ML4VA: Machine Learning for Virginia

Manish Balamurugan (mb2mcc), Aneesh Vittal (akv6zr), Alip Arslan (aa8pss)

December 08 2022

1 Abstract

Virginia's diabetes rate is on par with the national average and growing quickly. In recent years, we've also seen a rise in the proportion of the population that can be classified as prediabetic. However, Type 2 diabetes is preventable with lifestyle changes and proper treatment. For this project we have decided to tackle the challenge of determining whether an individual is at risk of diabetes in order to ensure that they can undergo the proper screening, treatment, and life style changes in order to properly control this disease before serious complications arise. In this project we employ various classification models and experiment with optimizers and error functions to achieve the highest precision model. We believe that identifying an individuals likelihood for contracting diabetes may encourage Virginians to consider lifestyle changes and reduce our state's diabetic rate in the future.

2 Introduction

Millions of people across the United States face the consequences of dealing with diabetes on a daily basis. In the U.S., 11% of the adult population has diagnosed diabetes while a staggering 38% of adults have diabetes, but have not been diagnosed ([4]). In Virginia, the statistics aren't much different. Over 10% of adults have diagnosed diabetes while 33% are undiagnosed [1]. Diabetes creates risk of long term complications and is financially burdensome. In 2017, diabetes related medical expenses were estimated to be over $6 billion. More than just being statistics, these numbers show the scale of diabetes' impacts and the way it affects people's quality of life. Using Machine Learning, we could create a model to classify patients that are at risk for diabetes in an effort to promote early treatment. Prior research on this topic, such as the *Classification and prediction of diabetes disease using machine learning paradigm*[2] study, performed classification using a variety of models. The dataset used for the study only contained responses that were diabetic or not diabetic. In our study, we aim to introduce data on prediabetic patients as well to analyze performance. Additionally, the study entitled *An Ensemble Approach to Predict*

Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study[3] performed ensemble learning to make a similar prediction. However, the dataset used in this study concerned other diseases and conditions rather than viewing it from a lifestyle perspective. These are the approaches we aimed to test in this experiment.

3 Method

We wanted to explore a variety of different models including Linear Regression, Logistic Regression, Decision Trees, Random Forest Trees, Bayesian Learning, and Neural Networks. By analyzing metrics across these models, we will be able to come to a conclusion on the efficacy of these machine learning models to solve our problem as defined.

Since our dataset had numerical values indicating whether or not the patient had diabetes, we tried regression as well as classification approaches. In order to convert the multiclass problem posed by the dataset into a binary classification problem, we converted all prediabetic labels to diabetic labels. Seeing as our goal is to recommend diabtes screenings to patients who would be considered either prediabetic or diabetic, we were able to combine these labels. To aim for better performance, we wanted to use the features that were correlated to the label. We dropped features that were negatively correlated and proceeded to split the dataset. The dataset was split according to a 70:30 ratio of training data to testing data. At this point, we also separated the labels from the dataset in order to train our model. Upon examination, we noticed that the data had already been cleaned and had no missing values. All features of the dataset were numerical as well. Since we didn't have to encode any categorical values, we proceeded by passing the data through a Standard Scaler to scale features to unit variance. Once the data was scaled, we passed it into the models.

4 Experiments

For this experiment, each model is implemented and hyperparameter tuned utilizing GridSearchCV. This refers to the process of performing hyperparameter tuning in order to determine the optimal values for a given model by generating all hyperparameter combinations and evaluating each model using the Cross-Validation method. Across our models we utilized a confusion matrix to determine each model's architecture performance. For certain models - such as linear and logistic regression - there was not a significant impact when performing hypertuning while for others the converse can be said such as the Random Forest model. For figures such as epochs we made our training across all the models (we utilized 10 epochs for each model). For most our models we used the pretained models from the Sci-kit learning machine learning library for python.

For our CNN implementation we created a custom CNN with its architecture shown below:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 16)                304

 dense_1 (Dense)             (None, 16)                272

 dense_2 (Dense)             (None, 1)                 17

=================================================================
Total params: 593
Trainable params: 593
Non-trainable params: 0
```

For the first two layers we used a RELU activation function and for the last layer we used a sigmoid function.

5 Results

As mentioned earlier, we utilized 6 different models in order to ensure that our approach is a viable solution. Each model was hypertuned. We generated plots of both training and validation loss and accuracy in order to further analyze and extract any meaningful insights from our models - we then reported each models performance utilizing a confusion matrix. Across our models we saw an accuracy of approximately 82% with our better performing models performing 84% and better.

Of models we performed this trial with, three models stood out in particular:

1. Logistic Regression
   (a) Precision: 0.8168839533921777
   (b) Recall: 0.8491603595080416
   (c) Accuracy: 0.8491603595080416
2. Random Forest
   (a) Precision: 0.8232207675424282
   (b) Recall: 0.8523927783033743
   (c) Accuracy: 0.8523927783033743

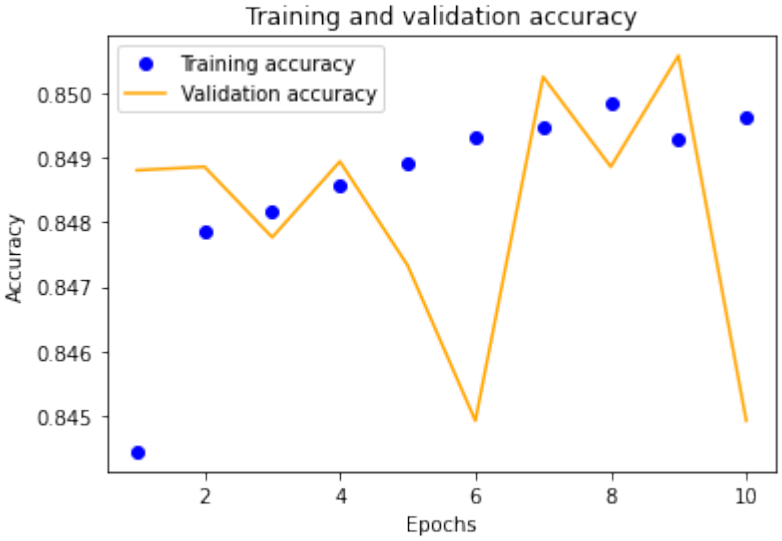
3. Convolutional Neural Networks

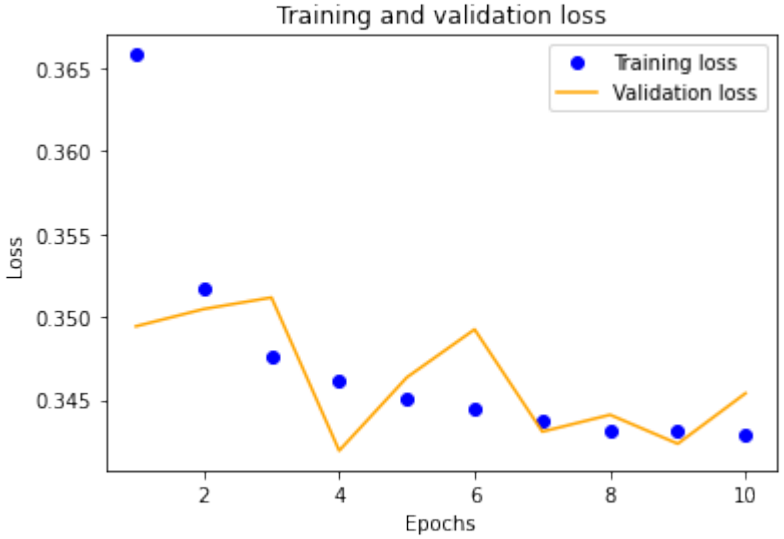
   (a) Precision: 0.8244024510880571

   (b) Recall: 0.8472485020498266

   (c) Accuracy: 0.8505885601043701

5.1





We used a confusion matrix to report the performance of the models used in this project and also generated our models training/validation loss and accuracy. The charts above correspond as a visual of the loss and accuracy generated for the CNN model that was employed.

6 Conclusion

From analyzing the results across our models, we see a consistent accuracy of around 84% or higher. This shows that we can effectively use Machine Learning models to help patients in Virginia get tested for diabetes early on in order to combat long term complications. Considering that prior research viewed this problem from the perspective of relating other illnesses to diabetes, our results show that one's lifestyle can be used to screen for diabetes as well. However, it is important to note that lifestyle choices may not be the best predictor of diabetes. In future research, perhaps a study combining lifestyle data with the patient's medical history may find better results. Additionally, seeing as the dataset was sourced from responses to a telephone survey, there may be better results obtained from a source with less potential for bias as some respondents may not accurately describe their lifestyles.

References

[1] *The Burden of Diabetes in Virginia.* https://diabetes.org/sites/default/files/2021-11/ADV$_2$021$_S tate_F act_s heets_V irginia_r ev.pdf$.

[2] *Classification and prediction of diabetes disease using machine learning paradigm.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6942113/.

[3] *An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9324493/.

[4] *National Diabetes Statistics Report.* https://www.cdc.gov/diabetes/data/statistics-report/index.html.

7 Contributions

Aneesh: Data cleaning, setting up models, tuning models, video presentation, final report

Manish: Data cleaning, setting up models, video presentation, final report.

Alip: Tuning models, setting up models, video presentation, final report