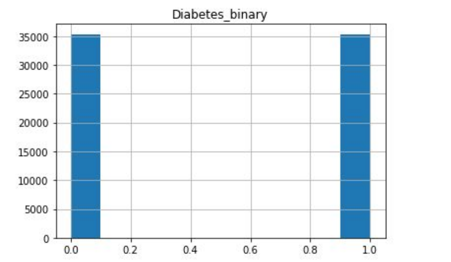
Datathon 2021 - Diabetes Prediction Model

By: Ismail, Amna, Mark

Dataset



The dataset used in our project utilizes results from the 2015 Behavioral Risk Factor Surveillance System, a telephone survey conducted by the CDC annually. The survey collects information from Americans regarding risk behaviors, chronic health conditions, and healthy habits, practises, and preventative measures. The modified datasets are a consolidated version of the original BRFSS 2015 dataset available on Kaggle, which contains responses from 441,455 individuals and 330 features; features being direct survey questions or calculated variables based on responses.

The specific dataset we used is diabetes _ binary _ 5050split _ health _ indicators _ BRFSS2015.csv, a clean dataset using 70,692 responses to the CDC's 2015 survey. It has 21 features and is balanced. It has a 50-50 split in respondents being classified as having either no diabetes or having prediabetes or diabetes. The variable Diabetes_binary has 2 classes, 0 for no diabetes, and 1 for prediabetes or diabetes.

The Research Question

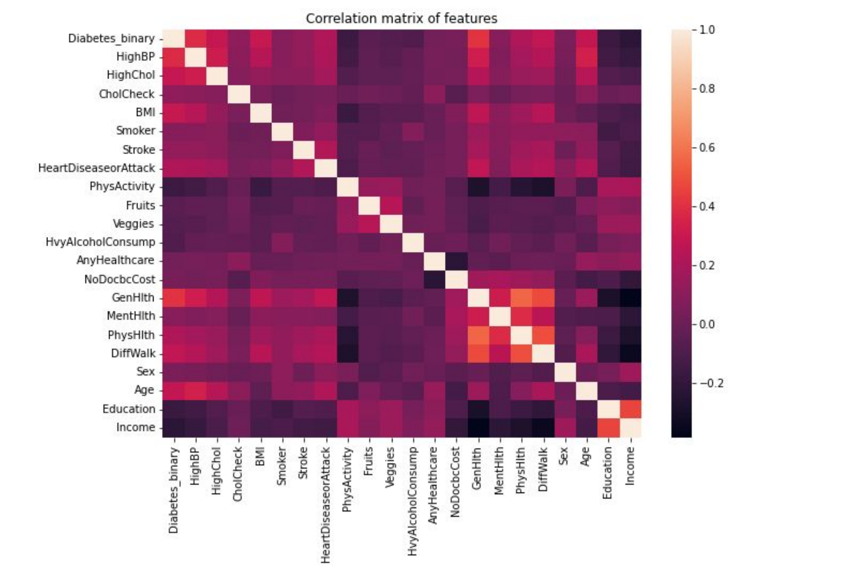
Based on existing information regarding an individual’s health conditions and lifestyle habits, can we accurately diagnose whether they have diabetes or pre-diabetes? And with what level of accuracy?

Predictors

The following columns from the dataset were used as predictors:

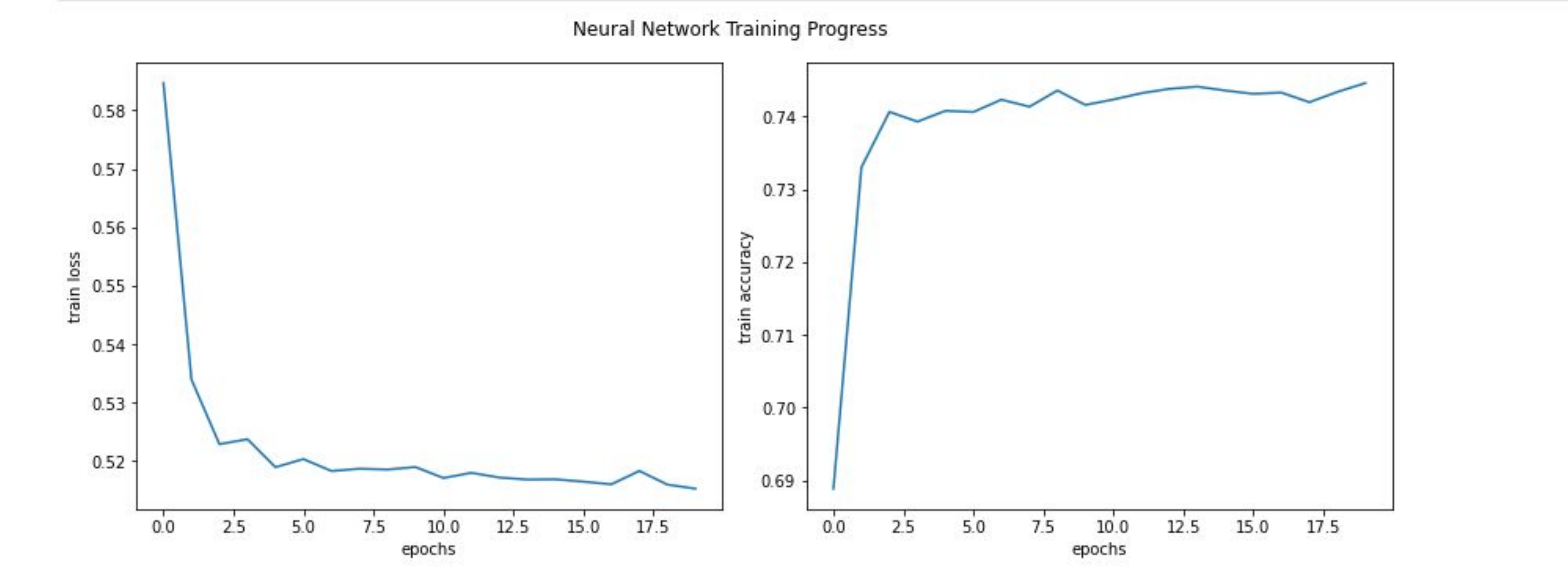
- HighBP
- HighChol
- BMI
- HeartDiseaseorAttack
- GenHlth
- PhysHlth
- DiffWalk
- Age

To select predictors, we generated a heat map with a correlation matrix of predictors and selected columns with a correlation coefficient of 0.25 or higher.



Model Training

- Accuracy increases as we pass through epochs
- Total of 20 epochs used
- SVM: 74.6% accuracy
- Decision Tree: 71.0% accuracy
- Random Forest: 71.0% accuracy
- Neural Network: 74.4% accuracy



Impact

The international Diabetes Federation has listed diabetes as one of the leading global health crises in the 21st century. In 2019, 1 in 3 Canadians were found to have either diabetes or prediabetes. While more research is being done on preventative measures for Type 1 diabetes, the onset of Type 2 diabetes can be delayed or prevented through a number of means, including healthy behavior interventions, such as physical activity, certain dietary patterns, and weight loss. Models such as ours can help individuals gauge how much risk they have for developing diabetes, or help gauge for themselves whether they need to seek treatment or intervention.

Conclusion

- Support Vector Machine performed best at predicting diabetes
- Small difference between Neural Network - 0.2%
- SVM - much longer runtime
- Decision Tree and Random Forest - very fast

To consider:

- Criteria for selecting columns - correlation
- Add/remove columns to increase accuracy? HeartDiseaseorAttack, PhysActivity, Income
- Real world applications - medical, insurance

THANK YOU!