**Diabetic Data Analysis**

Eric Schneider, [eschneider@bellarmine.edu](mailto:eschneider@bellarmine.edu)

Adham Howaidy, [ahowaidy@bellarmine.edu](mailto:ahowaidy@bellarmine.edu)

**ABSTRACT**

This paper explores a diabetic data csv file from a Machine Learning Repository. It is diabetic data from 130 hospitals in the United States. The paper talks about the goals and hypothesis of this entire project. This project goes through of the background of the original diabetic data and why it was made in the first place. It goes through the graphs and tables of the dataset. It talks about methods during the experiments and certain tools used too. It also talks about several problems encountered when working with this dataset because it is impossible if you get no error at all, so it talks about those problems in details. This paper then explores the results that we came with and discusses them as well as comparisons made between models and results. Finally, it goes through future work or if any improvements need to be made whether it is an improvement for people who want to create a dataset like this one or improvements to the models themselves.

1. **INTRODUCTION**

This data set that we were provided with is about diabetic data which will help us analyze factors related to readmission as well as other results pertaining to patients who have diabetes. Our goal for us is to predict if age predicts the times spent in hospitals, and we also want to see if tree diagrams for insulin levels whether they are steady or not steady.

1. **BACKGROUND**

First, we need to know what a hospital is first. A hospital is an establishment giving clinical and surgical therapy and nursing care for the sick or harmed people. The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. The length of stay was at least 1 day and at most 14 days, laboratory experiments were performed during the encounter, and medications were administered during the encounter.

1. **EXPLORATORY ANALYSIS**

This data set is made up of 101766 samples with 50 columns of varying data types. 25 columns however lacked the necessary amount of data to gain succinct answers, so only 25 of the 50 columns will be used. Time in hospital graph shows that it is skewed to the right while the age of patients’ graph shows that is skewed to the left.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| admission\_type\_id | Int64 |
| admission\_source\_id | Int64 |
| discharge\_disposition\_id | Int64 |
| race | Object |
| gender | Object |
| age | Object |
| time\_in\_hospital | Int64 |
| payer\_code | Object |
| medical\_specialty | Object |
| num\_lab\_procedures | Int64 |
| num\_procedures | Int64 |
| number\_medications | Int64 |
| number\_outpatient | Int64 |
| number\_inpatient | Int64 |
| number\_emergency | Int64 |
| Number\_diagnoses | Int64 |
| citoglipton | Object |
| metformin | Object |
| glipizide | Object |
| glyburide | Object |
| pioglitazone | Object |
| insulin | Object |
| change | Object |
| diabetesMed | Object |
| readmitted | Object |

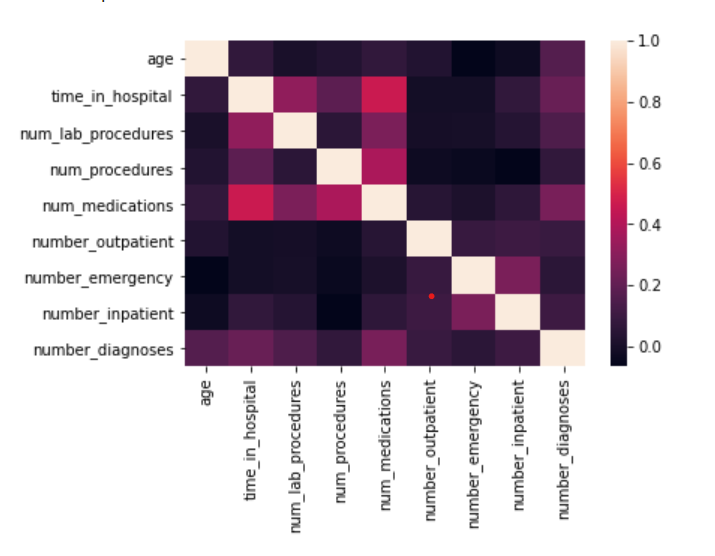
**

Figure 1: Correlations from the dataset (heatmap)

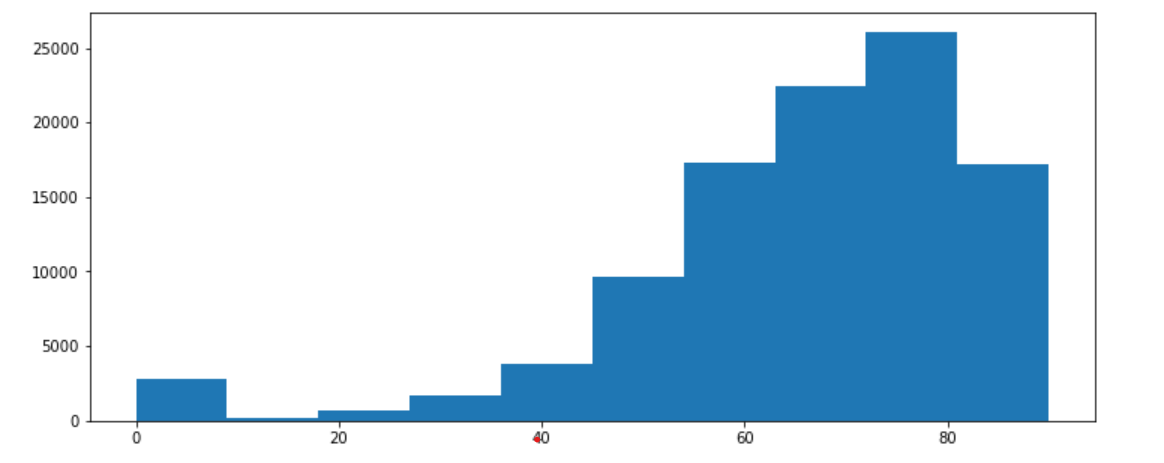
**

Figure 2: Distribution of age (histogram)

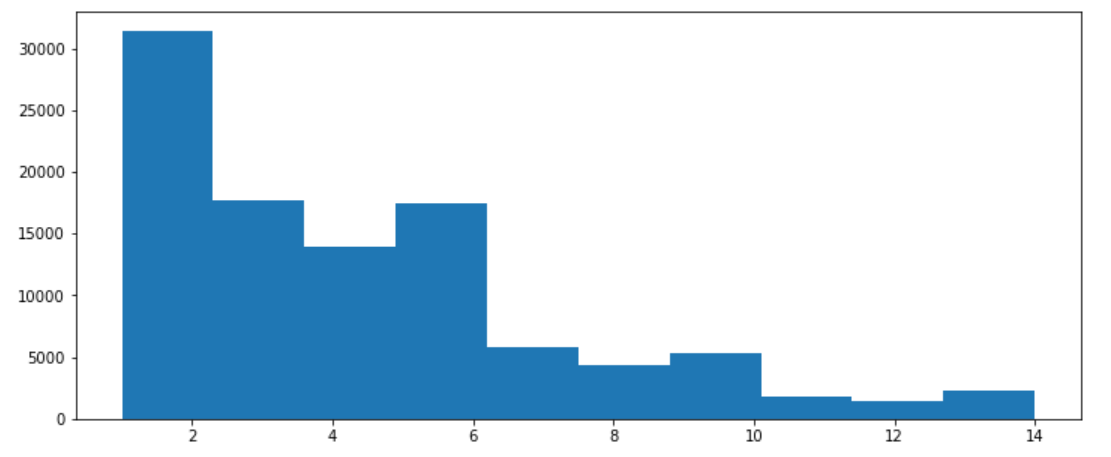


Figure 3: Distribution of time (histogram)

1. **METHODS**
   1. *Data Preparation*

First, we looked at all of the columns. We examined which ones we needed for further analysis and which ones needed to be dropped just because they weren’t really needed, or because they had way too many missing values, or because they just did not make any sense. For instance, we had columns under the names of “diag\_1, diag\_2, and diag\_3.” There was no given information of what those columns were and what the values meant, so we decided to dropped them.

* 1. *Experimental Design*

You will run your model several times with different parameters to see what different results you get. In a table, describe your experimental parameters. Three or four experiments are sufficient. This is where you will describe how you divided your data into train, validate and test data sets. For example:

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | Test size of 20% with a max depth of 8 |
| 2 | Test size of 10% with a max depth of 8 |
| 3 | Test size of 20% with no depth |
| 4 | Test size of 10% with no depth |
| 5 | Test size of 20% with a max depth of 5 |
| 6 | Test size of 10% with a max depth of 5 |

* 1. *Tools Used*

Describe all of the software tools you used to perform your data preparation and model implementation.

The following tools that were used for this analysis were Python v3.9.7 (64-bit) running the Anaconda 2021.11 for Microsoft Windows 11 was used for all of the analysis and implementation. The following libraries for python were also used including Pandas 1.4.2, Numpy 1.22.0, Matplotlib 3.5.1, Seaborn 0.11.2, SKLearn 1.0.2. We used Pandas for data analysis and data cleaning. We used Numpy to be able to work on mathematical operation on arrays. We used Matplotlib and Seaborn for data visualization to be able to generate graphs from the data that we have, they provide informative and really good-looking statistical graphs with many different colors. We used SKLearn for the decision tree and the linear model regression which is basically machine learning.

1. **RESULTS**
   1. *Classification Measures*

|  |  |  |  |
| --- | --- | --- | --- |
| **Experiment number** | **Accuracy in the form of F1 Scores (Overall Accuracy)** | **F1=0** | **F1=1** |
| 1 | 82 | 79 | 84 |
| 2 | 82 | 79 | 84 |
| 3 | 80 | 78 | 81 |
| 4 | 81 | 80 | 83 |
| 5 | 82 | 79 | 84 |
| 6 | 82 | 79 | 84 |

Note: The dataset has an equal steady amount of insulin levels in patients as to unsteady amount of insulin levels. =1 means steady and =0 means either high or low insulin levels.

* 1. *Discussion of Results*

The first model is better because it’s evenly split between the 5th and 6th trials. While the 5th and 6th trials were equal, they carried a lower entropy which provided a much higher information gain. The second model didn’t work at all because it was trying to use linear regression as well as support but it failed horribly.

* 1. *Comparison of Models*

Each model showed promising predictions and had an accuracy level of at least 0.8 accuracy level; however, the highest accuracy levels were 0.84 for experiment numbers 1,2, 5, and 6.

We used the naïve bayes for the second model and it had a very bad prediction indicating that all the hospital patients are white since that the race of 75% of patients. The model, unable to make true predictions on race from any other part of the dataset, naturally selected the higher proportion.

* 1. *Problems Encountered*

There was an error in reshaping the fit x and y vectors for the decision tree. We needed to reshape the array from 1 dimension into a 2D array before separating the dataset into individual x and y vectors and then use .t to fit the x and y vectors into a 2D array again. The second model tried to use a linear regression model and support vector machine but it was an uttermost failure. Running models were very slow due to the sheer amount of data entries provided totaling 101767 rows. Even after aggregating the insulin column to remove null values, there were still over 50000 data entries for the model to consider

* 1. *Limitations of Implementation*

There a large amount of data entries that lack the test for chemical levels in blood samples. Though there are entries for these categories with a 0 but that doesn’t necessarily indicate unsteady levels in the blood sample. it could also mean that some of those experiments simply were not run. This means that the 0 is non-steady or non-existent.

* 1. *Improvements/Future Work*

A good way to improve our model is to write an algorithm that includes only the entries that had the tests done or systemically go through each test and eliminate the entries that didn’t have any values. This method can only improve the model to an extent however since each consolidated column could reduce the total entries for the model by up to 95%. This is because 95% was the threshold for incomplete data used in the cleaning phase to deduce whether a column had enough entries to prove useful.

1. **CONCLUSION**

In the end, we were able to run some models like the decision tree which actually turned out to be a very good model and it gave us at least 0.8 accuracy or a bit higher, but the naïve bayes definitely needs a lot more work in its algorithm because just like mentioned earlier, it predicted that the patients in the hospitals were white which is actually false since approximately ¼ is African-American, Non-White, or unidentified. We tried using linear regression and support vector machine but unfortunately could not work. For the support vector machine, it kept loading for a very long time so we assumed it wouldn’t load at all and had to stop the kernel from running.

*Division of Labor*

Eric worked on the notebook (coding) while Adham worked on the entire word document. Both people worked on searching the web for troubleshooting and research for the analytical models