

Performance of ML Algorithms for Diabetes Diagnosis

Category: Medicine & Health

Aayan Mukul

The Gwinnett School of Mathematics, Science, and Technology

970 McElvaney Ln NW, Lawrenceville, GA 30044

Table of Contents

1. Introduction.....	3
1.1 Rationale.....	3
1.2 Research Question.....	4
1.3 Hypothesis & Null Hypothesis.....	5
2. Methodology.....	5
2.1 Procedure.....	5
Part A. Dataset.....	5
Part B. Data Preprocessing.....	6
Part C. Training Data & Testing Data.....	6
Part D. Data Collection.....	7
Part E. Procedure Diagram.....	7
2.2 Variables.....	7
3. Results.....	8
3.1 Data Tables.....	8
3.2 Graphs + Models	9
4. Discussion.....	13
5. Conclusion.....	15
6. Acknowledgment of Major Assistance	16
7. References.....	17

1. Introduction

1.1 Rationale:

In Native American societies, diabetes is a major health issue that results in dangerous consequences. According to the Indian Health Service (IHS), Native Americans have the highest rate of type 2 diabetes among all racial and ethnic groups in the United States of America (Indian Health Service, 2016). In addition, Native Americans are nearly three times as likely to develop diabetes compared to non-Native Americans (CDC, 2019).

Several reasons are responsible for the high rate of diabetes in Native American communities. One of the main issues is lack of access to healthcare. Many Native American communities are located in rural areas with limited access to medical facilities. This results in difficulty for people to access necessary healthcare, including diabetes diagnosis and treatment.

Another reason is that Native Americans are more likely to carry risk factors for diabetes—particularly obesity (Indian Health Service. 2021). Obesity is a major risk factor for diabetes, and Native Americans are more likely to be obese than other racial and ethnic groups in the United States (CDC, 2019). This results from a lack of access to healthy food options in Native American communities, as Native Americans are more likely to have a diet high in refined grains and sugar, which increases the risk of diabetes (Schure *et. al.*, 2019).

The consequences of diabetes are detrimental in Native American communities. Native Americans with diabetes are more likely to experience health issues such as heart disease, blindness, kidney failure, and amputations (National Research Council, 1996). In addition, diabetes is one of the leading causes of death in Native American communities (Indian Health Service, 2011).

Machine learning algorithms could potentially be used to improve the diagnosis and treatment of diabetes in Native American communities. These algorithms can analyze patterns in electronic health records to diagnose Native Americans for diabetes. Electronic health records can provide a wealth of information about a person's health, including risk factors for diabetes such as age, weight, and family history. By training a machine learning model on the electronic health record data, the algorithms can identify patterns that are predictive of diabetes and use this information to identify patients who are at high risk of developing diabetes and intervene earlier to prevent the disease.

Several different types of machine learning models can attempt to identify patterns that are predictive of diabetes; however, each model has varying results. Because different models have varying levels of accuracy and different models are suited for different types of data, it is important to select the best type of machine learning model to be deployed into healthcare facilities to predict diabetes in Native American communities.

In conclusion, diabetes is a major global health issue, especially in Native American communities. The lack of access to healthcare and high rates of obesity contribute to the large rates of diabetes in these communities. Machine learning algorithms have the ability to help improve the diagnosis and treatment of diabetes in Native American communities, and research is needed to determine which machine learning model is best suited for deployment in this scenario.

1.2 Research Question:

What is the accuracy of machine learning models for predicting diabetes based on data collected from Native American tribes?

1.3 Hypothesis & Null Hypothesis:

Hypothesis:

Logistic regression will have the greatest accuracy for predicting diabetes in Native American tribes because it returns a binomial result—diabetic or not—and avoids confounding effects by analyzing the association of all the variables relative to each other, resulting in the greatest accuracy through training itself from a multivariable perspective (Sperandei, 2014).

Null Hypothesis:

There will be no significant difference in the accuracy of predicting diabetes diagnosis in Native American tribes between machine learning models.

2. Methodology

2.1 Procedure

Part A. Dataset

In this research, the data about signs of diabetes is not collected by the researcher. The dataset is publicly available on the website Kaggle. The dataset is taken from the National Institute of Diabetes and Digestive and Kidney Diseases which collected data on female Pima Indians. The dataset used in this research study contains the data of the Native Americans who lived there and contains information on the number of pregnancies, glucose and insulin levels, body mass index (BMI), and other statistics about them.

Part B. Data Preprocessing

Data preprocessing is an extremely important step to enhance the quality of the data. This step helps in the feature extraction and analysis, resulting in the best possible machine learning model outcome. To begin, after acquiring the dataset and importing it into the Python notebook, import all the required libraries: pandas, numpy, MatLab, seaborn, and sk-learn. Afterward, identify and handle all the missing values in the used dataset and split the dataset into training data and testing data—in this research, the entire dataset had zero null or missing values. The data instances will aid the machine learning model in diagnosing the female Pima Indians with diabetes and help get better machine learning model results.

Part C. Training Data and Testing Data

During the experiment, the dataset was divided into training and testing data. The dataset contains 768 rows and 9 columns. The ratio of training data and testing data is 80:20. To visualize a relationship prior to training, a correlation heatmap and pair plot was generated and implemented. The heatmap and pair plot help to visualize the entire dataset and the total instances.



Figure A: Heatmap

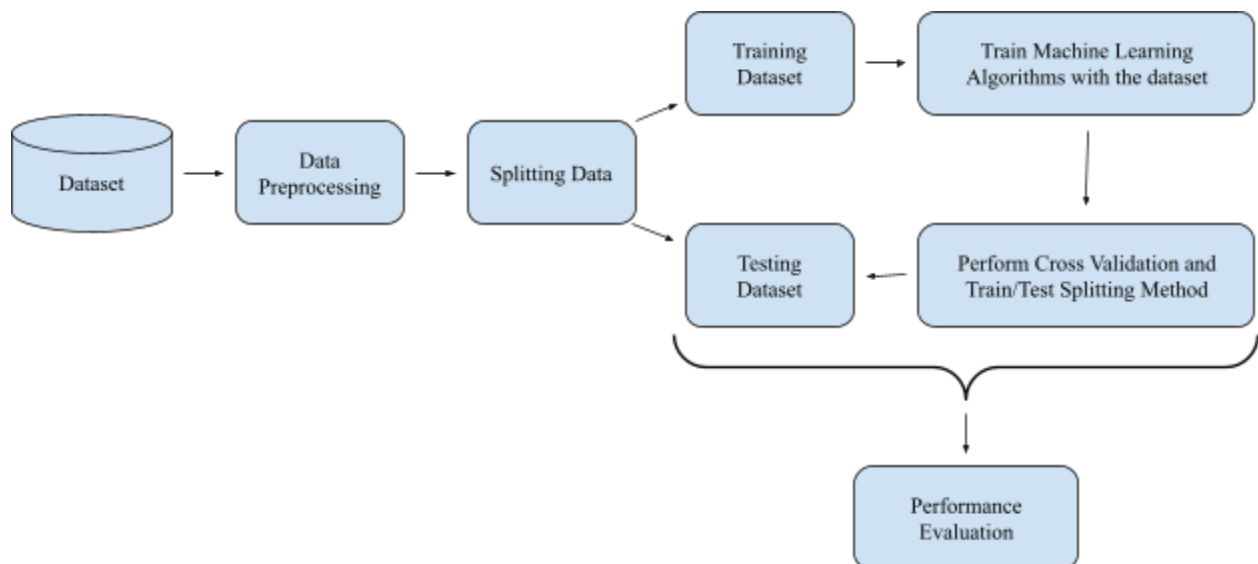


Figure B: Pairplot

Part D. Data Collection

Begin by importing all of the machine learning models into the Python notebook and create a function that runs each model through the training data and testing data. Create a confusion matrix based on the results of the testing data and calculate the accuracy, precision, recall, F1 score, and AUC score. Once calculated, log all of the results into a table and compare results between models.

Part E: Procedure Diagram



2.2 Variables

Independent Variable: The Machine Learning algorithms used in this experiment are the independent variables. The selected machine learning algorithms are Logistic Regression, Ridge Classifier, Ridge Classifier CV (Cross-Validation), Gaussian Process Classifier, Decision Tree Classifier, K Nearest Neighbors Classifier, Support Vector Classification, and Random Forest Classifier.

Dependent Variable: The performance metrics—accuracy, precision, recall, F1 score, and

AUC—used in our experiment that compares the performance of machine learning algorithms are the dependent variables.

Controlled Variables: The variables kept consistent throughout the experiment—dataset (training data and testing data), evaluation metric, model hyperparameters, and training procedure—are the controlled variables.

3. Results

The experiment results are obtained by splitting the data into training data and testing data. The most commonly used performance metrics such as accuracy, precision, recall, etc. are used to detect the confusion.

Accuracy is the best measurement that is useful in determining the best machine learning algorithms. The most accurate machine learning algorithms make the best decision. Precision refers to all the true positives divided by the total of true positive and false positives—calculating how precise and consistent the model is. The recall is called the measure of identifying the true positives correctly. Recalls help to identify how accurate the model is by using relevant data. F1 score combines the precision and recall score of a model to compute how many times a model made a correct prediction across the entire dataset. AUC represents the probability that the model is correct.

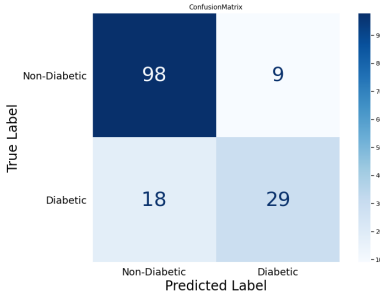
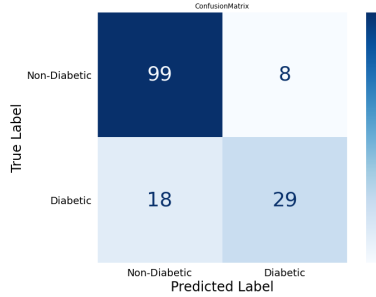
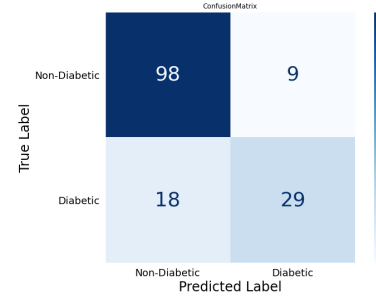
3.1 Data Tables

Model	Accuracy	Precision	Recall	F1 Score	AUC
RidgeClassifier	0.83117	0.78378	0.61702	0.69048	0.77113
LogisticRegression	0.82468	0.76316	0.61702	0.68235	0.76645

RidgeClassifierCV	0.82468	0.76316	0.61702	0.68235	0.76645
SupportVectorClassification	0.79221	0.7027	0.55319	0.61905	0.72519
DecisionTreeClassifier	0.78571	0.63462	0.70213	0.66667	0.76228
RandomForestClassifier	0.76623	0.7037	0.40426	0.51351	0.66474
KNeighborsClassifier	0.72078	0.53846	0.59574	0.56566	0.68572
GaussianProcessClassifier	0.61688	0.38889	0.44681	0.41584	0.5692

Figure 1: ML Models Measuring Metrics

3.2 Graphs + Models

Logistic Regression	Ridge Classifier	Ridge Classifier CV
 <p>ConfusionMatrix</p> <p>True Label</p> <p>Non-Diabetic</p> <p>Diabetic</p> <p>Non-Diabetic</p> <p>Diabetic</p> <p>Predicted Label</p>	 <p>ConfusionMatrix</p> <p>True Label</p> <p>Non-Diabetic</p> <p>Diabetic</p> <p>Non-Diabetic</p> <p>Diabetic</p> <p>Predicted Label</p>	 <p>ConfusionMatrix</p> <p>True Label</p> <p>Non-Diabetic</p> <p>Diabetic</p> <p>Non-Diabetic</p> <p>Diabetic</p> <p>Predicted Label</p>
Gaussian Process Classifier	Decision Tree Classifier	K Nearest Neighbor Classifier

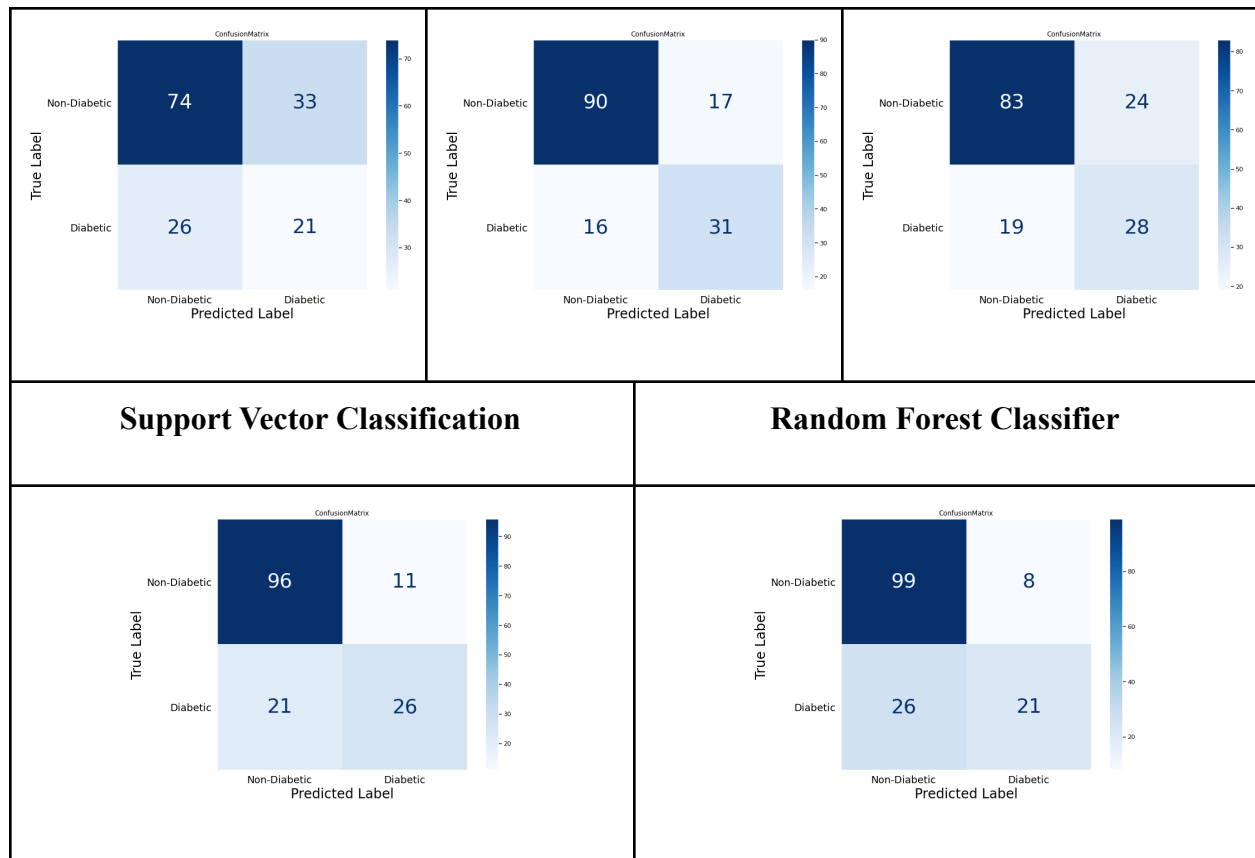


Figure 2: Confusion Matrix of Each ML Model

In Figure 2, the confusion matrix of each machine learning model is shown. The number of true positives, true negatives, false positives, and false negatives is stated in each quadrant. True negative and true positive are signs of accurate machine learning models.

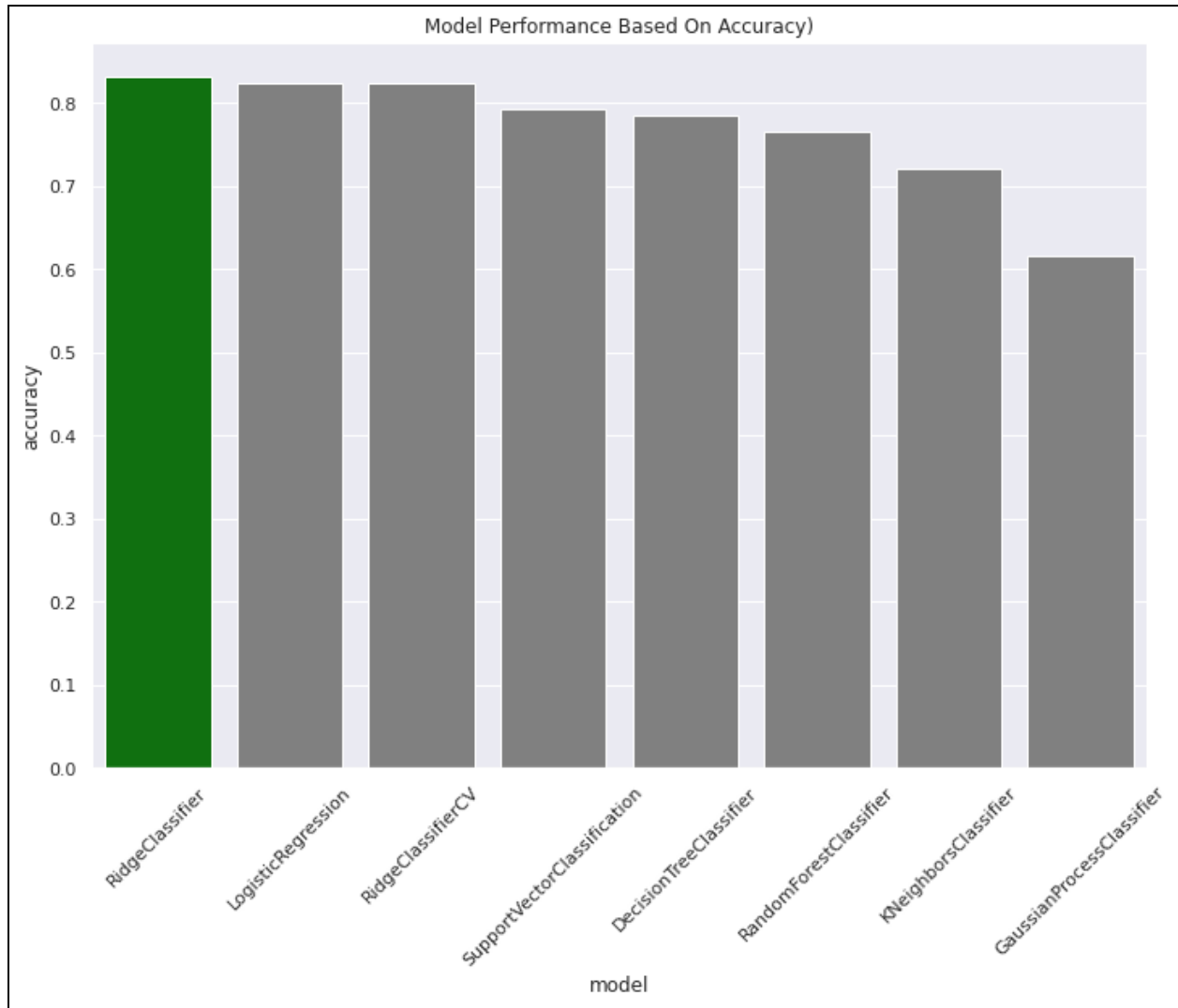


Figure 3: ML Model Performance Based on Accuracy

In Figure 5, the accuracy of each ML model is plotted using Python and SK-learn accuracy calculations. In the graph, the machine learning model “Ridge Classifier” has the greatest accuracy, suggesting that Ridge Classifier is the most accurate and deployable model for Pima Indians.

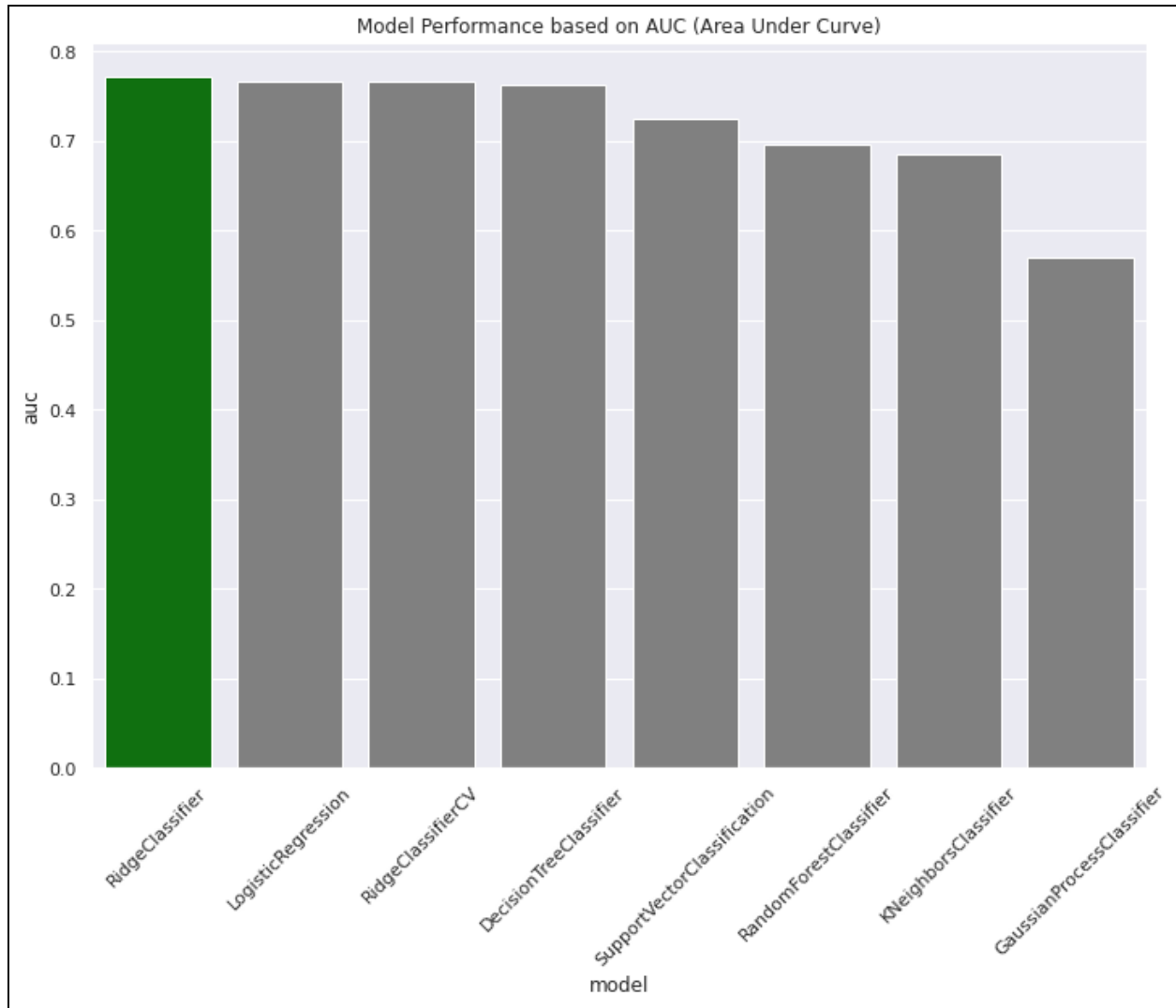


Figure 3: ML Model Performance Based on AUC

In Figure 5, the AUC score of each ML model is plotted using Python and area under the curve calculations. In the graph, the machine learning model “Ridge Classifier” has the greatest AUC score, suggesting that Ridge Classifier is correct the most often and possibly a deployable model for Pima Indians.

4. Discussions

The current study aimed to develop and validate machine learning models for diabetes diagnosis of Native Americans based on the most relevant information in determining diabetes derived from extensive research taken from female Pima Indians by the National Institute of Diabetes and Digestive and Kidney Diseases. For this aim, the Ridge Classifier, Logistic Regression, Ridge Classifier CV, Decision Tree Classifier, Support Vector Classification, Random Forest Classifier, K Nearest Neighbors Classifier, and Gaussian Process Classifier models were developed using a dataset taken from institutional research on female Pima Indians. The experimental results showed that the Ridge Classifier model had the best performance among the other seven machine learning techniques with an accuracy of 83.117%, precision of 78.378%, and ROC AUC of 77.113% (Figure 1). The results showed that logistic regression, ridge classifier CV, and decision tree classifier models also have a good prediction performance, as the ROC AUC for all is above 76%, and their diagnostic accuracy and efficiency is better than the remaining models trained using the same parameters (Figure 3).

Different studies have been evaluating the application of machine learning techniques in predicting diabetes and other diseases/sicknesses. In the case of COVID-19, Yadaw et. al. compared the performance of four machine learning algorithms—logistic regression, random forest, support vector machine, and extreme gradient boosting—for predicting COVID-19 mortality (2020). The researchers found that extreme gradient boosting happened to be the best model among all the models developed in terms of AUC with 91%. In another study concerning type 2 diabetes diagnosis, researchers found that the top-performing machine learning models were the decision tree and random forest classifier, with accuracy and ROS AUC of 99.0% (Fregoso-Aparicio et. al., 2021).

In the current study, some features such as glucose and insulin levels, BMI, age, and diabetes pedigree were of the highest importance; however, the number of pregnancies and skin thickness were of the lowest importance in predicting diabetes. However, from the perspective of medical facilities and physicians, awareness of this information may be significant in preventing diabetes through healthy methods. In machine learning techniques, not all of this information is important and can be ignored in analysis, allowing the model to predict diabetes with fewer factors. As a result, assessing diabetes can be a quicker process because the machine learning models do not need to account for insignificant information.

Several studies have also reported some important clinical features (predictors) for diabetes by performing research on triggers for diabetes and relationships among patient traits and diabetes. Clinical and biological equations state that body mass index (BMI), waist circumference, and baseline glucose level are signs of diabetes; in addition, for women, the triglycerides levels were another predictive measure of diabetes and possibly contribute to diabetes development (Balkau et. al., 2008). Parts of these predictive measures are already accounted for in the machine learning models; however, future data can be added, such as the waist circumference to the dataset, to increase the accuracy of the machine learning model.

The use of machine learning in the diagnosis of diabetes in Native American populations has the potential to greatly improve the efficiency and effectiveness of medical care in these communities. The Ridge Classifier algorithm, in particular, has demonstrated strong performance in terms of prediction accuracy, precision, recall, f1 score, and ROC AUC. By enabling the early identification of patients at risk of developing diabetes, this model can facilitate the optimal allocation of hospital resources and facilitate early prevention efforts, ultimately leading to improved health outcomes for Native Americans. Additionally, the use of this predictive machine

learning model can simplify the diagnosis process, enabling healthcare providers to more efficiently and effectively serve larger numbers of patients. Ultimately, the deployment of a valid and reliable predictive model such as the Ridge Classifier algorithm could significantly reduce the burden of diabetes in Native American communities, which suffer disproportionately from this disease.

5. Conclusion

Diabetes is a major health issue in Native American communities, with Native Americans having the highest rate of type 2 diabetes among all racial and ethnic groups in the United States and being nearly three times as likely to develop the disease compared to non-Native Americans. Access to healthcare and high rates of obesity are among the main contributing factors to the high prevalence of diabetes in Native American communities. Machine learning algorithms have the potential to improve the diagnosis and treatment of diabetes in these communities by analyzing patterns in electronic health records to identify patients at high risk of developing the disease.

In this study, machine learning techniques were employed to develop predictive models for diabetes diagnosis, evaluating their performance using a set of nine clinical features. Among the eight machine learning algorithms tested, the ridge classifier model demonstrated the highest classification accuracy, achieving a score of 83.12% and an AUC of 77.11%. The proposed model showed strong potential for predicting diabetes in female Pima Indians, and could serve as a valuable tool for expediting the diabetes diagnosis process in Native American communities. Additionally, the model has the potential to identify high-risk patients for diabetes at an early stage, enabling clinicians to initiate preventative treatment measures to mitigate the likelihood of the patient developing the disease.

As a result of the dataset taken from a specific tribe of Native Americans, a limitation occurs that is based on a limited dataset of electronic health records from Native American tribes. This may not be representative of the larger Native American population, and it is possible that the results of this study may not generalize to other Native American communities. Additionally, the use of machine learning algorithms to predict diabetes diagnosis is only one aspect of improving healthcare in Native American communities. Other factors, such as access to medical facilities and cultural barriers to healthcare, may also need to be addressed in order to effectively address the high prevalence of diabetes in these communities. Finally, the use of machine learning algorithms for predicting diabetes is still a relatively new field of research, and it is possible that further developments in this area may lead to the creation of more accurate and effective models. While the accuracy of the machine learning model in this study was relatively high, it is important to recognize that it may not be suitable for sole reliance in clinical practice. Instead, it may be advisable to consider complementary approaches or to further validate the model's performance in a real-world setting. In the future, it would be valuable to investigate the utility of other machine learning models, such as neural networks, for predicting diabetes diagnosis in order to determine the most accurate and deployable algorithm for use in medical settings. This could ultimately lead to improved healthcare outcomes for patients.

6. Acknowledgment of Major Assistance

Over the course of the research, which lasted from September 2022 to December 2022, I would like to express my sincere gratitude to my parent, Mursida Rahman, for their unwavering support and assistance throughout the research process for this paper. Their diligent monitoring and guidance helped ensure that the project was completed on time and to the best of my ability.

7. References

- Balkau, B., Lange, C., Fezeu, L., Tichet, J., de Lauzon-Guillain, B., Czernichow, S., Fumeron, F., Froguel, P., Vaxillaire, M., Cauchi, S., Ducimetiere, P., & Eschwege, E. (2008). Predicting Diabetes: Clinical, Biological, and Genetic Approaches: Data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). *Diabetes Care*, 31(10), 2056–2061. <https://doi.org/10.2337/dc08-0368>
- Centers for Disease Control and Prevention. (2019). *NHIS - Tables of Summary Health Statistics*. Centers for Disease Control and Prevention. <https://www.cdc.gov/nchs/nhis/SHS/tables.htm>
- Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome*, 13(1). <https://doi.org/10.1186/s13098-021-00767-9>
- Indian Health Service. (2016, October). *Special Diabetes Program for Indians | Fact Sheets*. Indian Health Service. <https://www.ihs.gov/newsroom/factsheets/diabetes/>
- Indian Health Service. (2019, October). *Disparities | Fact Sheets*. Indian Health Service. <https://www.ihs.gov/newsroom/factsheets/disparities/>
- Sandefur, G. D., Rindfuss, R. R., Cohen, B., & National Research Council (U.S.). Committee On Population. (1996). *Changing numbers, changing needs : American Indian demography and public health* (Vol. 12). National Academy Press.
- Schure, M., Goins, R. T., Jones, J., Winchester, B., & Bradley, V. (2019). Dietary Beliefs and Management of Older American Indians With Type 2 Diabetes. *Journal of Nutrition Education and Behavior*, 51(7), 826–833. <https://doi.org/10.1016/j.jneb.2018.11.007>

- Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265.
- Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 24(1), 12–18. <https://doi.org/10.11613/bm.2014.003>
- Yadaw, A. S., Li, Y., Bose, S., Iyengar, R., Bunyavanich, S., & Pandey, G. (2020). Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *The Lancet Digital Health*, 2(10), e516–e525. [https://doi.org/10.1016/s2589-7500\(20\)30217-x](https://doi.org/10.1016/s2589-7500(20)30217-x)