict whether an individual requires diabetes medication. We will be using Logistic Regression, Naive Bayes, and Decision Tree as our main machine learning models. In order to achieve the best results possible, we will be performing extensive data analysis and hyperparameter tuning to models where it applies. We will discuss and explore our findings in the Result section of this paper.

II. Background and related work

A. Technical background helpful for understanding your report (if relevant)

The main technical aspect of our project is dealing with Spark, which we will use on the databricks platform. Our implementation of the machine learning models is similar to those shown in class, and we have not used any major piece of technology outside this. However, in order to understand our findings, we will be using metrics such as accuracy, precision, recall, and f1 scores to convey the ability of our models. We will not spend too much time discussing these parameters, but will discuss what they mean for the model. To facilitate the readers understanding, we will also try to make as much use of visual charts and diagrams to convey our data analysis.

B. Review of existing work pertinent to your project

As discussed earlier, diabetes is a very common illness, and a very dangerous one at that. For this reason, there has been a long time interest in delving deep into the factors that play a role in this illness. Unfortunately, we were not able to find scientific papers that specifically use the same dataset as ours, but there is a lot of work being done similar to ours. For example, the paper by Kaur, H. and Kumari, V. (2022) [1] discusses the importance of accurately predicting the early onset of the illness as its cases are rapidly growing around the world. In this paper, they use five machine learning models to develop trends and detect patterns with risk factors regarding diabetes. Though we used a different dataset, our goal is similar to that which is in the paper.

III. Methodology

A. Experiment setup

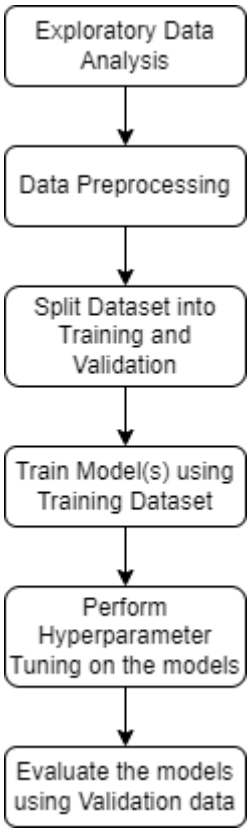

Fig. 1 . Activities flow diagram.

Firstly, we had to do some feature engineering on our dataset. We chose certain features of higher importance and value. In these data entries there were empty fields denoted by a '?' which was replaced with 'None'. Then we separated numerical features and categorical features. This was done so we could implement hot one encoding on the categorical data. With this cleaner dataset we could begin our experiment.

B. Experimentation factors

- **Logistic Regression:** Logistic Regression is used when the dependent variable is categorical. For our purposes our regression model is binary as a patient is either taking or not taking diabetes medication.

- **Naive Bayes:** Naive Bayesis a probabilistic model that is used for classification tasks. This model used the Bayes theorem, the assumption made here is that the predictors/features are independent. That is, the presence of one particular feature does not affect the other. Our model uses 36 base features and one target feature. Now the assumption made by the Bayes theorem mostly holds true as the features hold weak

correlation as shown by the heat map below.

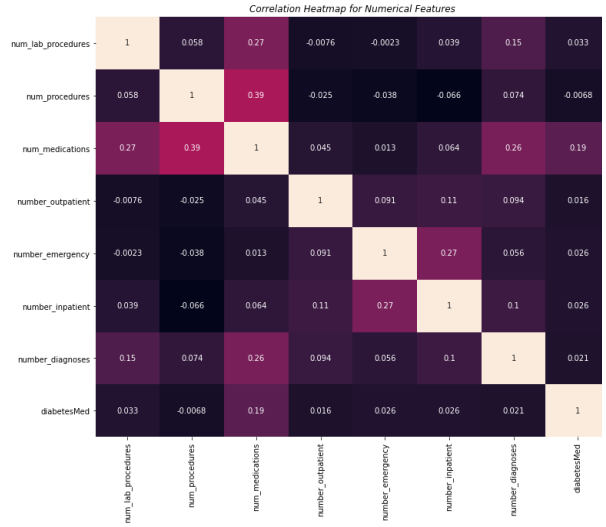

Fig. 2 . Heat map of features

- **Decision Tree:** In a decision tree model, as the name suggests a decision tree is used to visually and explicitly represent decisions and decision making. For our purposes, it would determine whether a patient would use diabetes medication based on the features that we feed the model.

- **Hyperparameter tuning:** We did hyper-tuning for all of our models. To do this we used CrossValidator to perform hyper tuning. For logistic regression, we hyper-tuned the regularization parameter and elastic net parameter. For Naive Bayes, we hyper-tuned a smoothing parameter. Lastly, we hyper-tuned the max depth parameter for the Decision tree. Through the hyperparameter tuning, we were able to achieve better metrics across the board.

- **Details on data split:** The dataset we used is split into two parts which are called train-test and validation. The purpose of naming one part train-test is for hyperparameter tuning, and then to validate the quality of our model we used the validation set.

C. *Experiment process*

First, we did exploratory data analysis to discover the underlying implications of the information found in the data. Then we worked on feature engineering to improve the data that we would later feed our models. We then split the data into train-test and validation. Next we made our models which consisted of three models. These models are Logistic Regression, Naive Bayes, and Decision Tree. After this we used hyperparameter tuning on all of these models to achieve the best possible performance metrics.

D. *Performance metrics*

The key performance metrics that we will be using are the standard accuracy, precision, recall, and F-score. To briefly discuss the meanings of these metrics, refer to the table below:

TABLE I. Metrics and Definitions [2]

Metric	Definition
Accuracy	The fraction of predictions our model got right
Precision	What proportion of positive identifications was actually correct?
Recall	What proportion of actual positives was identified correctly?
F-Score	The harmonic mean of precision and recall

Fig. 3 . Metrics and Definitions

IV. RESULTS

A. *Key findings.*

a. *Exploratory data analysis*

For our exploratory data analysis, we built several charts to help us visualize the data and to perform feature engineering, choosing the features we believed would allow us to achieve the best results. Here are the diagrams along with a small description (if necessary).

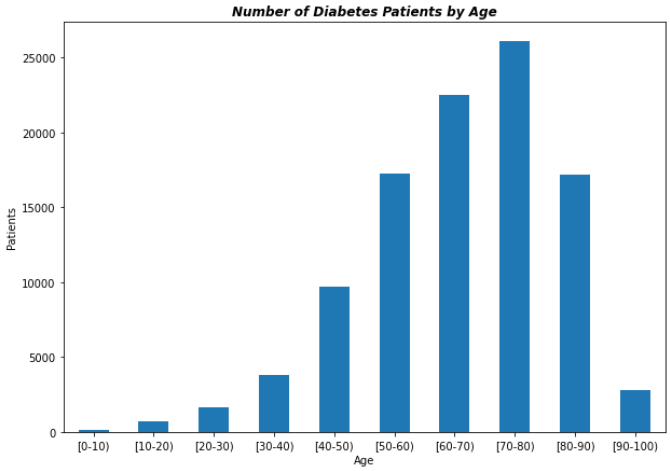

Fig. 4 . Number of Diabetes Patients by Age

The bar graph here shows the distribution of diabetes patients over different age groups. As expected, we observed that diabetes was far more predominant in older ages, which makes sense intuitively, as physical health is a large

contributor in diabetes, and we can assume that younger people are more physically active.

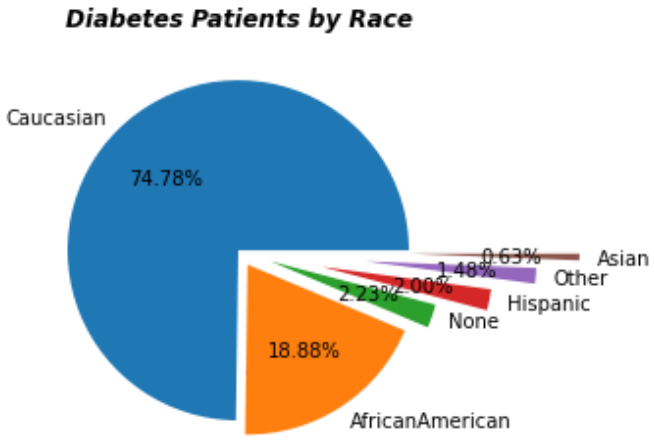

Fig. 5 . Diabetes Patients by Race

This pie chart shows the distribution of races of diabetes patients. The majority is caucasian and the smallest percentage is Asian. 2.23 % of the patients, which are designated as none are people we have no data on about their race. Through this chart, we can infer that Caucasians make up the majority of patients, but that African Americans make up a disproportionate number of patients. By comparing the ratio of African Americans in the population versus the number of diabetes patients. A quick google search shows the current percentage of African Americans in the U.S. is 13.6%, and the number of African American patients with diabetes is 18.9%. Thus, showing that African Americans are disproportionately affected by diabetes.

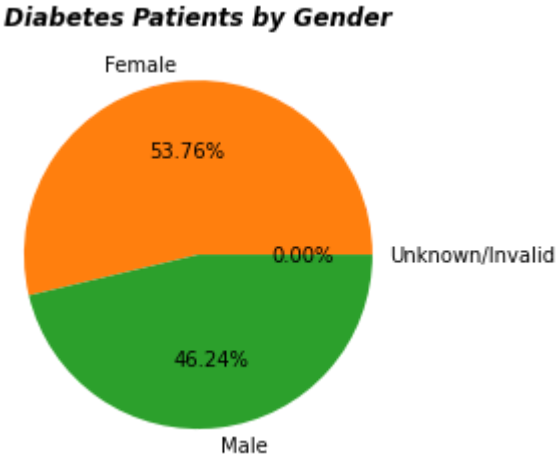

Fig. 6 . Diabetes Patients by Gender

The pie chart above is a simple visualization of the percentage of men and women that have diabetes. The data also showed that women were more likely to be diabetes patients than men.

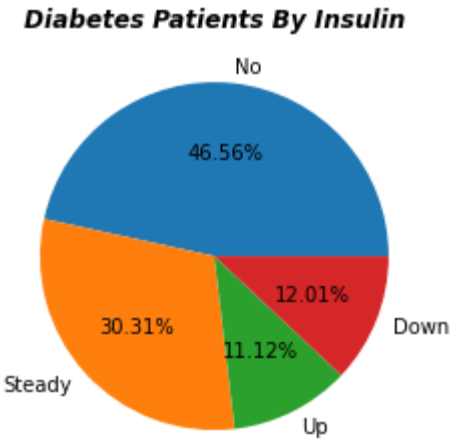

Fig. 7 . Diabetes Patients by Insulin Medication

This pie-chart here is perhaps more interesting than the previous. This chart shows the insulin the patient is taking. Nearly half the patients are not taking insulin medication. This could be due to the expensive cost of insulin in the United States, or perhaps do to the different types of diabetes, such as those with type two diabetes that don't require diabetes injections.

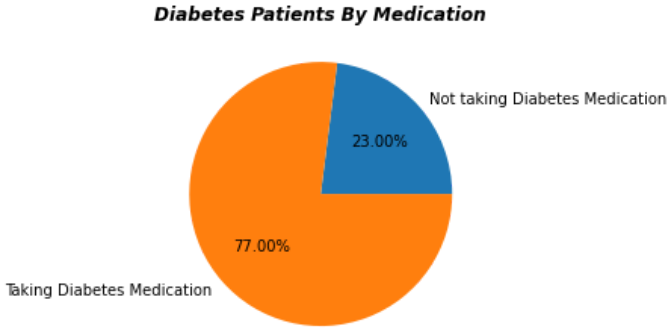

Fig. 8 . Diabetes Patients by Diabetes Medication

This pie-chart is basically what our project mainly revolves around. This represents the portion of patients that are on Diabetes medication versus those who are not.

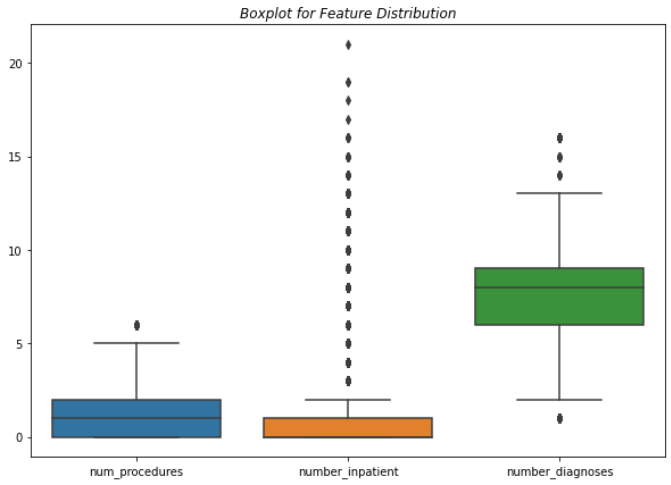

Fig. 9. Boxplot for Feature Distribution

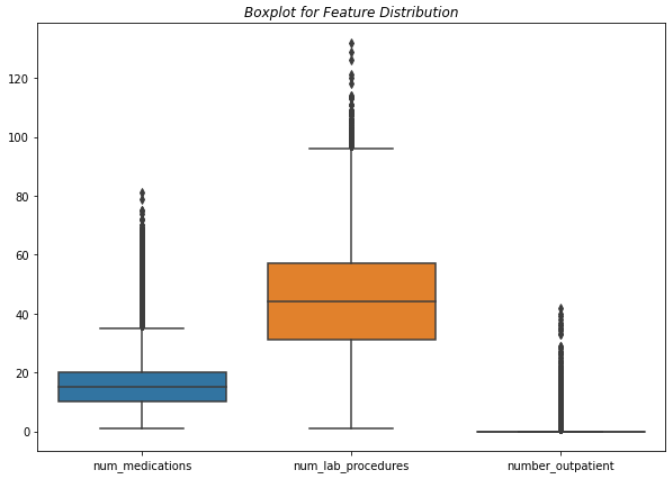

Fig. 10. Boxplot for Feature Distribution

The two boxplots above demonstrate the spread and skewness of the numeric features in the dataset. We found that the number of medications, the number of lab procedures and the number of our patients had many outliers. The skewness can be explained by the concentration in a few patients. We decided to keep the outliers to account for the patients that require more resources.

3.1 LR

3.1 A LR Baseline

```
Summary Stats
Accuracy = 0.7721745172978607
Precision = 0.7721745172978607
Recall = 1.0
F1 Score = 0.8714429755769662

[[    0.  4590.]
 [    0. 15557.]]
```

Fig. 11. Summary of metric stats for LR Baseline

Using the baseline Logistic Regression, we have created a model that seems to predict true for everything. Although this model has a relatively high Accuracy, Precision, Recall, and F1 Score.

The nature of the prediction is evident when the confusion matrix is observed. The confusion matrix shows that the model evaluated all cross-validation data as either True Positive or False Positive.

The fact that this model still has a 77% accuracy score is simply due to the fact that ¾ of the input data has a label of 1 indicating Diabetes Patients taking medication.

Thus even with a model that simply just predicts 1 for all input data, it gives a relatively high accuracy. But this can clearly be improved with threshold and hyperparameter tuning to make this model more representative of the data.

3.1 B LR - Threshold tuning for max F-Measured

```
Best Threshold: 0.6918576787358852

Summary Stats
Accuracy = 0.9978160520176701
Precision = 0.9989707301383082
Recall = 0.9982001671273382
F1 Score = 0.9985852999807086

[[ 4574.    16.]
 [   28. 15529.]]
```

Fig. 12. Summary of metric stats for LR Threshold tuning for max F-Measured

After performing Threshold tuning for maximum F-Measure score. We found that the best threshold is a value of 0.69, where with this threshold, the accuracy, precision, recall, and F1 score of the model all went above 99%. This indicates that the main issue with the baseline model is simply the threshold and changing this results in an almost perfect model.

3.1 C LR - Hyper Parameter Tuning

```
Best regParam:  0.01
Best elasticNetParam:  0.0

Summary Stats
Accuracy = 1.0
Precision = 1.0
Recall = 1.0
F1 Score = 1.0

[[ 4590.     0.]
 [    0. 15557.]]
```

Fig. 13. Summary of metric stats for LR Hyper Parameter Tuning

Out of curiosity, we also explored performing hyper parameter tuning on the model using the best parameter to see the possible effect of this. We found that by having a Regularization Parameter of 0.01 and an elastic Net Parameter of 0 we can create a model that perfectly predicts the unseen CrossValidation Dataset.

3.2 Naive Bayes
3.2 A - Baseline

```
Summary Stats
Accuracy = 0.9597955030525637
Precision = 0.9613339172871176
Recall = 0.9876582888731761
F1 Score = 0.9743183259353202

[[ 3972.   618.]
 [  192. 15365.]]
```

Fig. 14. Summary of metric stats for Naive Bayes Baseline

The second model type we build using our dataset is the Naive Bayes model, the baseline model without any modifications results in an accuracy rate of 96% with all the accuracy, precision, recall, and F1 Score to be above 96%. From those metrics and our confusion matrix, we found that this model appears to work well for all of the unseen CrossValidation data.

3.2 B - Hyper Parameter Tuning

```
Best smoothing:  1.0
Summary Stats
Accuracy = 0.9597955030525637
Precision = 0.9613339172871176
Recall = 0.9876582888731761
F1 Score = 0.9743183259353202

[[ 3972.   618.]
 [  192. 15365.]]
```

Fig. 15. Summary of metric stats for Naive Bayes with hyperparameter tuning

Upon performing hyperparameter tuning on Naive Bayes we found that it didn't change the performance. We tuned the smoothness of the model and found that the best result is 1 which is used in the baseline model. Thus, the hyperparameter-tuned model has the same results as the baseline model.

3.3 Decision Tree
3.3 A - Baseline DT

```
Test Error = 0.0334541
Summary Stats
Accuracy = 0.9665458877252197
Precision = 1.0
Recall = 0.9566754515652118
F1 Score = 0.9778580814717477

[[ 4590.     0.]
 [  674. 14883.]]
```

Fig. 16. Summary of metric stats for Decision Tree Baseline

The final model type we have built was the Decision Tree Model, this model has a very high accuracy, precision, recall, and F1 Score. Evident from the confusion matrix and the precision score, the model seems to not have produced any False Positives but it appears that the False negative can be improved with further tuning.

3.3 B - Hyper Parameter Tuning

```
Best maxDepth:  9
Summary Stats
Accuracy = 0.997468605747754
Precision = 1.0
Recall = 0.9967217329819373
F1 Score = 0.9983581753211216

[[ 4590.     0.]
 [   51. 15506.]]
```

Fig. 17. Summary of metric stats for Decision Tree with hyperparameter tuning.

By applying hyperparameter tuning on the Decision Tree model, we found that by setting a max length of 9, our model increased the accuracy, recall, and F1 score to be over 99% while keeping a perfect recall score. The model has dramatically reduced the number of False Positives and made the model more accurate and more representative of the input data.

TABLE II. Models and their metrics summary.

Model	Accuracy	Precision	Recall	F1
LR - Baseline	77.2%	77.2%	100%	87.1%
LR - Threshold	99.7%	99.8%	99.8%	99.8%

LR - Hyper Parameter	100%	100%	100%	100%
NB - Baseline	95.9%	96.1%	98.7%	97.4%
NB - Hyper Parameter	95.9%	96.1%	98.7%	97.4%
DT - Baseline	96.7%	100%	95.7%	97.8%
DT 0 Hyper Parameter	99.7%	100%	99.7%	99.8%

Fig. 18. Models and their metrics summary.

Conclusion

Through our report we can make inferences about our dataset, and make predictions on whether a diabetes patient needs medication. In our project, we were able to use EDA to find important information about the patients in the dataset such as, their race distribution, gender distribution, and whether the patients needed a change to their insulin dosage. Our problem was to discover whether or not a diabetes patient would require medication for their illness, and we did this through running machine learning models on a set of data we found on Kaggle listing 10 years of diabetes patient data. It is important to be able to find whether a patient is in need of medication especially with the rising costs of pharmaceuticals. Through our process we were able to develop a highly accurate machine-learning model which was able to predict whether these patients would require medication. Our best model is the Logistic Regression model after both Threshold and Hyperparameter tuning. This solution would be able to help doctors determine whether a patient is in need of diabetes medication by re-affirming their decision or helping in giving a preliminary diagnosis.

References

[1] Kaur, H. and Kumari, V. (2022), "Predictive modeling and analytics for diabetes using a machine learning approach", Applied Computing and Informatics, Vol. 18 No. 1/2, pp. 90-100. https://doi.org/10.1016/j.aci.2018.12.004

[2] Google Machine Learning. (2022). “Machine Learning concepts: Classification”. https://developers.google.com/machine-learning/crash-course/classification/