Anastasiia
Ganshina

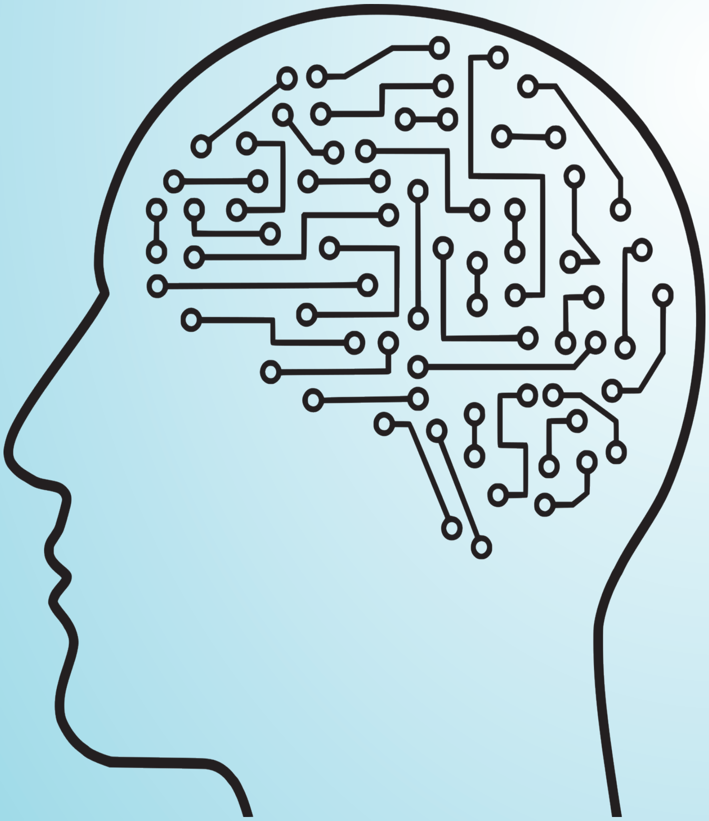
DIABETES DIAGNOSIS MACHINE LEARNING ALGORITHM

Around the world, and especially in the USA, the number of people with diabetes increases linearly over the years. This disability, unless diagnosed early and treated properly, leads to many complications, such as heart disease, nerve damage, vision loss, and other. Being diagnosed with the disability on time, lets a diabetic person prevent such complications and live a long and full life. In this publication, I will walk you through the process of development of this project as well as its end result: an algorithm that can predict whether person has diabetes or not.

TABLE OF CONTENTS

PROJECT PROPOSAL

While people cannot diagnose themselves with diabetes, modern algorithms have the ability to do so. The product proposed is an algorithm that takes some information about a person as an input, and as predicts whether the person has diabetes based on that information.



Besides the baseline algorithm explained above, the project might develop in the cluster of algorithms that will analyse the provided dataset and give a user more information about their diabetes status. Examples include: the change of sugar levels, insulin dosage, foods that spike the sugar levels, etc. Moreover, with enough time and effort, a project may pivot to an application that will track the person's diabetes. It may include, sugar tracking, sugar prediction, diabetes status.

Being prediabetic and having a family member dying from type 1 diabetes, make me very passionate about the disease and the project. With this micro-internship opportunity, I will be able to learn enough skills and make such an algorithm possible.

DATA EXPLORATION

In order for the algorithm to accurately predict if a person has diabetes, enough data regarding symptoms should be passed to the training set. The desired dataset had to include as many distinct unrelated data points as possible.

Before using the dataset in the algorithm, all duplicates in the set were removed and the order of the data points was randomized. This was done in order to ensure that the algorithm did not have the same training and testing data or any relations between the features (example: sorted by age) that could potentially corrupt the program.

Chosen symptoms are presented in the column to the right. More information about each of the features and their relevance can be found in the next pages.

Age

Gender

Polyuria

Polydipsia

Sudden weight loss

Weakness

Polyphagia

Genital thrush

Visual blurring

Itching

Delayed healing

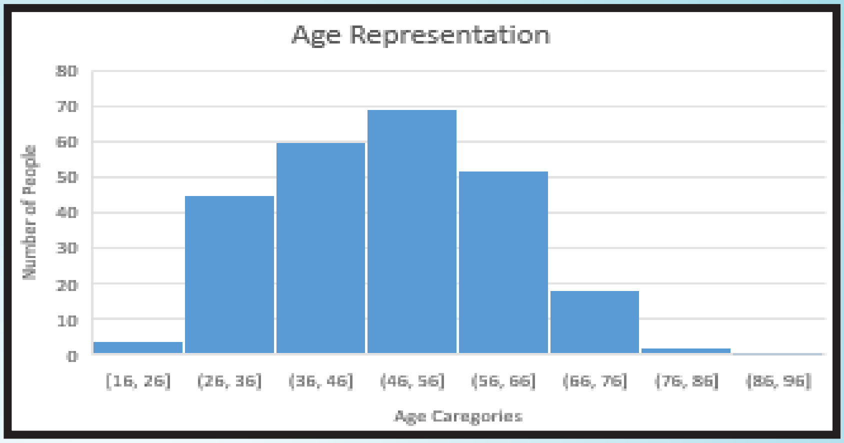
Partial paresis

Muscle stiffness

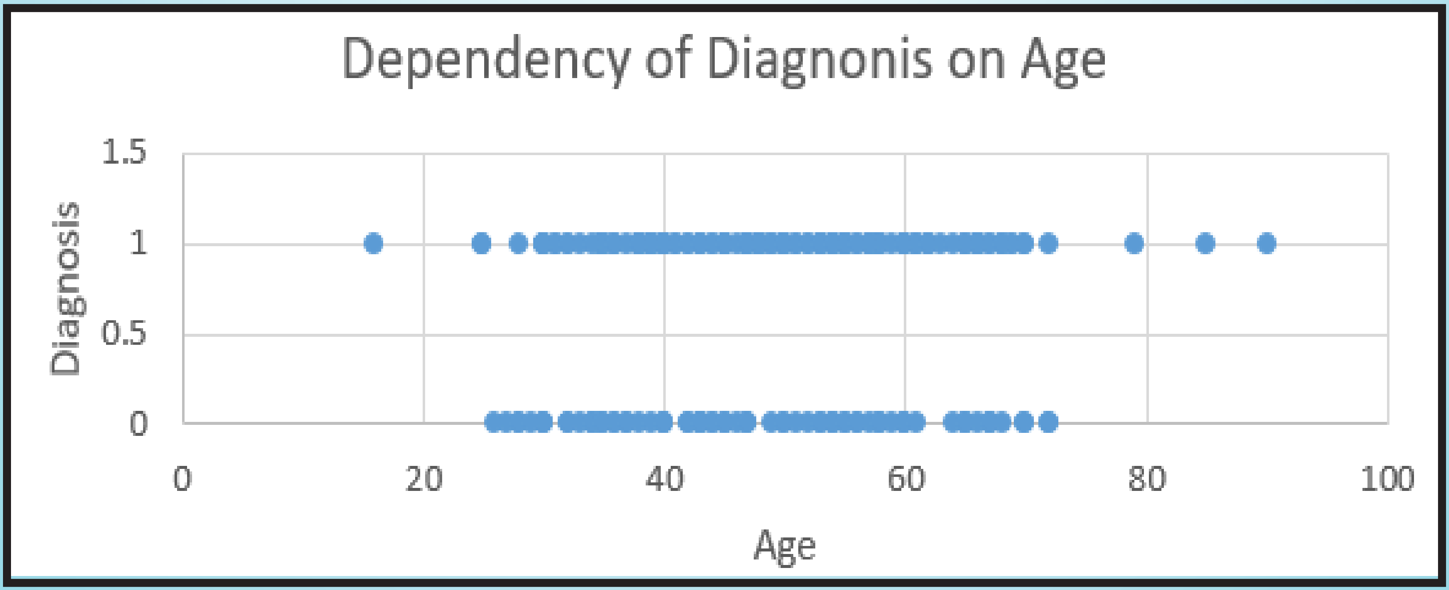
Alopecia

Obesity

AGE

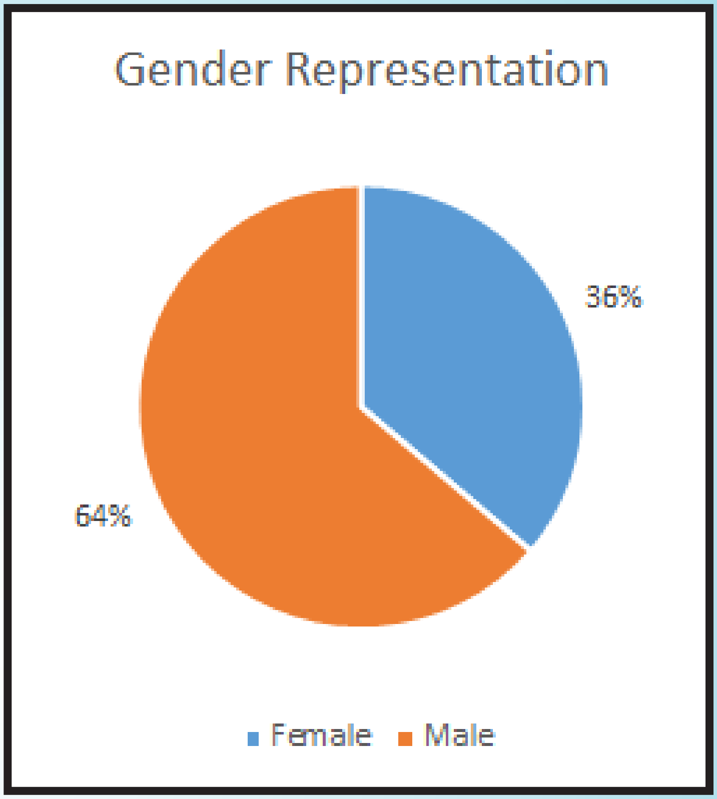


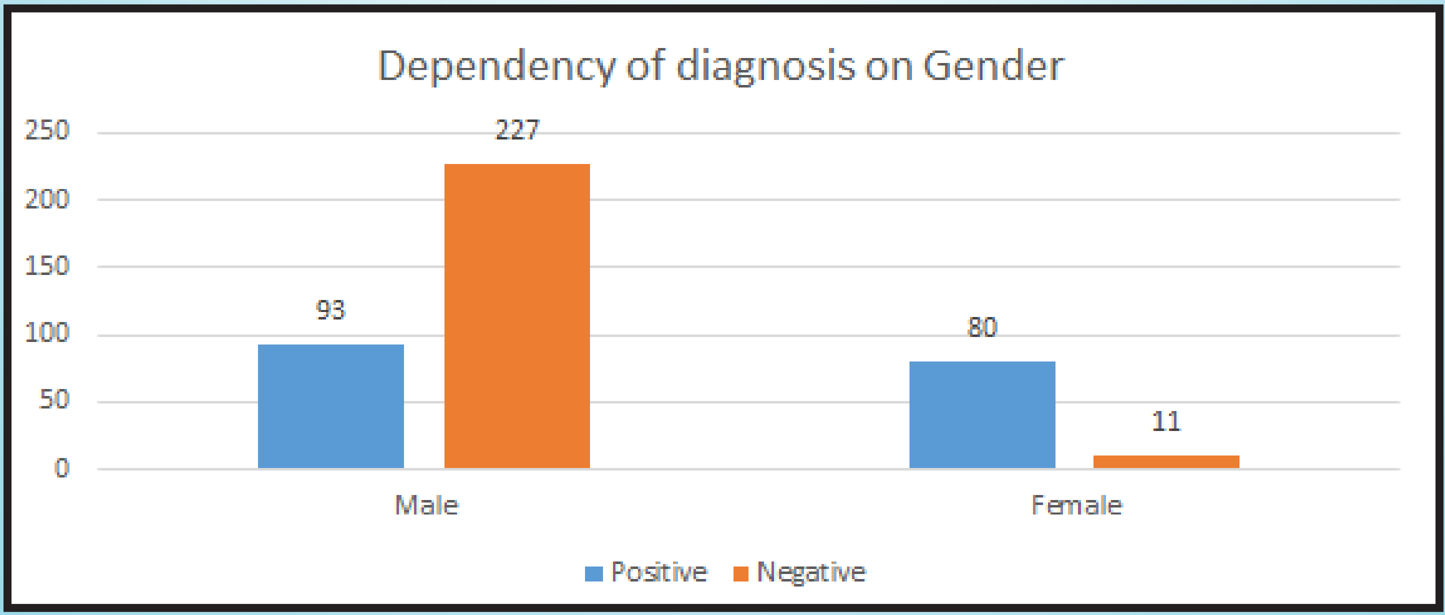
According to the Center for Disease Control (CDC), type 1 diabetes is mostly prevalent in young adults, teens and children of ages 4 to 14. While other cases are possible, they are rare. Type 2 diabetes, on the other hand, is diagnosed among adults over the age of 45. As we can see in the analyzed dataset, the majority of the positive cases layers on the category of ages from 45 to 65 which means that age is relevant in this algorithm. Another characteristic that should be noted is that the dataset does not provide data from younger people which leads to a conclusion that the majority of cases in this study is type 2 diabetes.



GENDER

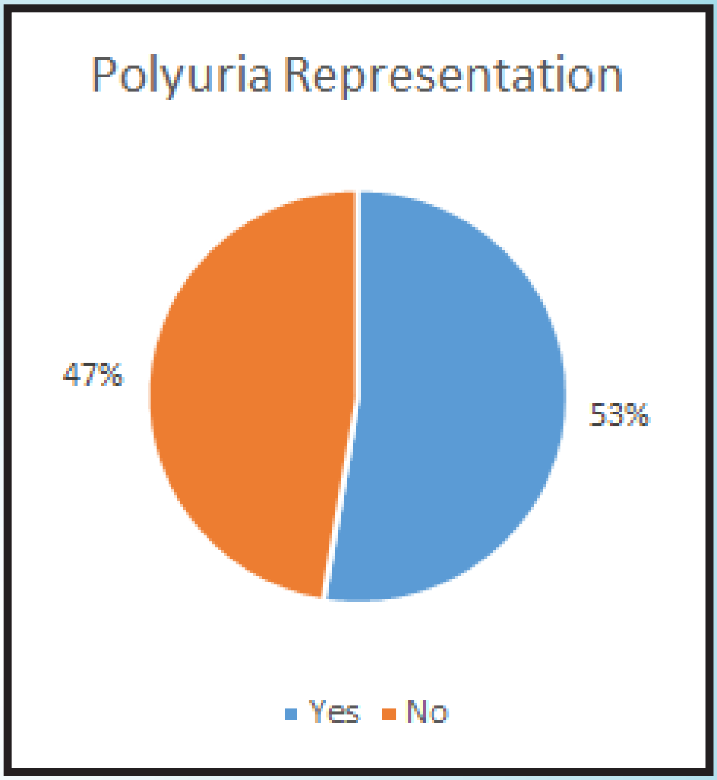
According to NIH, there is almost no bias in diabetes diagnosis on patient's sex. In our dataset, however, we can clearly see the bias of women being more likely to be diabetic. As a result, this feature will not be included in the Machine Learning Algorithm to eliminate its bias.

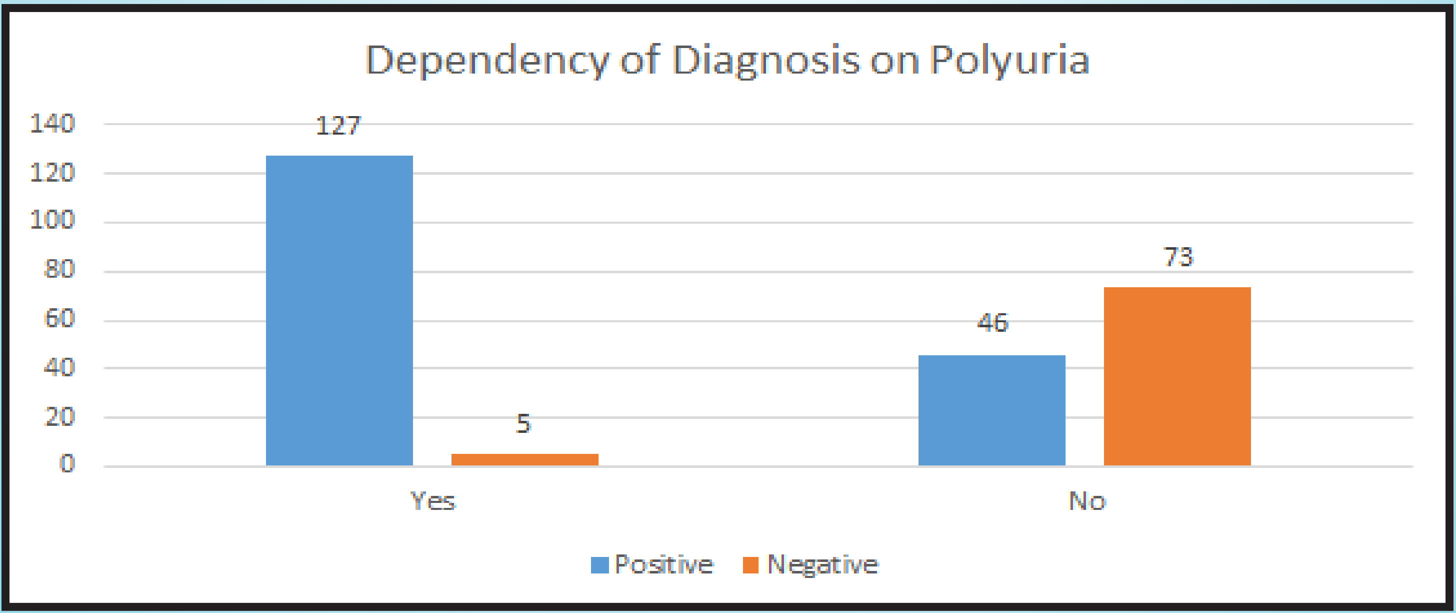




POLYURIA

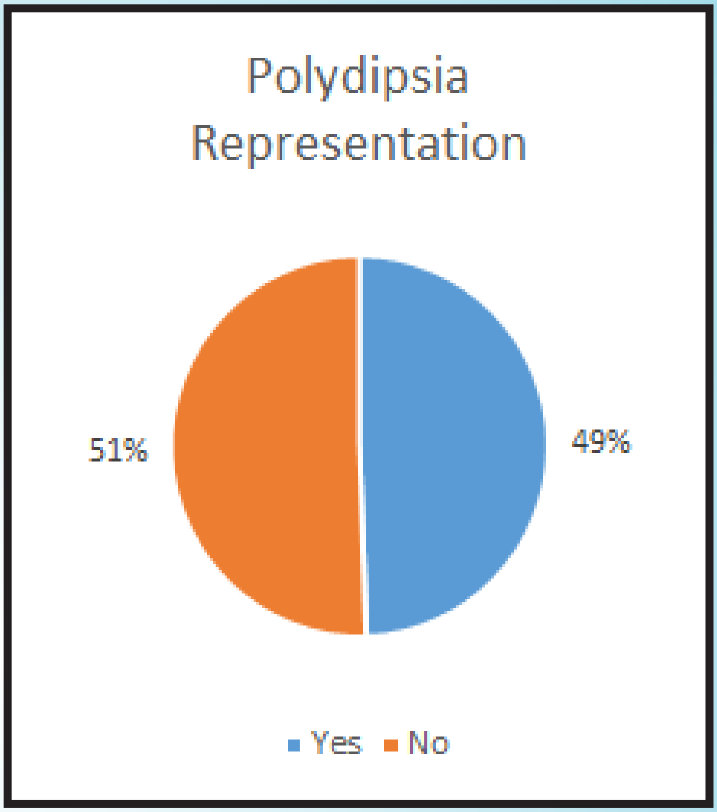
Polyria is a condition in which a person produces large volumes of dilute urine. According to the CDC, production of large volumes of urine is a symptom of diabetes. As a result, this feature is relevant to the study. In the provided dataset, 53 % of people in the study have polyuria, and the majority of them tested positive for diabetes. This means that this data will improve the accuracy of the algorithm.

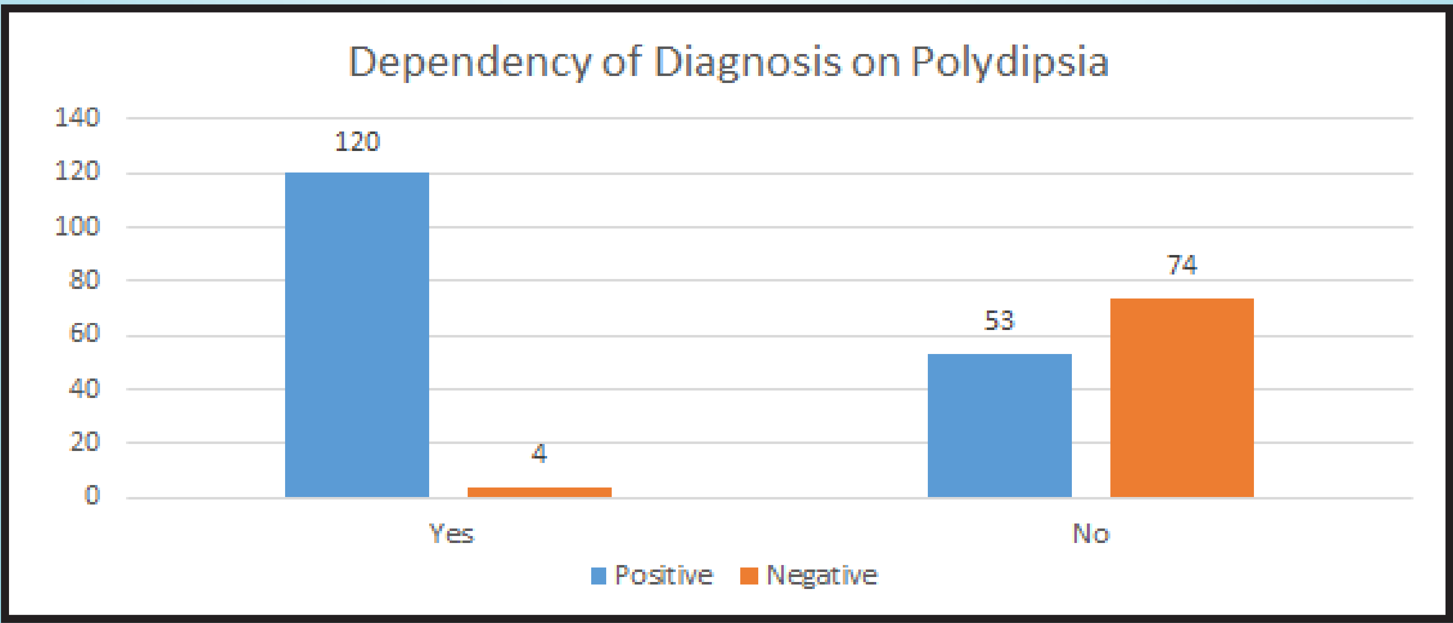




POLYDIPSIA

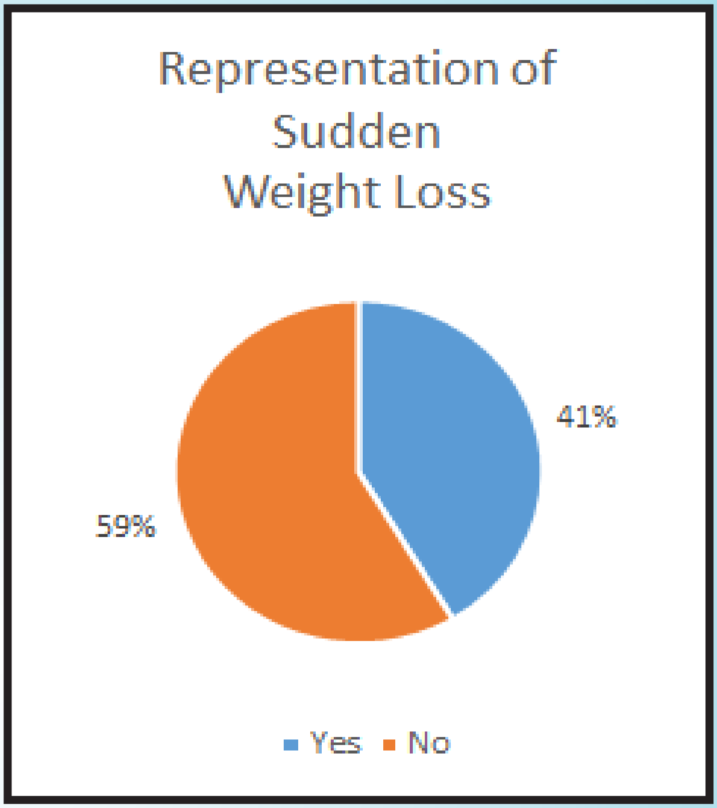
Polydipsia is defined as a great thrust due to an underlying disease: diabetes, in our case. According to CDC, thrust is a symptom of diabetes that means that it is relevant to the research. In this data set, 49% of people experience polydipsia and most of those people end up diagnosed with diabetes. As a result, this feature in this dataset will increase the accuracy of the algorithm.

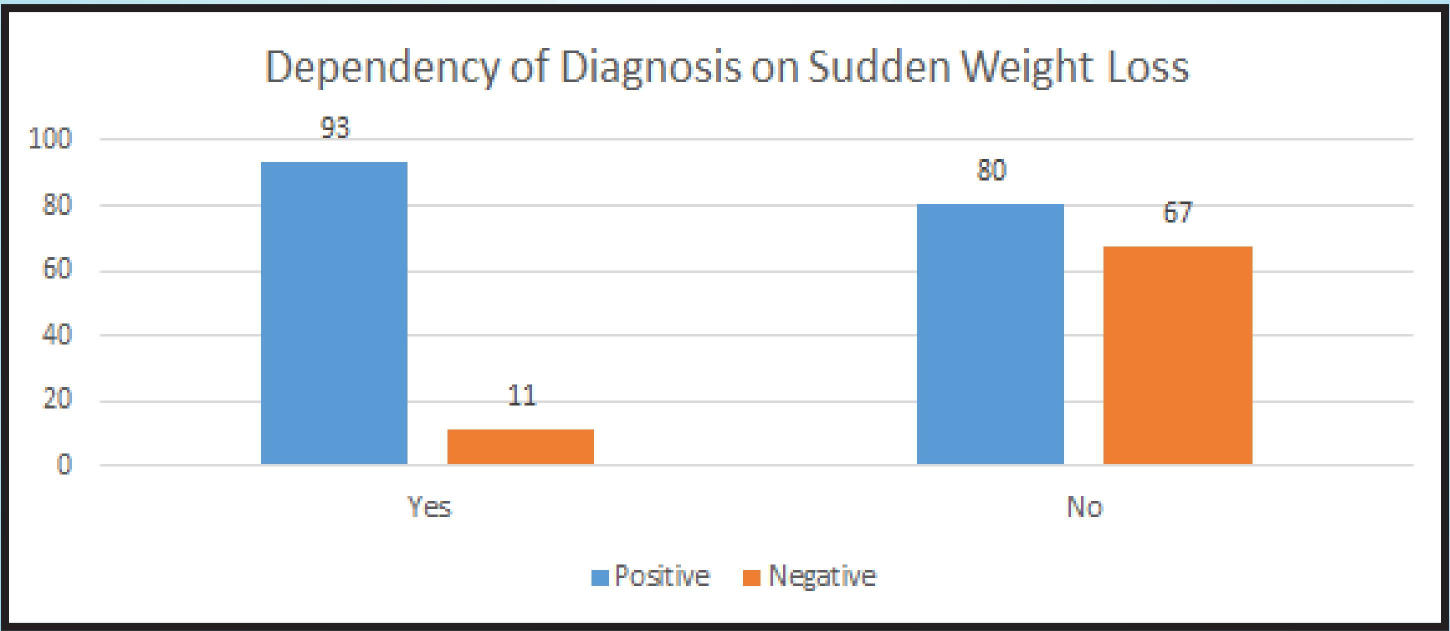




SUDDEN WEIGHT LOSS

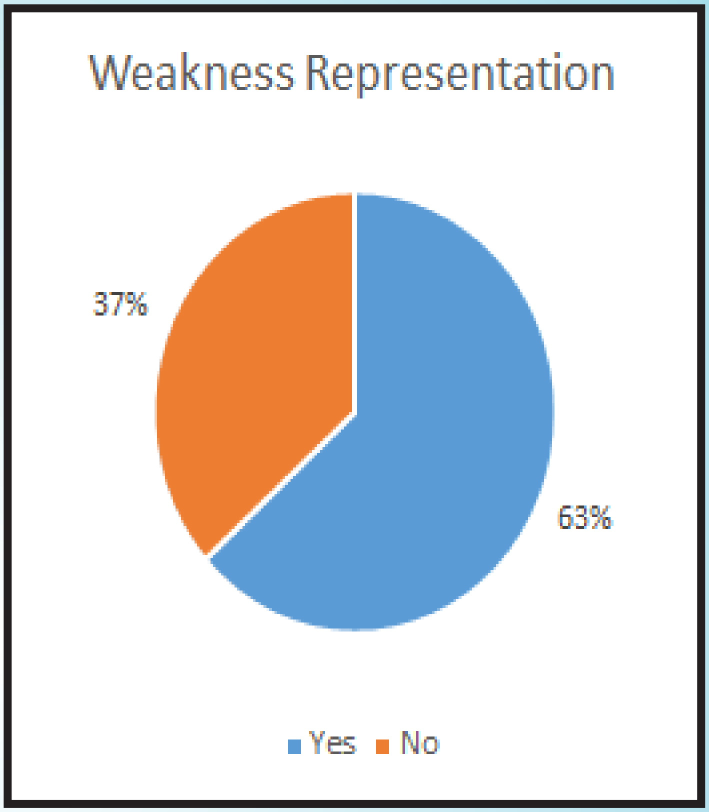
According to the Center for Disease Control, sudden weight loss is a symptom of both types of diabetes and prediabetes which means that this feature is relevant to the study. Provided data set includes 41 % of people who claim to have unintended weight loss and the majority of them have diabetes. This findings lead to the conclusion that the effect of this data on the algorithm will be positive.

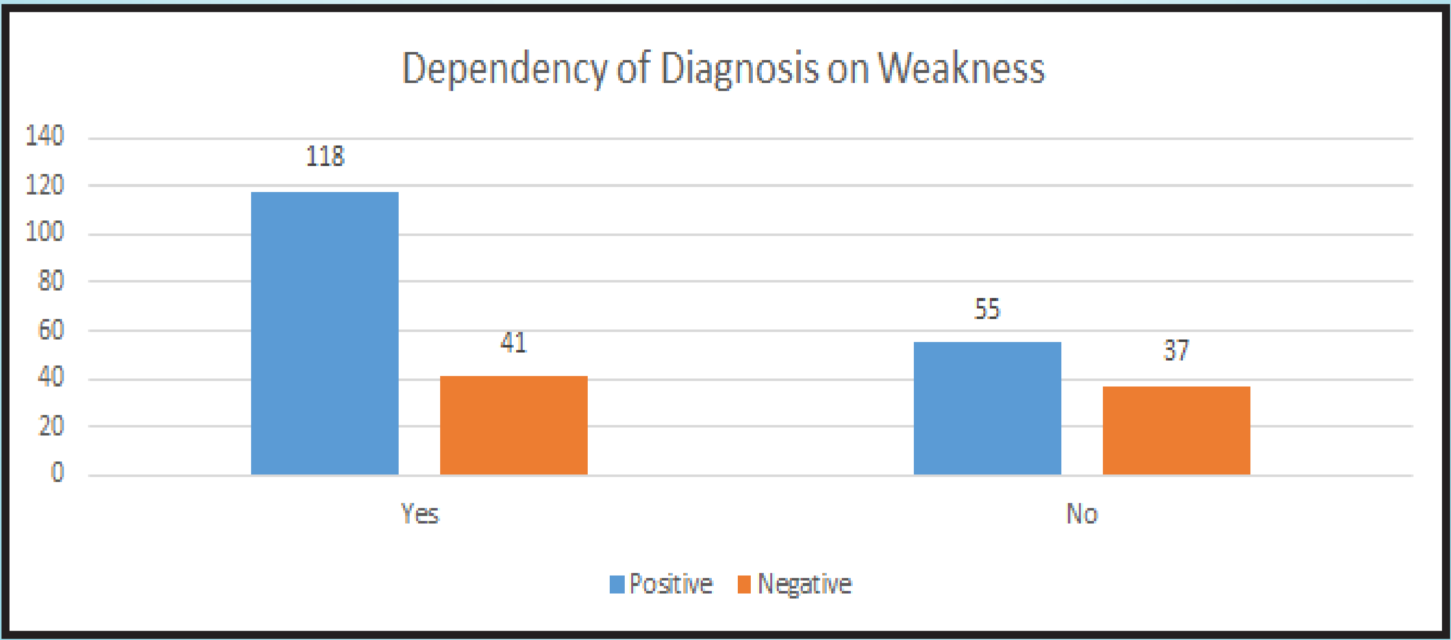




WEAKNESS

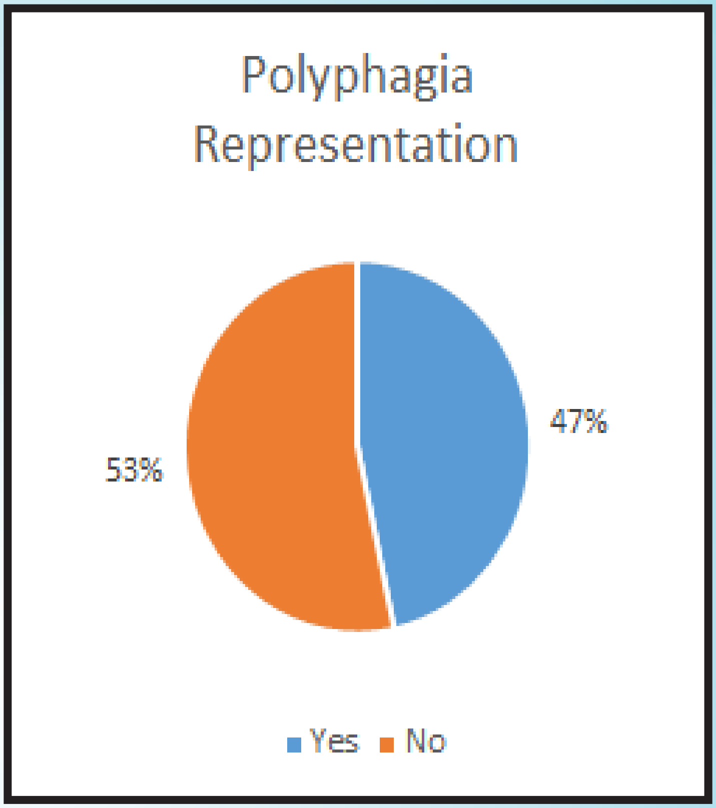
According to CDC, weakness is a symptom of diabetes that means that it is relevant to the research. In this data set, 63% of people experience weakness and most of those people end up diagnosed with diabetes. As a result, this feature in this dataset will increase the accuracy of the algorithm.

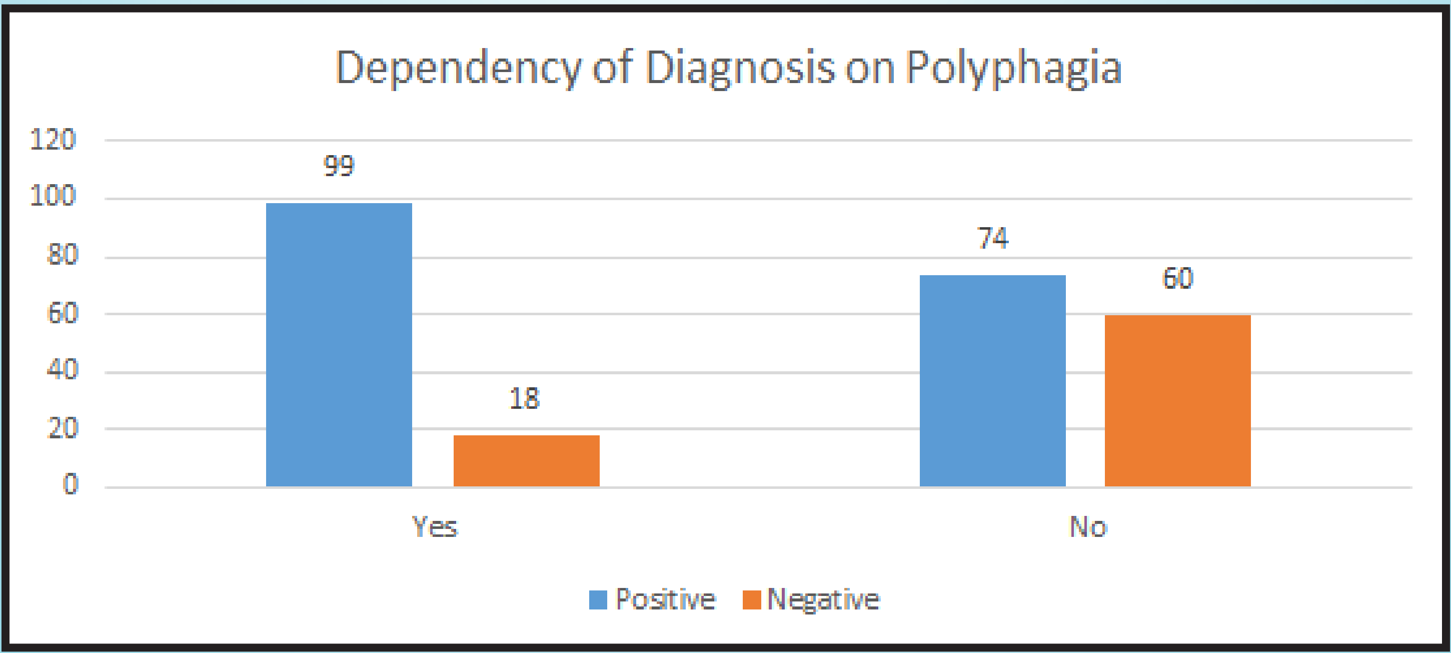




POLYPHAGIA

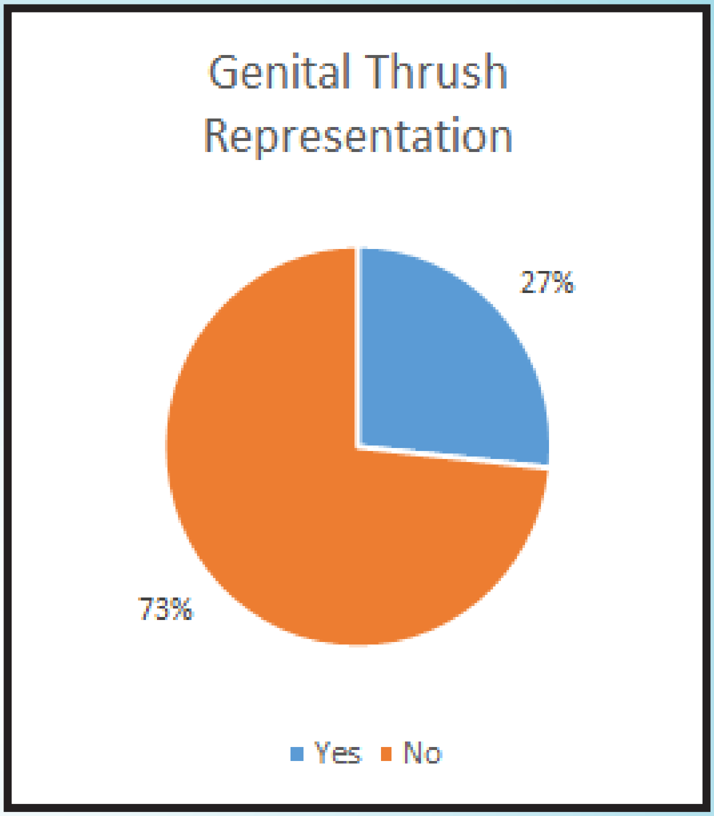
Polyphagia or, in other words, extreme hunger, is a symptom of diabetes and is relevant to this research project. In this dataset, 47% of people in the study claimed to have extreme hunger and 85% of them have diabetes. As a result the feature is helpful to this algorithm.

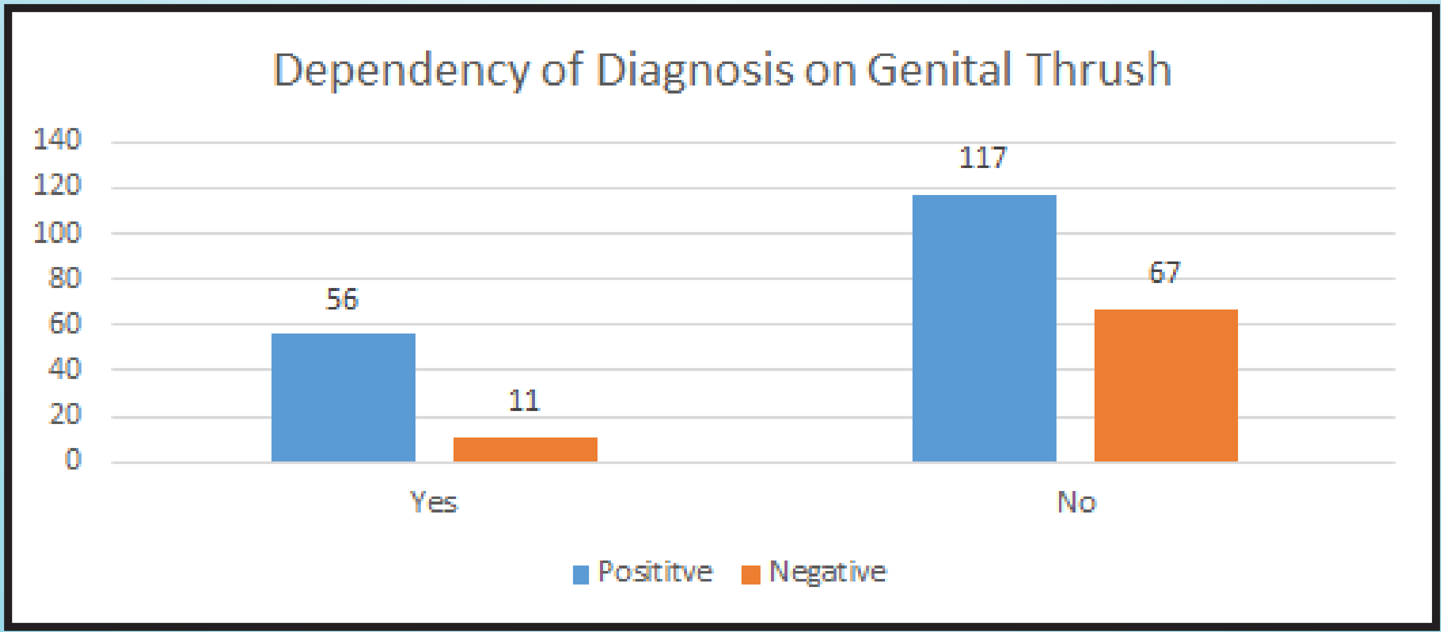




GENITAL THRUSH

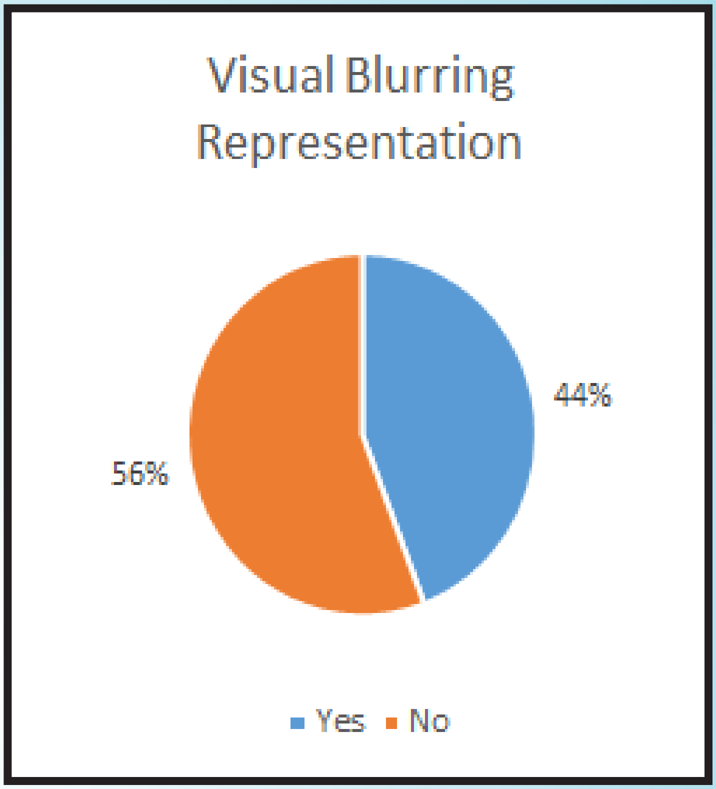
NIH provides information that infections, such as, vulvovaginal candidiasis occur more often in diabetics. Although it is not a symptom of diabetes, this condition is a consequence of it. As a result, this information may be useful in this study. The vast majority of people in the dataset did not experience genital thrush, but from those who did, the majority was diagnosed with diabetes. As a result, this information will be beneficial for the algorithm.

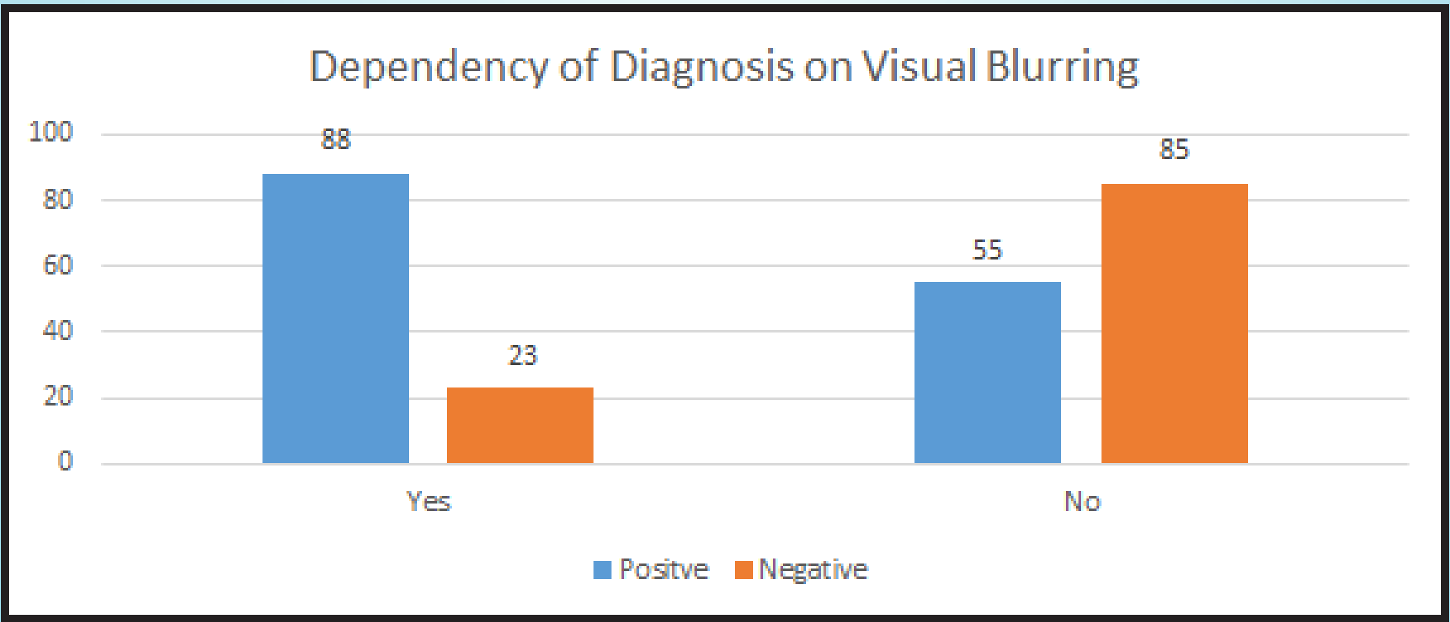




BLURRED VISSION

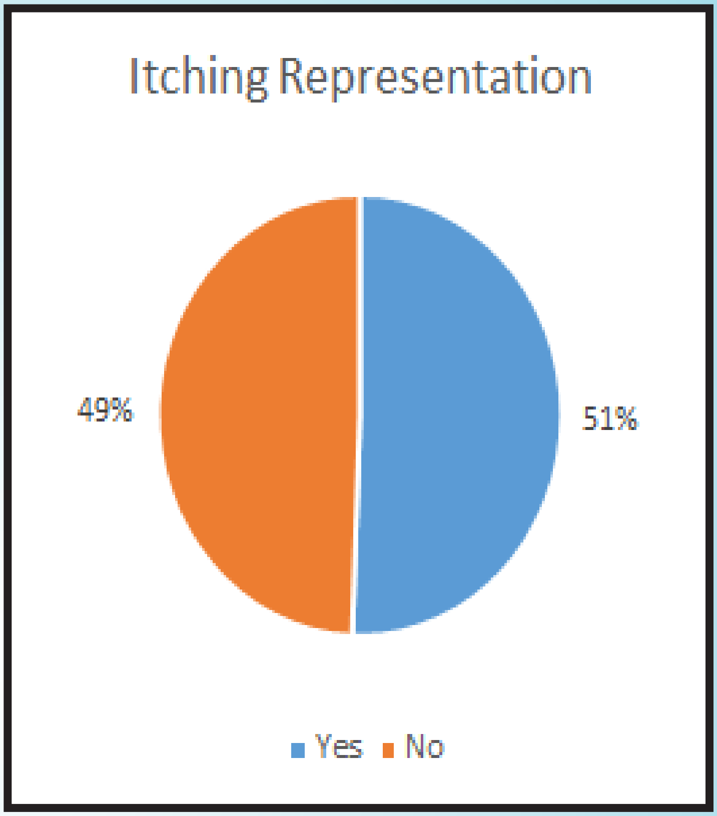
CDC claims that visual blurring is a consequence of diabetes which means that the feature is relevant to the study. In the dataset, 44% of people experience visual blurring and the majority of them are diagnosed with diabetes. As a result, this information will increase the accuracy of the algorithm.

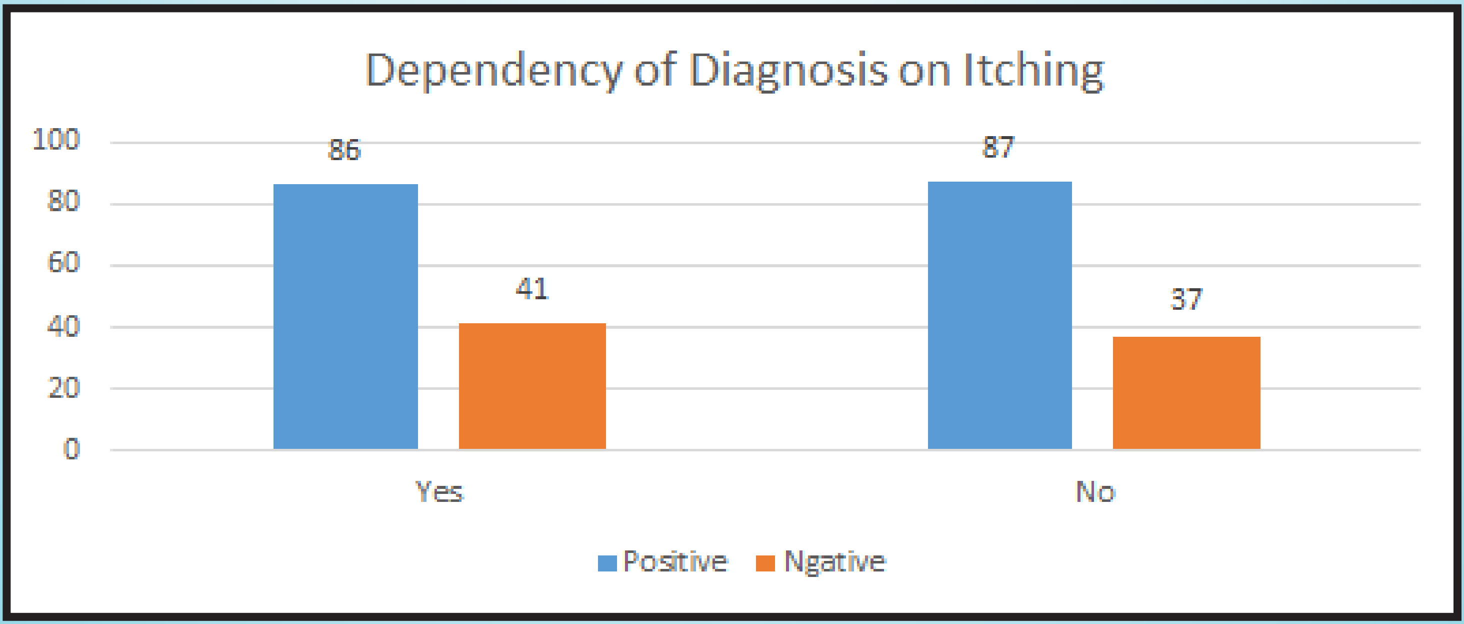




ITCHING

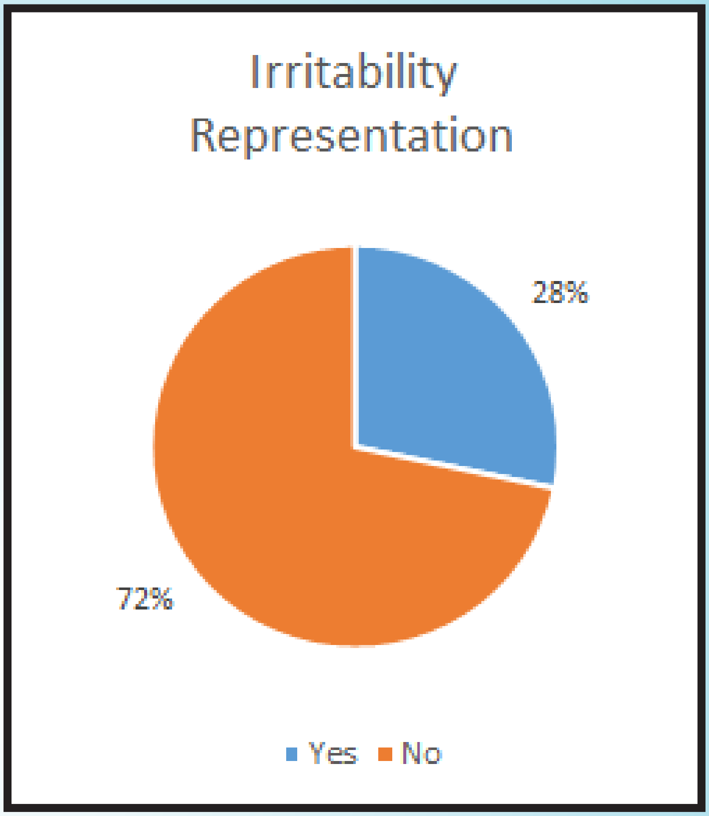
American Academy of Dermatology Association claims that itching and other dermatological illnesses may represent diabetes. THIs means that the data is helpful to this study. In the dataset, 51% of people claimed that they experience itching. The majority of them were also diagnosed with diabetes which means that this data will be useful to the algorithm.

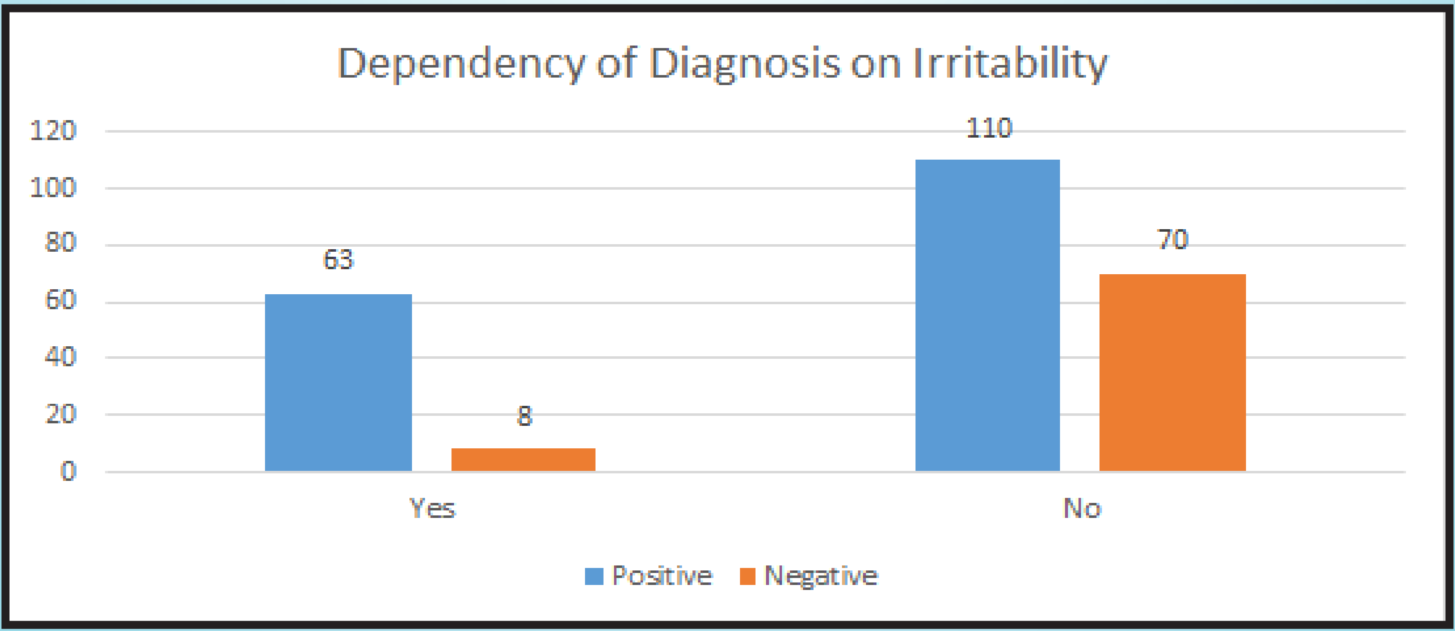




IRRITABILITY

According to the CDC, people with diabetes are 2 to 3 times more likely to experience mental health issues. As a result, irritability is a consequence of diabetes that is relevant to this study. As it can be concluded from provided graphs, the majority of people who claimed to be more irritable than usual, were diagnosed with diabetes. As a result, this data will be beneficial to the algorithm.

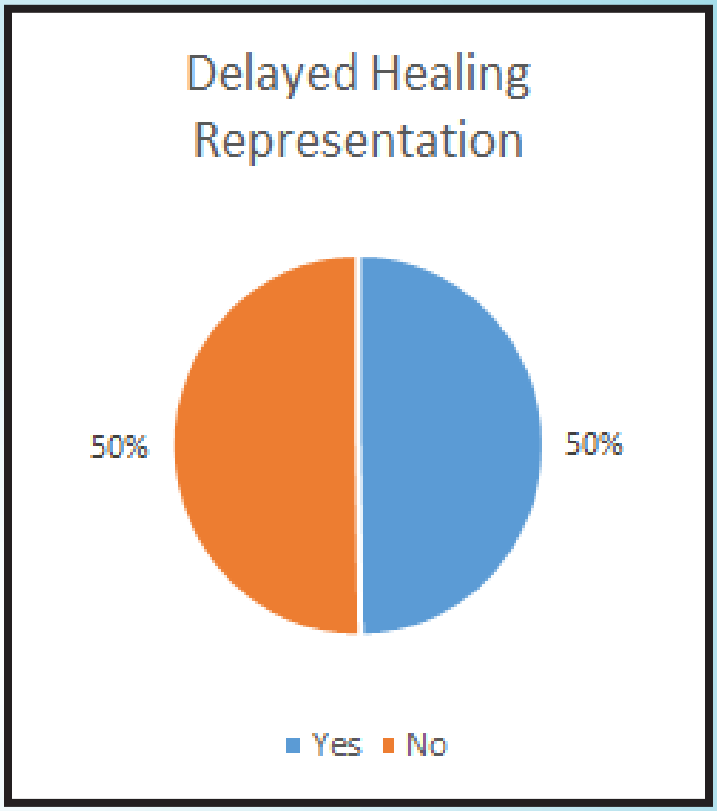


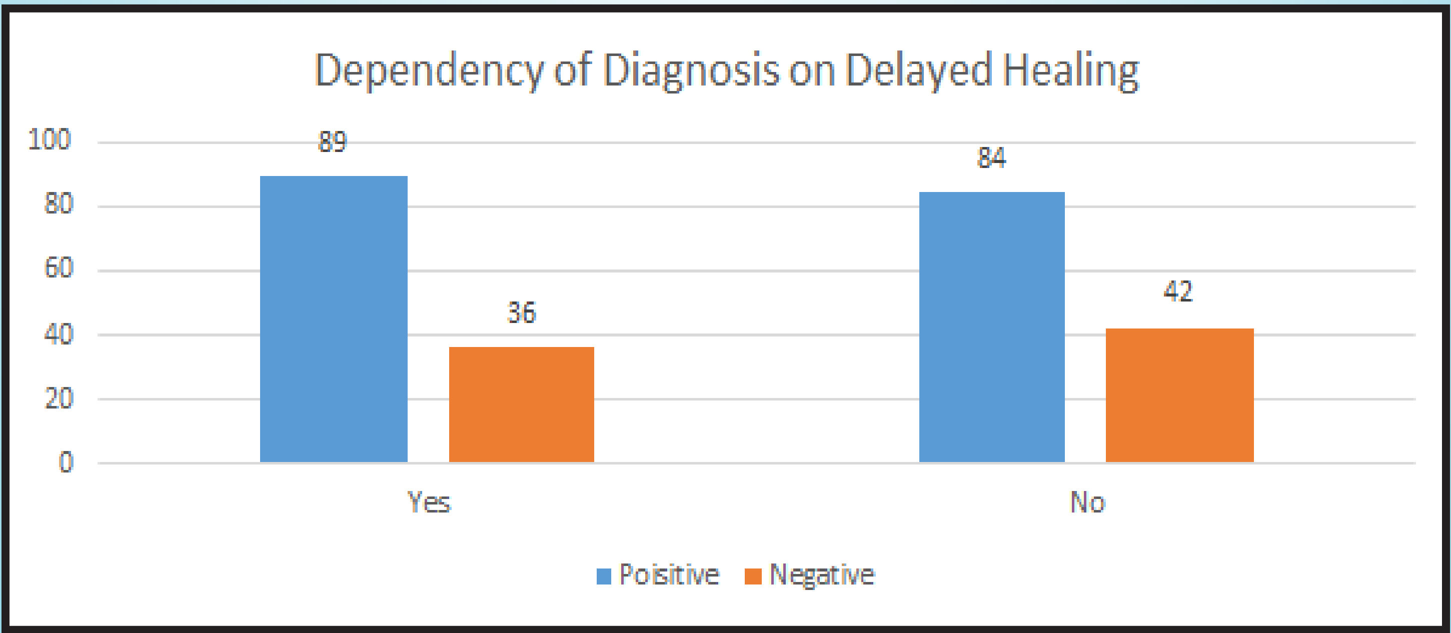


Anastasiia
Ganshina

DELAYED HEALING

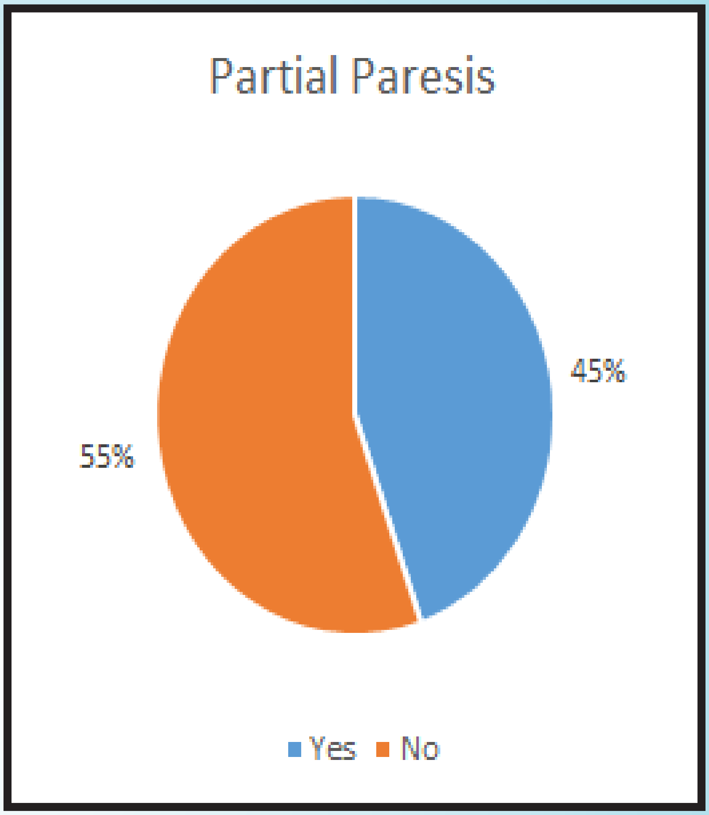
According to NIH studies, people with diabetes are more likely to have a poor immune response to wounds. As a result, this is a sign that a person might have diabetes. In the provided dataset the majority of people who had delayed healing response, were diagnosed with diabetes which means that the feature will be useful to the algorithm.

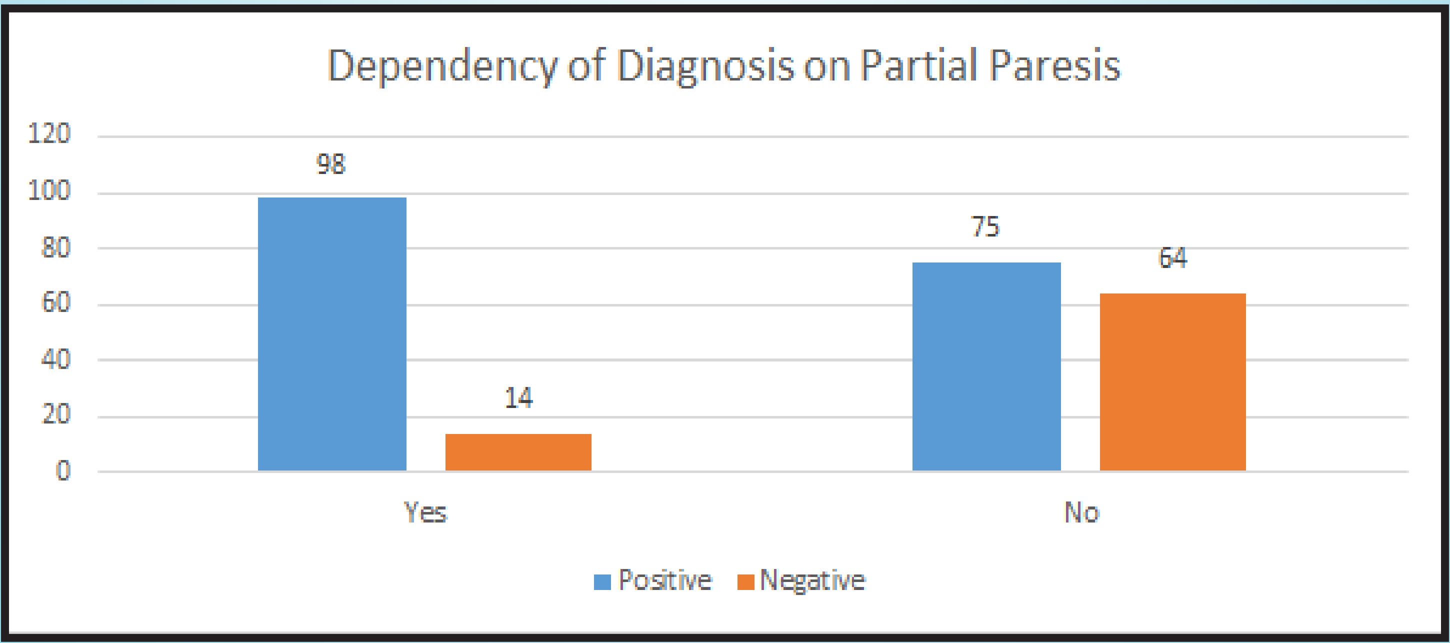




PARTIAL PARESIS

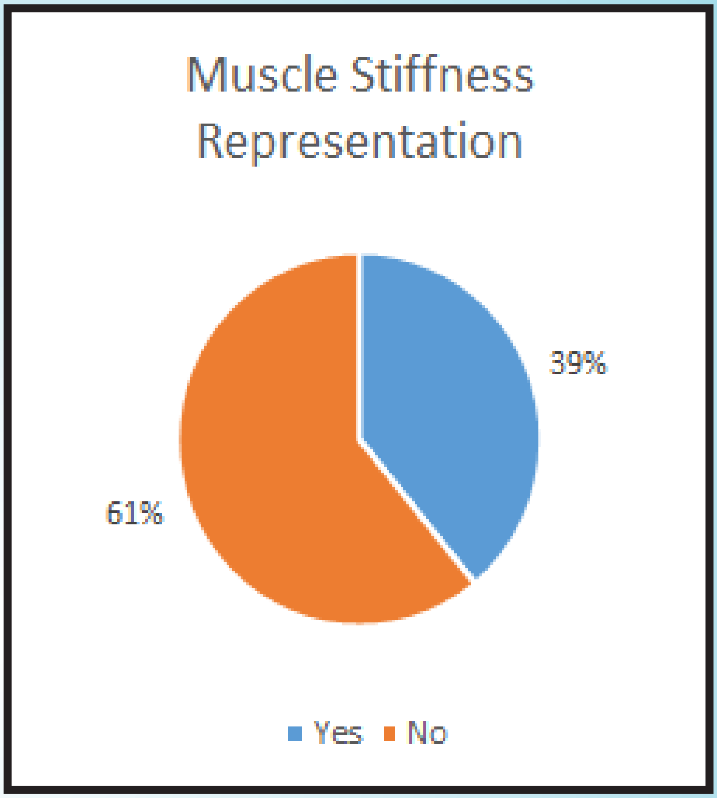
According to the CDC, paresis is a consequence of an untreated diabetes which is relevant to our study and might increase the accuracy of the algorithm. The dataset also has a diverse representation of people who have partial paresis, and the majority of those who do, were diagnosed with diabetes. As a result, the feature is very useful to the algorithm.

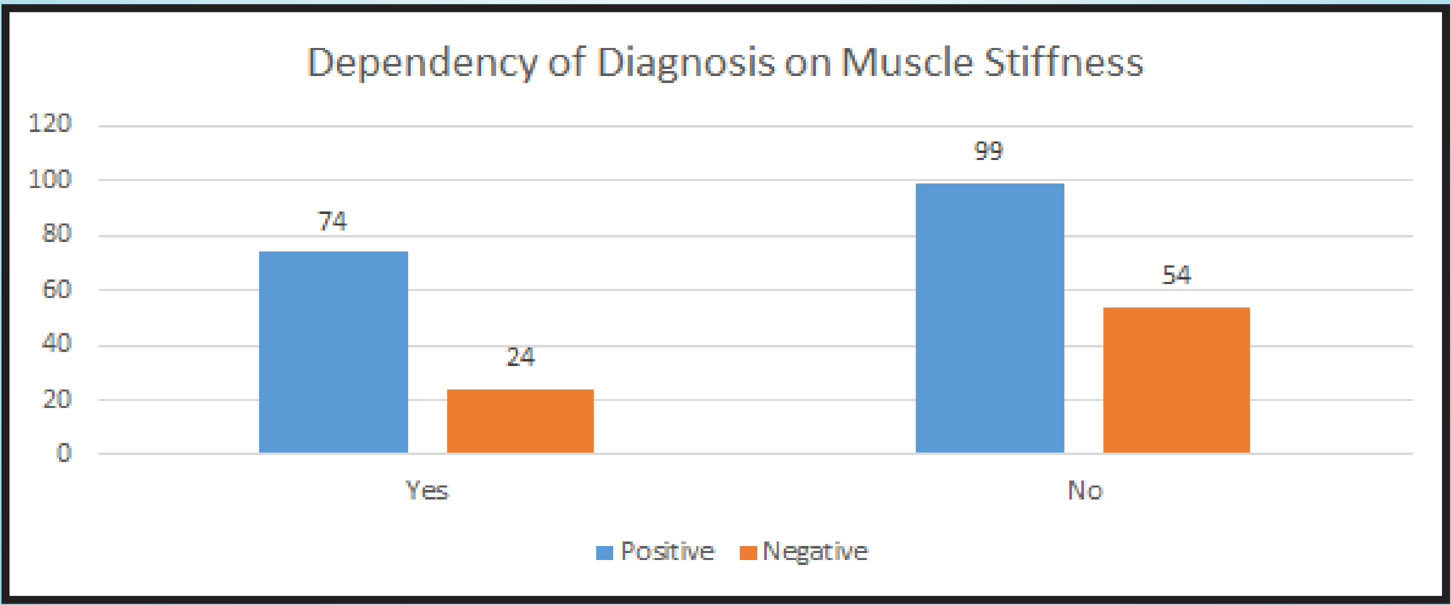




MUSCLE STIFFNESS

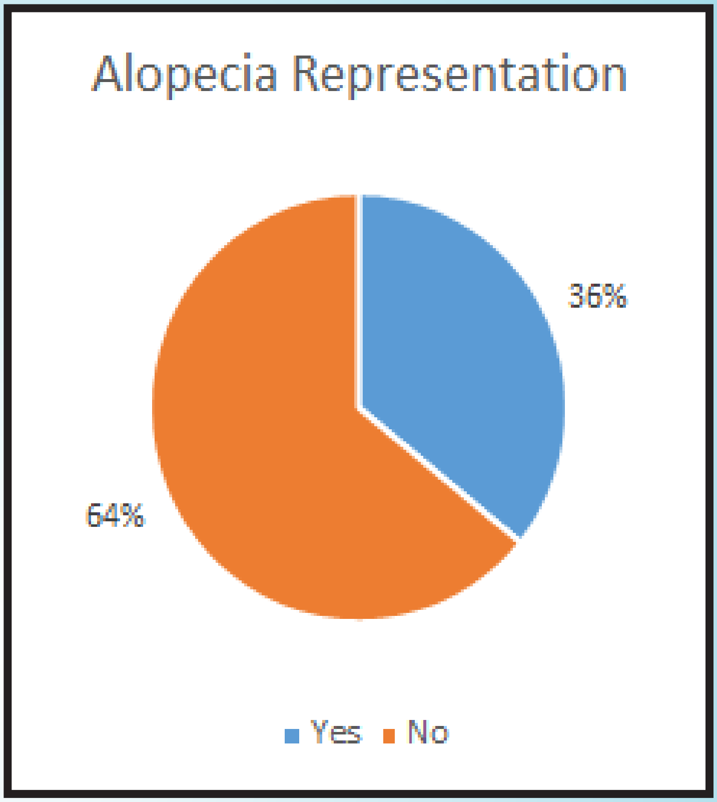
According to the CDC, muscle stiffness is a complication of diabetes which is relevant to the study. The graphs also show that in the studied dataset the majority of people who indicated that they have muscle stiffness, were diagnosed with diabetes. As a result, this data should be used to train the algorithm.

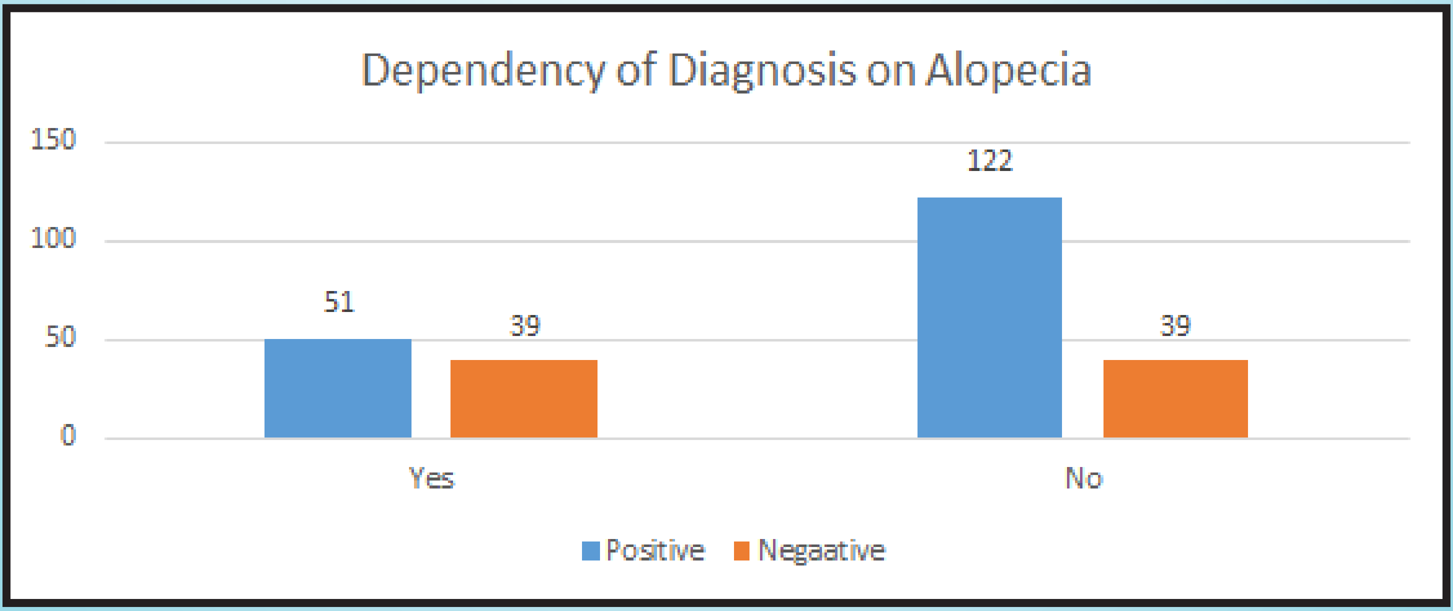




ALOPECIA

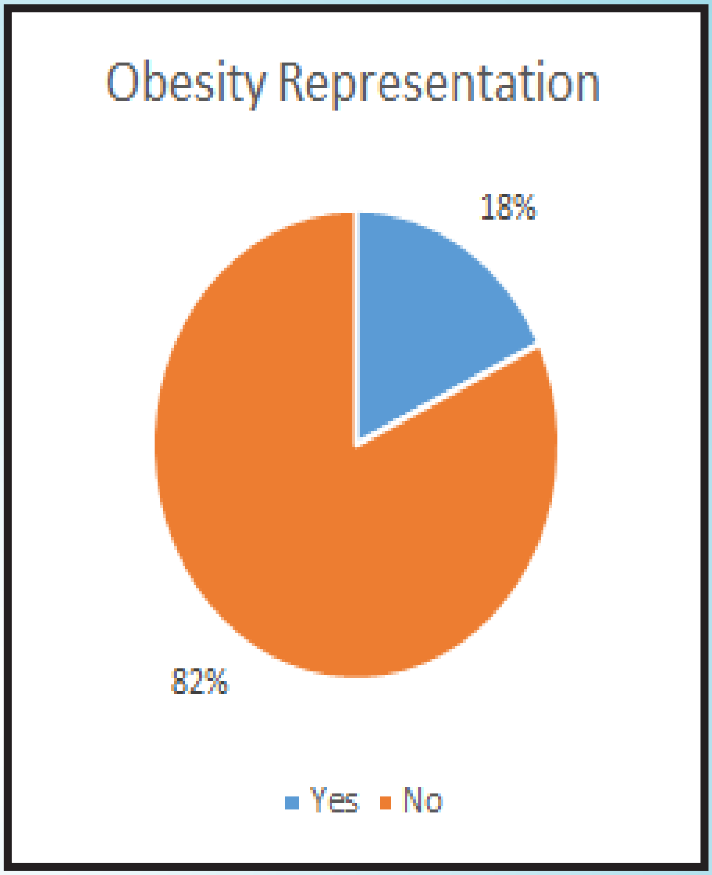
Alopecia is a condition which causes hair to fall off. According to NIH, alopecia is an early sign of metabolic processes that cause type 2 diabetes. As a result, this data should be included. However, because the data does not show strong dependency of diagnosis on whether a person has alopecia, the feature will not be used in the algorithm to prevent it from being corrupted.

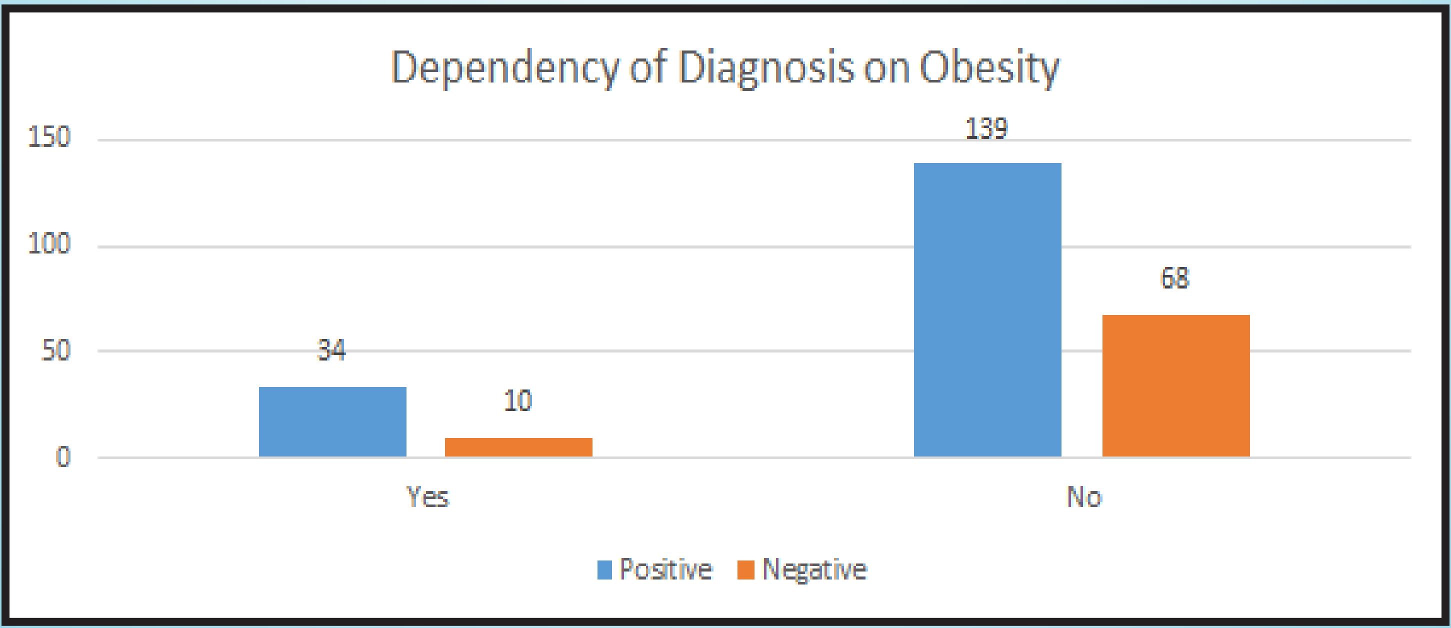




OBESITY

According to NIH research, obesity is the leading cause of type 2 diabetes. As a result this information should be included in this research. In the provided data, although the vast majority of patients in this data do not have diabetes, the majority of those who do, have diabetes diagnosed. This means that this data will help the accuracy of the algorithm to increase.





Anastasiia
Ganshina

ALGORITHM

The algorithm was designed by using binary logistic regression. This type of algorithm predicts to what class the outcome belongs based on estimated probability. This probability depends on a certain threshold that is determined by the training set.

This type of algorithm was chosen for several reasons:
1. The number of the features is less than the number of observations which prevents the function from overfitting.
2. The data is linearly separable which leads to high efficiency of the algorithm.

To Implement logistic regression, Python and its libraries were used. More specifically, the libraries that are used are numpy, pandas, and sklearn. The code can be found by following the QR code provided.

https://github.com/avganshina/diabetes_predictions/blob/main/diabetes_prediction.ipynb

To learn more about loginsic regression, consider these links:

1. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
2. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Anastasiia
Ganshina

FINDINGS

After deleting the features, such as gender and alopecia, that might potentially corrupt the algorithm, modeling the algorithm and dividing the dataset into testing and training sets, the finding shows that the algorithm is 76% accurate. This percent of accuracy is predicted and with more data in the training set, this percent of accuracy may increase.

Future developments of this project include data collection and development of the cluster of similar algorithms that can then be used to determine the types and severity of patients' diabetes.

RESOURSES

https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

https://pubmed.ncbi.nlm.nih.gov/11206408/

https://www.cdc.gov/diabetes/basics/symptoms.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC65518/#:~:text=Diabetes%20mellitus%20predisposes%20individuals%20to,marker%20of%20diabetes%20%5B9%5D.

https://www.aad.org/public/diseases/a-z/diabetes-warning-signs

https://www.cdc.gov/diabetes/managing/mental-health.html

https://www.nih.gov/news-events/nih-research-matters/poor-immune-response-impairs-diabetic-wound-healing

https://www.cdc.gov/diabetes/managing/problems.html

https://www.cdc.gov/diabetes/basics/diabetic-ketoacidosis.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5073072/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066828/

AKNOWLEDGEMENTS

This project would not be possible without Jane Strode Miller who is a funder of the internship. The Jane Strode Miller Micro-Internship motivated me to start the idea I had in my mind for a long time.

Special thank you to Bastian Center for Career Success at Knox College for promoting this internship and connecting me with two mentors: Kameron Wells and Jacob Scholl.

This project might be considered a continuation of my life goal: helping people struggling with an incurable disease via creating a software that lets them manage the condition.