HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 4 (2022)

Predicting Diabetic Retinopathy Using Machine Learning

Nathan Zhang

Watchung Hills Regional High School, Warren, NJ, USA

ABSTRACT

Over the last 3 decades, advances in laser surgery and intraocular drug delivery have decreased the risk of diabetic retinopathy. However, the use of artificial intelligence and other forms of data analysis in retinopathy screening has not had the same advances. This project aimed to analyze the National Health and Nutrition Examination Survey and to develop the best machine learning model to predict the disease. Violin plots were created to compare the distribution of diabetic retinopathy diagnosis with gender. The plot showed that females who had said they were not taking insulin were less likely to be diagnosed with diabetic retinopathy. Another violin plot showed that those with hypertension who were taking insulin were less likely to have diabetic retinopathy. Histograms were also created to show the distributions of the variables, color-coded by retinopathy diagnosis. Specifically, it was found that the bigger the time gaps since the diabetes diagnosis, the more likely a person suffers from diabetic retinopathy. Finally, multiple machine learning models were tested and these were the most accurate in predicting diabetic retinopathy with an 80% accuracy, LinearSVC, CalibratedClassifierCV, and Logistic Regression.

Introduction

Diabetic Retinopathy is the leading cause of visual impairment among adults, with an estimated thirty-seven million people globally with this debilitating disease [1]. Currently, the best prevention for Diabetic Retinopathy is early detection through screening by ophthalmologists. However, with the number of patients with diabetes expected to increase from 422 million people in 2021 to 642 million in 2040, the burden and cost of screening will pose a significant challenge for all those involved [2]. Furthermore, patients living in poor and minority communities are unlikely to even get a screening, considering its cost [3]. This is especially significant considering diabetes disproportionately affects ethnic minorities [4]. The use of machine learning to identify an individual's risk for diabetic retinopathy could improve the efficacy of screening and lower the costs of such a procedure, and allow these underserved communities to get treatment for the disease.

The motivations for this research and its importance stems from several existing studies. Firstly, Machine Learning is being used in many different applications in the medical field. Medical error is the third leading cause of death after heart failure and cancer, and so reducing this error could have a substantial effect on the survival rate of a disease. The increasing complexity of diagnostics and poorly coordinated medical care make misdiagnosis common. Through ML, precision medicine has the potential to decrease the misdiagnosis rate and to give more accurate treatments tailored to the specific patient. Kalavar et al. (2021) mark the many applications of Artificial Intelligence in diabetic retinopathy, including analysis of risk factors and analysis of color fundus photography and other retinal images [6]. Because treatments of diabetic retinopathy can include laser photocoagulation, intraocular corticosteroids, and anti-VEGF agents, identifying individuals most likely to get Diabetic Retinopathy early is paramount to avoiding the debilitating effects from the treatments. ML algorithms of previous studies could determine the risk of diabetic patients with no Diabetic Retinopathy developing it within two years up to 0.71. DL algorithms could predict the progression from early NPDR to PDR

as high as 0.968 [7]. These studies provide promise in an algorithm that can identify high-risk patients with Diabetic Retinopathy.

However, the most effective machine learning model for analyzing the risk of Diabetic Retinopathy is still unclear, and earlier studies fail to use many factors that could contribute to an individual's risk profile. Our study aims to use exploratory analysis through various relationships between variables to find what variables had the greatest correlation with diabetic retinopathy, and find machine learning models that most accurately predicted the disease.

Data and Discussion

This study used data from the National Health and Nutrition Examination Survey to analyze and predict the probability of Diabetic Retinopathy. NHANES is a series of probability surveys designed to obtain information on the health and nutritional status of the US population. NHANES data are collected every 2 years serving as 1 analytical cycle, including 5 cycles (NHANES 2013-2014, NHANES 2011-2012, NHANES 2009-2010, NHANES 2007-2008, NHANES 2005-2006 data) of samples of adults aged who have been diagnosed with diabetes. The data are collected through household interviews by the National Center for Health Statistics, CDC, and are intended to be representative of the US population. The specific variables our study included in our analysis were the following (presented in Table 1): if the patient has retinopathy (DIQ080), if the patient is now taking insulin (DIQ050), gender (RIAGENDR), race/ethnicity (RIDRETH1), education level (DMDEDUC2), marital status (DMDMARTL), family poverty income ratio (INDFMPIR), age told of hypertension or high blood pressure (hyperten), age at screening (RIDAGEYR), health insurance status (insurance), A1C level (DIQ280), age at screening minus age told having diabetes (dm_y5).

Table 1. Variables from NHANES dataset [8]

Name	**Description**	**Value**
DIQ050	Are you now taking insulin	1 - Yes 2 - No
DIQ080	Has a doctor ever told that diabetes has affected eyes or that had retinopathy	1 - Yes 2 - No
RIAGENDR	Gender	1 - Male 2 - Female
RIDRETH1	Race/Hispanic origin	1 - Mexican American 2- Other Hispanic 3 - Non-Hispanic White 4- Non-Hispanic Black 5 - Other Race Including Multi-Racial
DMDEDUC2	highest grade or level of education completed by adults 20 years and older	If **DMDEDUC2**=5 then DMDEDUC2=5; else DMDEDUC2=1;
DMDMARTL	Marital Status	If **DMDMARTL**>=2 then DMDMARTL=2;

INDFMPIR	Ratio of family income to poverty	If **INDFMPIR**<=1 then INDFMPIR=1; else INDFMPIR=2;
hyperten	Whether you have hypertension	If BPXSY1>=140 or BPXDI1>=90 then **hyperten**=1; else hyperten=0;
dm_y5	Age at screening minus age told having diabetes	**dm_y5**=(RIDAGEYR-DID040)/5
controlled	Level of control over diabetes.	if DIQ280<7 then controlled=1; else if DIQ280<9 then controlled=2; else controlled=3;

These variables were compiled into a CSV file that was analyzed using Python language through Google Colab, the IDE used in the study. Two libraries NumPy and Pandas were imported into Colab, with NumPy handling arrays and matrices and pandas handling data processing. Furthermore, matplotlib and seaborn libraries were imported to handle data visualization.

Next, the .unique() method was used on the output variable DIQ080 to display the possible values of the variable. To calculate how many patients in the data suffered from diabetic retinopathy, the .value_counts() method was used to return the count for each unique value in the column DIQ080. With the count, the relative frequency of people who had the disease could be determined by setting the normalized parameter to true.

To see if there were outliers present in the data, boxplots were made of the data frame. To make these, a function named hist_all was created, taking the inputs data and outcome. First, the outcome variable DIQ080 removed a list of the dataframe's columns. The resulting list was assigned to a variable Xcol. Then, using matplotlib.pyplot's subplots function, the number of subplots, and the figure size were set. By looping through Xcol, the resulting figure below could be created through matplotlib.pyplot.

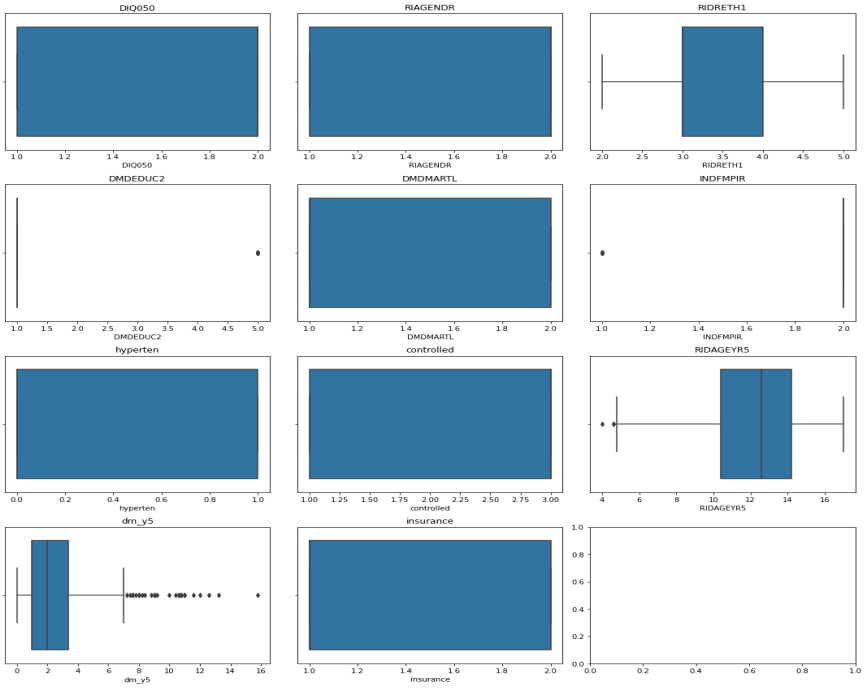

Figure 1. Boxplots of the variables from the NHANES dataset

In figure 1, we can see that the median age of screening is around 65 (RIDAGETR = Age/5). Q1 occurs slightly after the age of 50, and Q3 occurs slightly after 70. The maximum is around 85, while the minimum is around 25. Two outliers occur around 20. Furthermore, from the INDFMPIR boxplot, we can see that the majority of people in the dataset have a family income to poverty ratio greater than 1. It can also be inferred from the DMDEDUC2 boxplot that most of the people in the dataset are not college graduates. Also, from the dm_y5 boxplot, the median number of years between diabetes diagnosis and screening is 2 years, with Q1 being around 1 year and Q3 being around 3 years. The maximum occurs at around 7 years, with numerous outliers occurring from slightly above 7 years to 16 years. The minimum is slightly above 0 years.

To show the distribution of each variable's data, histograms were made that were color-coded with the response from the output variable, DIQ080. The histograms were created in a similar manner to the boxplots.

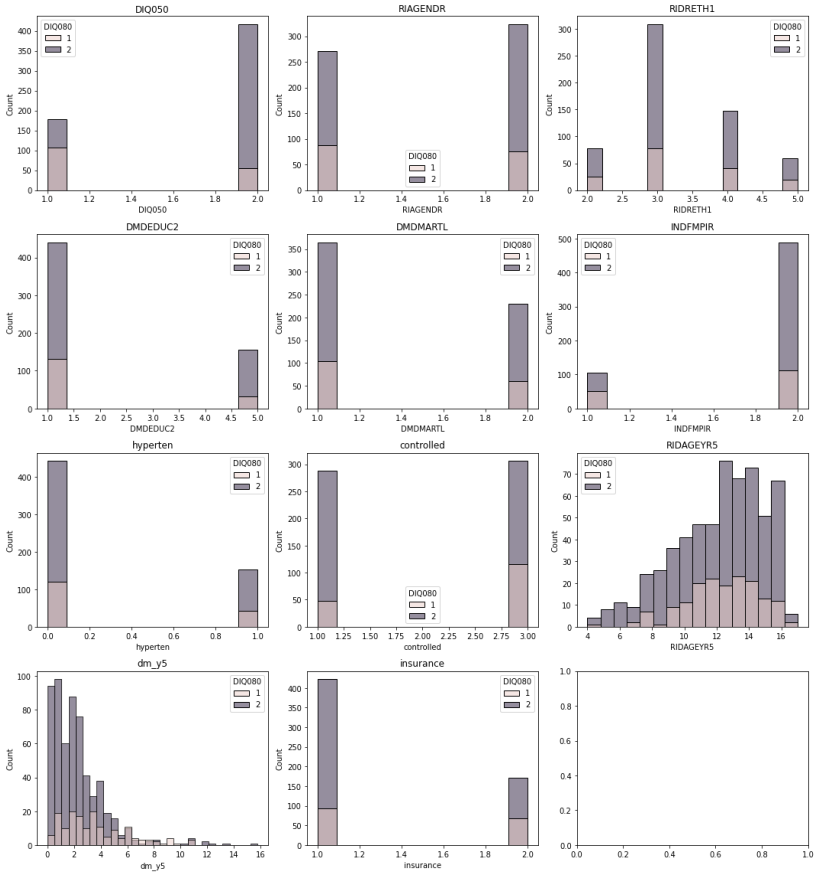

Figure 2. Histograms of the NHANES variables color-coded by the DIQ080

From figure 2, it is shown that a greater proportion of people who say they take insulin have been diagnosed with diabetic retinopathy than those who say they do not. There does not seem to be a significant difference between genders and races with a diagnosis of diabetic retinopathy. It does seem, however, that a lesser proportion of college graduates have diabetic retinopathy than those who are not. Marital status seems to have no effect on the diagnosis of diabetic retinopathy. But it is clear that low-income families have a higher incidence of diabetic retinopathy than those of higher income. Hypertension or high blood pressure seems to not affect the diabetic retinopathy diagnosis. Also, a higher proportion of individuals with an A1C level between 7 and 9 got diabetic retinopathy compared to those with an A1C level below 7. There does not seem to be any correlation between age at screening and diabetic retinopathy diagnosis. There does seem to be a correlation

between the years between screening and diabetes diagnosis and diabetic retinopathy diagnosis, where the larger the number of years the higher proportion gets diabetic retinopathy. Also, it seems that a higher proportion of people that are covered by Medicare or Medicaid have diabetic retinopathy compared to those who have private insurance.

Next, a correlation heat map was constructed using .corr() on the data frame to calculate the pairwise correlation between all the variables in the CSV file. To visualize the matrix, we use the heatmap function from seaborn.

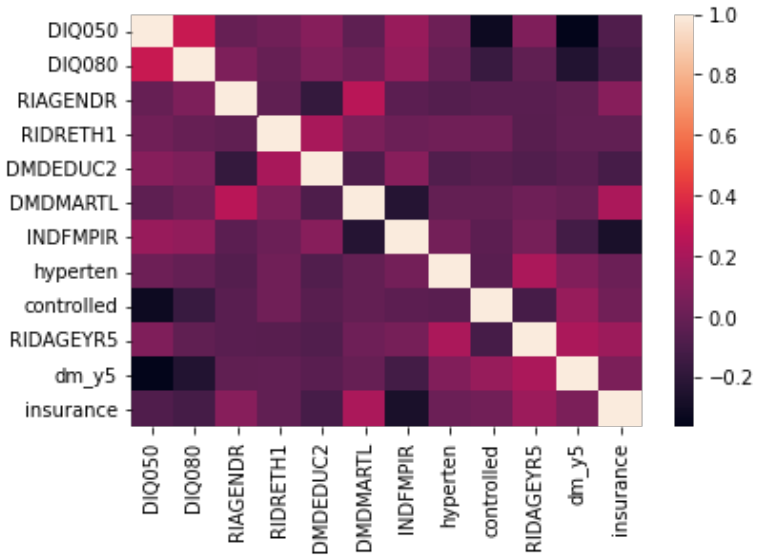

Figure 3. Correlation heatmap of dataframe

In figure 3, DIQ050 and DIQ080 have a strong positive correlation with each other. On the other hand, pairs of variables including dm_y5 and DIQ050, controlled and DIQ050, INDFMPIR and insurance, DMDMARTL and RIAGENDR and DMDEDUC2 all have no correlation. Furthermore, RIAGENDR and DMDMARTL have a positive correlation with each other.

Because most of the columns are categorical, violin plots were constructed between the variables of the data frame. Violin plot has a vertical kernel density plot for each category and a small box plot to summarize important statistics. For example, DIQ050 can be on the x-axis, DIQ080 on the y axis, with the “hue” parameter set to “RIAGENDR”, split=‘True’.

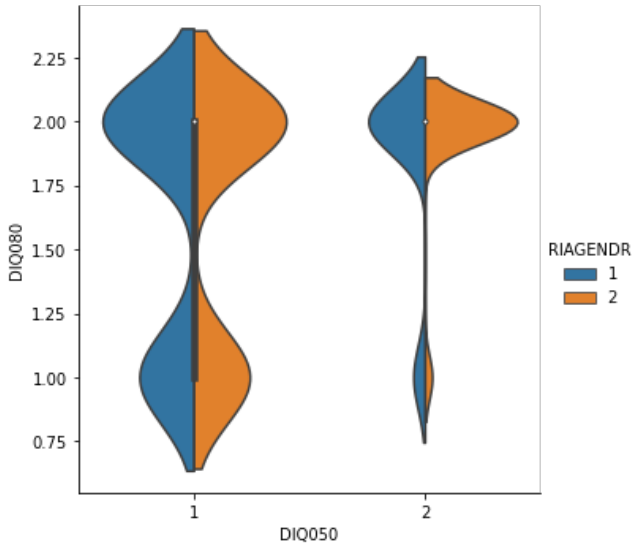

Figure 4. Violin plot comparing DIQ050 and DIQ080 with hue='RIAGENDR' and split='True

Figure 4 shows that there are much more people who do not take insulin and are females who do not have diabetic retinopathy than people who do not take insulin and are male. Furthermore, there does not seem to be any difference between males and females who take insulin in their diabetic retinopathy diagnosis. Although it can be clearly seen that those who take insulin are much less likely to not have diabetic retinopathy than those who do not take insulin. There doesn't seem to be much of a difference between genders for those that take insulin.

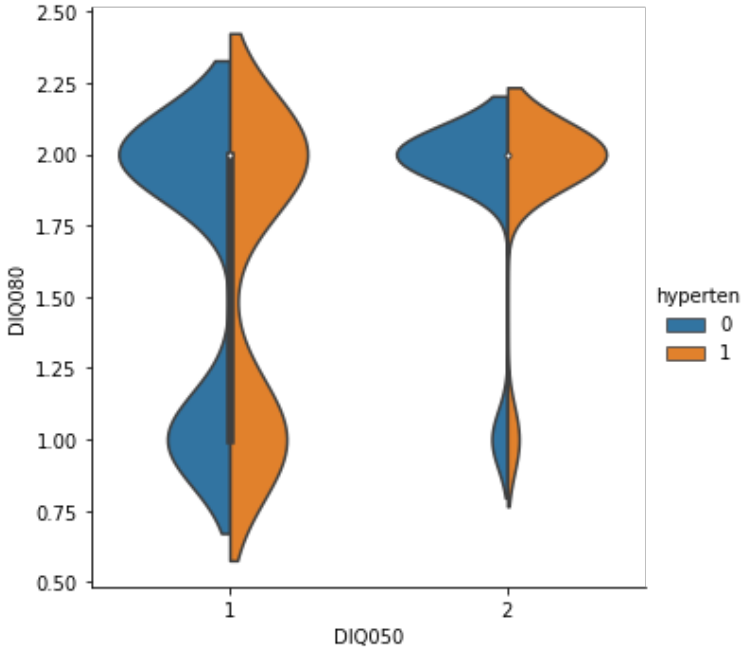

Figure 5. Violin plot comparing DIQ050 and DIQ080 with hue='hyperten' and split='True'

Figure 5 shows that those who do not take insulin and do not have diabetic retinopathy have a higher proportion with no hypertension than those with hypertension. And another plot by changing the x-axis to 'INDFMPIR' and the hue parameter to 'DMDEDUC2'.

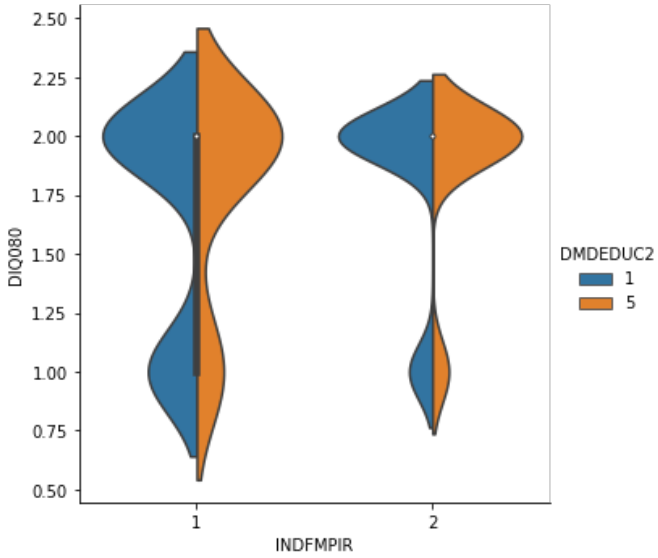

Figure 6. Violin plot comparing INDFMPIR and DIQ080 with hue='DMDEDUC2' and split='True'

From figure 6, it is clear that for those with diabetic retinopathy, a greater proportion of them are not college graduates.

Models

Table 2. ML Models

Nearest centroid classifier	A classification model that assigns to observations the label of the class of training samples whose mean (centroid) is closest to the observation [9]. Accuracy of 0.69
Bernoulli Naive Bayes	It is a variant of Naive Bayes that follows BernoulliNB event model for discrete data. BernoulliNB is designed for binary/boolean features [9]. Accuracy of 0.78
Gaussian Naive Bayes	It is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data [9]. Accuracy of 0.77
Decision Tree Classifier	It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome [9]. Accuracy of 0.75
Bagging Classifier	It is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction [9]. Accuracy of 0.77
Light GBM Classifier	Light GBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification and many other machine learning tasks [9]. Accuracy of 0.77
Random Forest Classifier	It is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset [9]. Accuracy of 0.79
AdaBoost Classifier	It is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases [9]. Accuracy of 0.78
Extra Tree Classifier	Extremely Randomized Trees Classifier(Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it's classification result [9]. Accuracy of 0.72
Logistic Regression	It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation [9]. Accuracy of 0.80
Linear Discriminant Analysis	It is closely related to analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements [9]. Accuracy of 0.79
XGBoost Classifier	XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning [9]. Accuracy of 0.75
Linear SVC	The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is [9]. Accuracy of 0.80
Calibrated Classifier CV	Probability calibration with isotonic regression or logistic regression [9]. Accuracy of 0.80
Perceptron	It is an algorithm for supervised learning of binary classifiers [9]. Accuracy of 0.69
Quadratic Discriminant Analysis	Quadratic Discriminant Analysis (QDA) is a generative model. QDA assumes that each class follow a Gaussian distribution [9]. Accuracy of 0.76

Dummy Classifier	A dummy classifier is a type of classifier which does not generate any insight about the data and classifies the given data using only simple rules [9]. Accuracy of 0.70
Label Spreading	It is similar to the basic Label Propagation algorithm, but uses affinity matrix based on the normalized graph Laplacian and soft clamping across the labels [9]. Accuracy of 0.69
Label Propagation	It is a semi-supervised machine learning algorithm that assigns labels to previously unlabeled data points [9]. Accuracy of 0.69
Ridge Classifier	It is based on Ridge regression method, converts the label data into [-1, 1] and solves the problem with regression method [9]. Accuracy of 0.79
Ridge Classifier CV	Ridge classifier with built-in cross-validation [9]. Accuracy of 0.79
SGD Classifier	Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression [9]. Accuracy of 0.74
Extra Trees Classifier	This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [9]. Accuracy of 0.74
K Neighbors Classifier	The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems [9]. Accuracy of 0.78
C-Support Vector Classification	The implementation is based on libsvm. The fit time scales at least quadratically with the number of samples and maybe impractical beyond tens of thousands of samples [9]. Accuracy of 0.78
Passive Aggressive Classifier	It is a family of Machine learning algorithms that are popularly used in big data applications. It works by responding as passive for correct classifications and responding as aggressive for any miscalculation [9]. Accuracy of 0.63

Conclusion

In this study, LinearSVC, CalibratedClassifierCV, and Logistic Regression was found to be the most accurate algorithms in predicting diabetic retinopathy, achieving 80% accuracy. Contrary to the traditional method of screening, which many low-income diabetics cannot get, machine learning can give patients an understanding of their diabetic health without having to go through the hassle of a clinic checkup. Furthermore, predicting diabetic retinopathy with an ML algorithm can help reduce human error in the diagnosis of the disease, and therefore deaths that can result from it. Further research can be done in improving the accuracy of the ML algorithm and adding more variables to the dataset for a complete picture of the patient's health. The data analysis process in this study can be applied to other diseases to improve diagnostic quality and lower its cost making it more accessible to the underserved community.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] World Health Organization. (n.d.). Diabetes. World Health Organization. Retrieved December 22, 2021, from https://www.who.int/health-topics/diabetes#tab=tab_1

[2] Predicting the risk of developing diabetic retinopathy using deep learning. thelancet. (2021, January). Retrieved December 22, 2021, from https://www.thelancet.com/action/showPdf?pii=S2589-7500%2820%2930250-8

[3] Avidor, D., Loewenstein, A., Waisbourd, M., & Nutman, A. (2020, April 6). Cost-effectiveness of diabetic retinopathy screening programs using telemedicine: A systematic review. Cost effectiveness and resource allocation : C/E. Retrieved December 24, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7137317/

[4] Barsegian, A., Kotlyar, B., Lee, J., Salifu, M. O., & McFarlane, S. I. (2017). Diabetic retinopathy: Focus on minority populations. International journal of clinical endocrinology and metabolism. Retrieved December 24, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5945200/#R10

[5] Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. Q. (2020, January 1). Artificial Intelligence with multi-functional machine learning platform development for better healthcare and Precision Medicine. Database : the journal of biological databases and curation. Retrieved December 24, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7078068/

[6] Wu, J.-H., Liu, T. Y. A., Hsu, W.-T., Ho, J. H.-C., & Lee, C.-C. (2021, July 3). Performance and limitation of machine learning algorithms for diabetic retinopathy screening: Meta-analysis. Journal of medical Internet research. Retrieved December 22, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8406115/

[7] Kalavar, M., Al-Khersan, H., Sridhar, J., Gorniak, R. J., Lakhani, P. C., Flanders, A. E., & Kuriyan, A. E. (2020). Applications of artificial intelligence for the detection, management, and treatment of diabetic retinopathy. International ophthalmology clinics. Retrieved December 23, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8514105/

[8] Centers for Disease Control and Prevention. (n.d.). *Nhanes Questionnaire Data*. Centers for Disease Control and Prevention. Retrieved July 24, 2022, from https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire

[9] Learn. scikit. (n.d.). Retrieved December 28, 2021, from https://scikit-learn.org/stable/index.html