



PROJECT FINAL REPORT

Diabetes 130-US hospitals for years 1999-2008 Data Set



MAY 5, 2020

REZA MARZBAN

Table of Contents

Chapter 1: Introduction	2
Chapter 2: Related Works	3
Chapter 3: Technical Methods.....	4
3.1: Data introduction	4
3.2: Preprocessing and Data Cleaning	6
3.3: Regression.....	7
3.4: Classification	7
Chapter 4: Experimental Results	8
4.1: Regression Results.....	8
4.2: Classification Results.....	9
4.2.1: Naïve Bayes	9
4.2.2: Decision Tree.....	10
4.2.3: Logistic Regression	10
4.2.4: Artificial Neural Networks (Multilayer Perceptron).....	10
4.2.5: Model comparison	10
Chapter 5: Conclusions and Discussions	12
References.....	14

Chapter 1: Introduction

With the rapid growth of big data, data mining in healthcare provides health system an efficient way to improve services and optimize profit. According to the American Diabetes Association, diabetes affects more than 34 million Americans with cost 327 billion annually for cost of treatment. This estimate imposes the significant financial burden of diabetes on society. In addition to that, diabetes can cause long-term effects on patients, such as blindness, amputation, and heart disease, as a result, it needs to be studied comprehensively. So, applying predictive models on diabetic patients could potentially inform the management of diabetes-related problems.

In recent years there has been many interesting data mining papers in the healthcare, but there are still unanswered challenges. As we are in the Big Data era, we have been introduced to a huge amount of data that can be useful in our analyses. Nowadays every single patient has a record in the hospital database with so many features that contains from primitive patient information to hospital advanced test results. These features bring both opportunity and challenges. As we have access to patient's information we can generate statistical models with high accuracy and performance that we could not achieve without enough data. Having too much data has many challenges as it would be very hard to process all data in a single machine CPU, and that is the reason behind the usage of cluster technologies like Hadoop and Spark. Another challenge is that; it is very hard to cleanse various types of data.

In this project, we used Diabetes 130-US hospitals for years 1999-2008 Data Set¹. It contains around 102,000 records and 50 columns, features, or variables. We have decided to check the following two hypotheses:

- 1- Predict Time in hospital variable which is an integer number of days between admission and discharge (**Regression**).

Business view: If we can predict time in the hospital, we can optimize hospital room usage, so we will have available rooms for all incoming patients. We want to find the variables that are significant in predicting the patient's length of hospital stay. This understanding would help hospital administration to more dynamically reallocate and optimize hospital beds.

- 2- Predict Change of medications variable which Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change". (**Classification**)

Business view: Hospitals can create an online predictor for patients, and after inserting their data, if they are classified into "Change" medication, they need to go to the hospital to get checked for further diagnosis and prescription. If the prediction is "no change" the patient does not need to revisit the hospital. It will enable hospital administration to reduce the number of patients that do not need to come to hospital in the first place. This will help the physicians to utilize their patient visiting hours in a more efficient way rather than wasting time on invalid cases or postponing the required cases to later dates due to crowd. This also helps in utilizing the hospital resources and equipment's in a more efficient way such that the resources are always available for the needed cases.

After implementing, tuning, and evaluating various models on these two hypothesis, Result showed that the best R squared that we can get from linear regression on hypothesis 1 is around **30.50%** which means there is a relation but the prediction is not good enough. The best Accuracy that we get, was **72.20%**

¹ <http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008>

by our Multi-Layer Perceptron algorithm which was better than our Decision Tree, Naïve Bayes and Logistic Regression. It means that in 72% of the cases, our model predicts correctly. The rest of this paper is organized as follows: in [Chapter 2](#), we review the literature. In [chapter 3](#), we cover the technical aspects of our models. In [chapter 4](#), we present the results of our predictors and compare them and finally in [chapter 5](#), we summarize the paper.

Chapter 2: Related Works

The dataset we analyze, was originally used to predict the readmission rates of each patient from the Diabetes 130-US hospitals for years 1999-2008 Data Set [1]. The platform named KNIME [2] (Konstanz information miner) that is used for this study, is a modular environment, which enables easy visual assembly and interactive execution of a data pipeline.

Although, previously there were several studies attempted by the scientists to create models in order to prove the hypothesizes in diabetics, most of them were focused on proving a relation or correlation among attributes. Stolker [3] studied and evaluated the relationship between A1C and glucose therapy intensification (GTI) in patients with diabetes. Britton [4] proved that there is no association between hemoglobin A1c and in-hospital mortality in patients with diabetes. Shah [5] showed the significance of hemoglobin A1c in predicting the outcome after cardiac resynchronization therapy in patients with diabetes and heart failure. Cios and Moore [6] have discussed the ethical and legal aspects of medical data mining including data ownership, fear of lawsuits, expected benefits, and special administrative issues.

In this study, several different methods and algorithms have been explored. Some of these algorithms are very primitive statistical models, and others are advanced machine learning algorithms. Also we have used Deep Learning or Artificial Neural Networks (ANN) in our process, so all the models can be compared and their performance can be assessed, and the best choice for our data is selected. We started with a naïve primitive Linear Regression [7]. Then Decision Trees [8] was used for classification. We also created a Naïve Bayes [9] model, and logistic regression [10]. In addition to these basic algorithms, Artificial Neural Networks or Multi-Layer Perceptron (MLP) [11], [12], [13] was tested as well. These algorithms is covered in detail in [Chapter 3](#), and their results is reviewed and compared in [Chapter 4](#).

Chapter 3: Technical Methods

3.1: Data introduction

Dataset used as benchmark for this project: Diabetes 130-US hospitals for years 1999-2008. It is published on UCI Machine Learning Repository in 2014. This data includes 102,000 observations and 50 variables as follows (table 1).

Feature name	Type	IV/DV	Description and values	% missing
Encounter ID	Primary Key	IV	Unique identifier of an encounter	0%
Patient number	Primary Key	IV	Unique identifier of a patient	0%
Race	Nominal	IV	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	IV	Values: male, female, and unknown/invalid	0%
Age	Interval	IV	Grouped in 10-year intervals: 0, 10), 10, 20), ..., 90, 100)	0%
Weight	Ratio	IV	Weight in pounds.	97%
Admission type	Nominal	IV	Integer identifier corresponding to 9 distinct values.	0%
Discharge disposition	Nominal	IV	Integer identifier corresponding to 29 distinct values.	0%
Admission source	Nominal	IV	Integer identifier corresponding to 21 distinct values.	0%
Time in hospital	Ratio	DV	Integer number of days between admission and discharge	0%
Payer code	Nominal	IV	Integer identifier corresponding to 23 distinct values.	52%
Medical specialty	Nominal	IV	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values	53%
No of lab procedures	Ratio	IV	Number of lab tests performed during the encounter	0%
No of procedures	Ratio	IV	Number of procedures performed during the encounter	0%
No of medications	Ratio	IV	Number of distinct generic names administered during encounter	0%
No of outpatient visits	Ratio	IV	Number of outpatient visits of patient in year preceding encounter	0%
No of emergency visits	Ratio	IV	Number of emergency visits of the patient in the year preceding the encounter	0%
No of inpatient visits	Ratio	IV	Number of inpatient visits of patient in year preceding encounter	0%
Diagnosis 1	Nominal	IV	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	IV	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	IV	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
No of diagnoses	Ratio	IV	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	IV	Indicates the range of the result or if the test was not taken.	0%
A1c test result	Nominal	IV	Indicates the range of the result or if the test was not taken.	0%
Change of medications	Nominal	DV	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
Diabetes medications	Nominal	IV	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
24 features for medications	Nominal	IV	For the generic names. the feature indicates whether the drug was prescribed or there was a change in the dosage.	0%
Readmitted	Nominal	IV	Days to inpatient readmission.	0%

Table1: variable descriptions

The statistics summary of numerical variables is addressed as follows:

Column Name	Min	Max	Mean	Std. Deviation	Variance	No. Missing
num_lab_procedures	1	132	43.0956	19.6744	387.0805	0
num_medications	1	81	16.0218	8.1276	66.0573	0
num_procedures	0	6	1.3397	1.7058	2.9098	0
number_diagnoses	1	16	7.4226	1.9336	3.7388	0
number_emergency	0	76	0.1978	0.9305	0.8658	0
number_inpatient	0	21	0.6356	1.2629	1.5948	0
number_outpatient	0	42	0.3694	1.2673	1.6060	0
time_in_hospital	1	14	4.3960	2.9851	8.9109	0

Table2: Numerical features statistics

The Health Facts data was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West (16). Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100, and bed size of 14 hospitals is greater than 500. As our analysis is going to be used in US, the data is covering the whole variance of distribution of patients in different geographical places. so, demographics variables are also included in the model.

According to our data, the distribution of race is very close to the actual race distribution in US; it is implication of validity of the data collected. In the table 3 we have included the actual distribution and our data distribution of races for comparison. The available difference is due to the fact that our data is collected from 1999 – 2008, and actual distribution is from 2018. Moreover, these differences can be justified because of nature of races and their tolerance against special diseases meaning that some races may be more vulnerable to special disease. The data is almost equally distributed in gender with 53.76% females and 46.24% male patients.

Race levels	Count	Percentage	Actual race Percentage in US ²
Null (Missing)	2,273	2.23%	-
African-American	19,210	18.88%	12.00%
Asian	641	0.63%	6.00%
Caucasian (White)	76,099	74.78%	60.00%
Hispanic	2,037	2.00%	18.00%
Other	1,506	1.48%	4.00%
Grand Total	101,766	100%	100%

Table3: Patient Race Distribution

² reference: <https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>

3.2: Preprocessing and Data Cleaning

In the first step of the preprocessing stage, we dropped the following 9 columns: “encounter_id, patient_nbr, discharge_disposition_id, admission_source_id, payer_code, weight, medical_specialty, examide, citoglipton”. While 5 of them were simply IDs which were not useful as input to our statistical models, the rest were removed due to high rate of missing value (e.g. Weight had 97% missing value!). As a second step, we transformed the age group to age mean by replacing the age range (like [30-40]) to the midpoint (35).

In the third step, we faced a new challenge with the three columns “Diag1, Diag2, Diag3” that contains the patient’s primary and other diagnoses. Initially, these columns had ICD-9 codes as values and there were around 960 such unique levels (for 960 different diseases). In order to reduce dimensionality and work with data more effectively, it is needed to handle categorical value with large number of levels. We used ICD-9 standard³ and grouped these codes into 19 different classes which resulted in the reduction of the number of unique levels from 960 to 19. We have created columns for each 19 levels and used one-hot encoder representation for the same. Please find below the distribution of the above described 19 classes in figure 1.

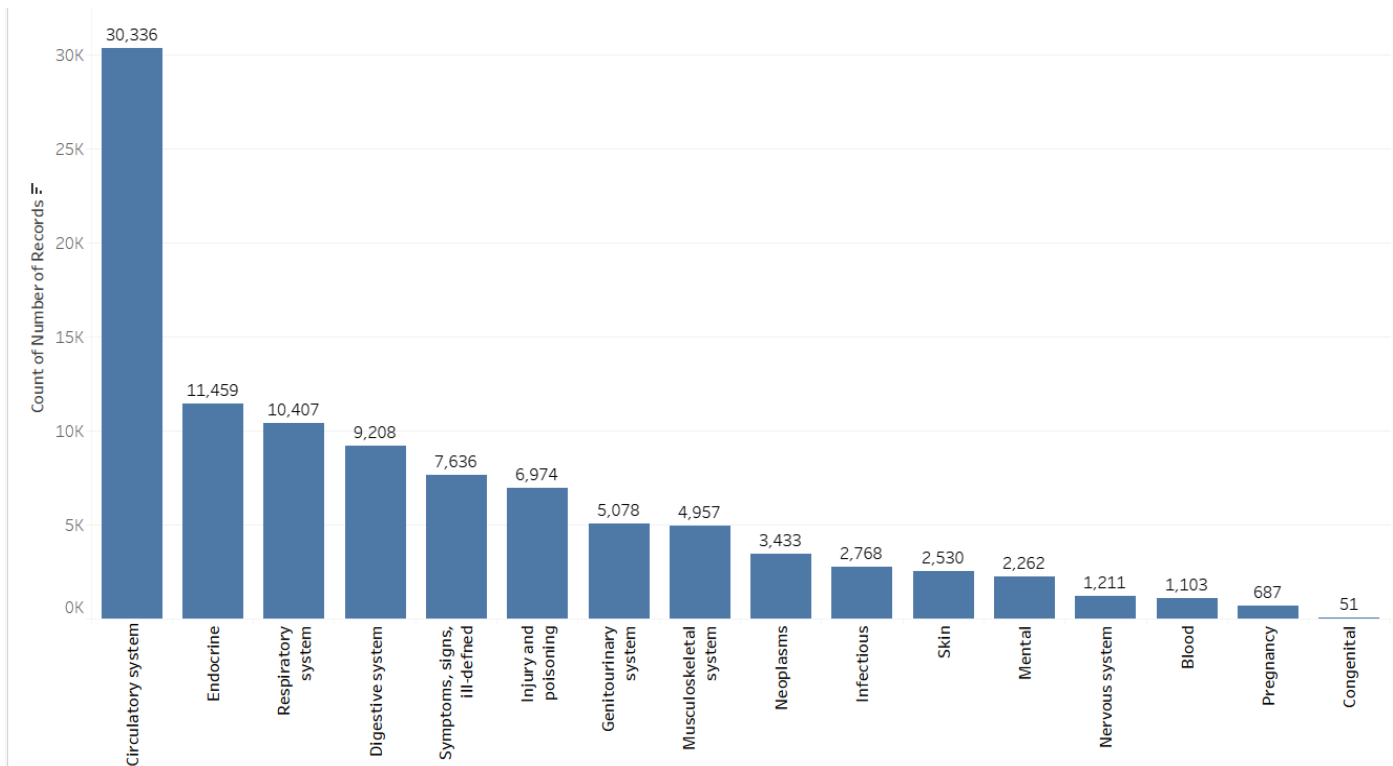


Figure1: Diagnosis classes Distribution

In the fourth step, we transformed the gender feature by assigning 1 to male and 0 to female and used the same one-hot encoder representation for the rest of the nominal features as well. At the end of our preprocessing phase, we partitioned our entire data randomly to training set, and validation set with the ratio of 80:20 respectively.

³ https://en.wikipedia.org/wiki/List_of_ICD-9_codes

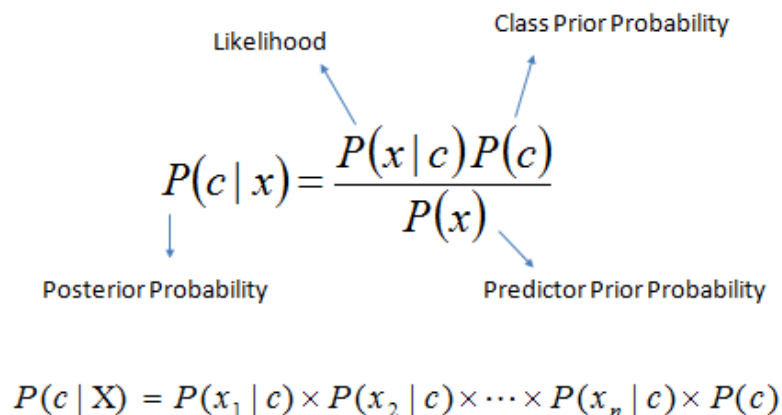
3.3: Regression

In our Regression part, we tried to implement a model to predict each patient “**time in hospital**” which is a number between 1 to 14. We set up a Multiple Linear Regression, and fitted it on train set, evaluated on validation set. Initially, we used all available preprocessed features, and then dropped insignificant variables one by one. After that we chose the most significant features according to their p-value and also their coefficient. Our final model used 25 significant features. In [section 4.1](#), all significant factors and the performance of this part have been discussed.

3.4: Classification

Our second hypothesis was that there is a relation between our independent variables and “**change in medication**”. It is a binary target that shows whether the patient’s medication routine was changed or not. In order to do so, four different Machine Learning algorithms have been implemented and compared. Algorithms used are Naïve Bayes, Decision Tree, Logistic Regression, and Artificial Neural Network (Multilayer Perceptron).

The **Naïve Bayes** algorithm assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence. $P(c|x)$ is the posterior probability of class (target) given predictor (attribute).



The diagram shows the Naïve Bayes formula with labels for its components. The formula is $P(c | x) = \frac{P(x | c)P(c)}{P(x)}$. Arrows point from the labels to the corresponding parts of the formula: 'Likelihood' points to $P(x | c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c | x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

The **Decision Tree** is a decision support tool that uses a tree-like model of decisions and their possible consequences. It is one way to display an algorithm that contains only the conditional control statements.

The **Logistic Regression** is very similar to linear regression, the only difference is that, at the end, sigmoid function is applied on the output to get a number between 0 and 1, which shows the probability of belonging to the positive class (“change” in our specific case.)

The **Artificial Neural Network (Multilayer Perceptron)** is an estimation of human brain neurons. Each perceptron mimics our neurons, and the network has layers that contains many perceptron. Our ANN model contain **6 layers** and each layer have **20 neurons** (perceptron).

[Section 4.2](#) entails the performance of all models mentioned and their comparison.

Chapter 4: Experimental Results

4.1: Regression Results

Score Name	Training set	Test Set
R_Square	0.301	0.305
Mean Absolute Error	1.898	1.889
Mean Squared Error	6.241	6.176
Root Mean Squared Error	2.498	2.485
Mean Signed Difference	-1.458	-0.013
Mean absolute percentage Error	0.665	0.663

Table4: Regression Performance Results

According to table 4, the R_square of our model is around 0.305. which means that our model can predict around 30% of variation in the target variable. As it is obvious from the table 4, all independent variables are significant at the level of 0.05 except three following variables: Change, Congential_diag, Genitourinary_diag.

Age is one of the significant independent variables in explaining Time_in_Hospital. The Coefficient of Age is equal to 1.645, which means if a patients Age, increases by one unit, the time in hospital also increases around 1.645 days which make sense as older patients need more time for recovery.

Another significant IV is num_lab_procedures which indicates the number of lab tests patient needs to perform in their stay. As num_lab_procedures increase by one unit, the time in hospital increases by 3.958 days. Again this seems to be right, as a patient who needs more lab test done should stay longer in hospital for their test result, analysis and further instruction. Other variable details like their coefficients, and t-value is included in table 5. Our linear regression has 25 Independent variables and an intercept (a.k.a. bias).

Variable	Coeff.	Std. Error	t-value	P> t
age	1.645	0.055	30.113	0.000
num_lab_procedures	3.958	0.062	63.756	0.000
num_procedures	0.538	0.035	15.339	0.000
num_medications	10.984	0.106	103.970	0.000
number_emergency	-3.988	0.787	-5.068	0.000
number_diagnoses	1.369	0.077	17.829	0.000
change	-0.010	0.018	-0.529	0.597
pregnancy_diag	0.328	0.112	2.926	0.003
Endocrine_diag	-0.052	0.021	-2.437	0.015
Infectious_diag	0.435	0.038	11.362	0.000
NEoplasms_diag	0.519	0.039	13.177	0.000
circulatory_diag	-0.416	0.024	-17.474	0.000
Respiratory_diag	0.240	0.024	9.812	0.000
Injury_diag	0.190	0.033	5.775	0.000
Skin_diag	1.053	0.036	29.297	0.000
musculoskeletal_diag	-0.238	0.036	-6.695	0.000
Digestive_diag	0.177	0.029	6.184	0.000
Congenital_diag	-0.263	0.175	-1.503	0.133
Genitourinary_diag	0.050	0.026	1.944	0.052
symptoms_diag	-0.372	0.026	-14.126	0.000
Mental_diag	0.823	0.038	21.583	0.000
Nervous_diag	0.322	0.046	7.050	0.000
V_diag	0.617	0.037	16.810	0.000
E_diag	-0.572	0.070	-8.149	0.000
blood_diag	-0.117	0.038	-3.096	0.002
Intercept	-0.737	0.059	-12.411	0.000

Table5: Regression variables statistics

4.2: Classification Results

To predict the intention of change in medicine in treatment for patients, various predictive models are deployed to examine the accuracy rate. Those results are found in given tables as follows:

4.2.1: Naïve Bayes

The final test accuracy of our Naïve Bayes model is **54.20%**. you can find its confusion matrix in table 6.

Naïve Bayes Confusion Matrix		Predicted	
		Change	No_Change
Actual	Change	6	9324
	No_Change	0	11024

Table6: Naïve Bayes Confusion Matrix

4.2.2: Decision Tree

The final test accuracy of our Decision Tree model is **66.30%**. you can find its confusion matrix in table 7.

Decision Tree Confusion Matrix		Predicted	
		Change	No_Change
Actual	Change	5908	3422
	No_Change	3431	7593

Table7: Decision Tree Confusion Matrix

4.2.3: Logistic Regression

The final test accuracy of our Logistic Regression model is **71.90%**. you can find its confusion matrix in table 8.

Logistic Regression Confusion Matrix		Predicted	
		Change	No_Change
Actual	Change	7899	1431
	No_Change	4288	6736

Table8: Logistic Regression Confusion Matrix

4.2.4: Artificial Neural Networks (Multilayer Perceptron)

The final test accuracy of our ANN model is **72.30%**. you can find its confusion matrix in table 9.

Artificial Neural Network Confusion Matrix		Predicted	
		Change	No_Change
Actual	Change	7907	1423
	No_Change	4215	6809

Table9: ANN Confusion Matrix

4.2.5: Model comparison

In table 10, It is observed that, Artificial Neural Network's performance is better than any other algorithm, although in Naïve Bayes, Precision is higher, but that is due to the fact that, Naïve Bayes is predicting almost all as No_change.

	Accuracy	Recall	Precision	F-measure	Area Under Curve ⁴
Naïve Bayes	54.20%	00.10%	100.00%	00.10%	0.7730
Decision Tree	66.30%	63.30%	63.30%	63.30%	0.7018
Logistic Regression	71.90%	84.66%	61.10%	73.40%	0.8018
Artificial Neural Network	72.30%	84.74%	65.20%	73.70%	0.8029

Table10: Classification models statistics

According to table 10, the accuracy rate of Artificial Neural model is highest among other models. Therefore, Artificial Neural Network is selected as the best model.

As you can see in the ROC curve shown in figure 2, Logistic Regression and ANN algorithms are working almost identical, in some rare situations, ANN works a little better on our data, however logistic regression is much faster in training.

⁴ Calculated and shown in figure 2.

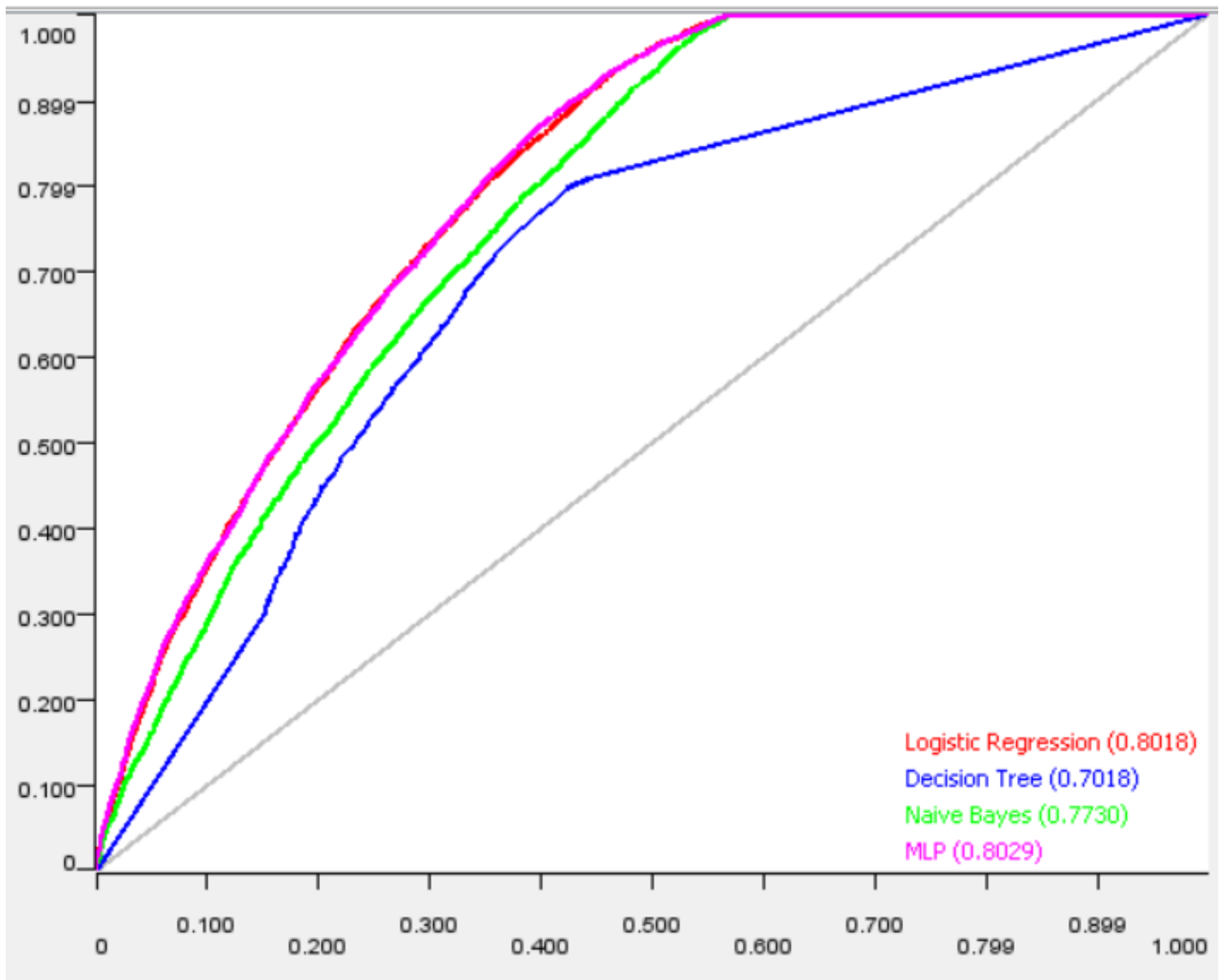


Figure2: ROC Curve and Area Under Curves (AUC)

Chapter 5: Conclusions and Discussions

According to American Diabetes Association's report in 2018, the cost of diagnosed diabetes increased from \$245 billion in 2012 to \$327 billion in 2017, which accounts for around 14% of healthcare cost. It means that one out of seven dollar spent on healthcare is spent for diabetes treatment. This cause a considerable pressure on the US economy.

Given that context, we used Diabetes 130-US hospitals for years 1999-2008 dataset as the benchmark to design, implement and compare several models. Predicting length of time needed for diabetic patients to be cured in the hospital as well as predicting the probability of change in their medicine are two important features which have been overlooked in the studies that have been done so far. These two predictions can considerably enhance the performance of hospitals in managing the situations and offering services effectively. These predictions are significantly important since the number of people who have been diagnosed as diabetics' people is high and is going higher. First we created a multiple linear regression to predict the time in hospital. We realized that the top 5 most important variables are num_medications, number_emergency, num_lab_procedures, age, number_diagnoses in order of importance. These have the highest coefficient absolute values in our regression model, and all of their p-values are equal to zero. num_medications seem to be the most important variable, and if we increase it by one unit though all other IVs do not change, the time in hospital increases by around 11 days.

In the classification models that we created to predict the change in medication, ANN seem to be most effective. Since Deep Learning algorithms are like black boxes, and are not easily interpretable, we interpret our other models. In our Logistic Regression, num_medications, number_emergency are significant with a high absolute value of coefficient.

In our Decision Tree, the first attribute that has been split on our data is diabetesMed, which shows if a patient is currently using diabetic medication or not. If the patient is not using it, the model predicts **no_change** immediately without any further analysis, but if they are using these medications, they need further information. It makes sense because if the patients are using diabetics medication like Insulin, their body is not as stable as others, and their need for medication changes by the time, so we need further info to predict. The next variable that has been split on our data is num_medication, which clearly shows that the increase in the number of medications for a patient will increase the probability of change in medication as well. If a patient is using diabetes medication and their normalized num_medication is less than 0.17, there is a 50/50 chance in the change of medication, but if it is above 0.17, the chance of changing medication increases to 68%.

Hence from the above observations, the Num_Medications plays a vital role as the increase or decrease in the number of medicines that the patient consumes will definitely impact the prediction of "Change In Medication". Hence, there should be more emphasis on the careful consideration of the Number of Medicines prescribed for the patients as this has a direct adverse impact for the "changeInMedicine" if they are wrongly prescribed. This may even lead to a considerable amount of waste of time and effort by the physicians. Also, it can result in the wastage of resources in the hospital which can be better utilized for other patients who really has the requirement or in much need of it. Focus on Num_Medications will definitely save the time, effort, resource and money in an efficient way that helps for better treatments which in turn helps to improve the hospital business. On the whole, we recommend that the patients who are using diabetic medications and has a high number of medication usage on a daily basis, must visit their hospital to update their prescription more frequently.

The future scope of this study can be extended in applying the learned approaches on different diseases such as infectious diseases. In addition to that, the future scope includes bringing up deep learning models with time series to see that after implementing this project, how well we can see the improvement in the changes that have been brought to the business as a result of it.

References

- [1] B. Strack *et al.*, “Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records,” *Biomed Res. Int.*, 2014, doi: 10.1155/2014/781670.
- [2] M. R. Berthold *et al.*, “KNIME: The konstanz information miner,” in *4th International Industrial Simulation Conference 2006, ISC 2006*, 2006, doi: 10.1145/1656274.1656280.
- [3] J. M. Stolker *et al.*, “Relationship between glycosylated hemoglobin assessment and glucose therapy intensification in patients with diabetes hospitalized for acute myocardial infarction,” *Diabetes Care*, 2012, doi: 10.2337/dc11-1839.
- [4] K. A. Britton *et al.*, “No association between hemoglobin A1c and in-hospital mortality in patients with diabetes and acute myocardial infarction,” *Am. Heart J.*, 2011, doi: 10.1016/j.ahj.2010.12.004.
- [5] R. V. Shah *et al.*, “Usefulness of hemoglobin A 1c to predict outcome after cardiac resynchronization therapy in patients with diabetes mellitus and heart failure,” *Am. J. Cardiol.*, 2012, doi: 10.1016/j.amjcard.2012.04.056.
- [6] K. J. Cios and G. William Moore, “Uniqueness of medical data mining,” *Artif. Intell. Med.*, 2002, doi: 10.1016/S0933-3657(02)00049-0.
- [7] V. A. Barbur, D. C. Montgomery, and E. A. Peck, “Introduction to Linear Regression Analysis,” *Stat.*, 1994, doi: 10.2307/2348362.
- [8] S. R. Safavian and D. Landgrebe, “A Survey of Decision Tree Classifier Methodology,” *IEEE Trans. Syst. Man Cybern.*, 1991, doi: 10.1109/21.97458.
- [9] H. Zhang, “The optimality of Naive Bayes,” in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, 2004.
- [10] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, “An introduction to logistic regression analysis and reporting,” *J. Educ. Res.*, 2002, doi: 10.1080/00220670209598786.
- [11] L. Noriega, “Multilayer perceptron tutorial,” *Sch. Comput. Staff. Univ.*, 2005.
- [12] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*. 2015, doi: 10.1038/nature14539.
- [13] R. Tadeusiewicz, “Neural networks: A comprehensive foundation,” *Control Eng. Pract.*, 1995, doi: 10.1016/0967-0661(95)90080-2.