Open programme

Heart diseases prediction

Prepared by

Alexandru Tcaciuc

Artificial Inteligence

Table of contents

Introduction

This report gives an overview on the actions made in the project and explains where the project should be orienting in the future.
This project has been requested by a private medical clinic who is interested in better and faster scheduling of the patients based on the data they can easily obtain. Specifically, heart diseases.
Heart diseases are the most widespread cause of death, which mostly affects the older part of the population.

These diseases can include:

- Abnormal heart rhythms, or arrhythmias
- Aorta disease and Marfan syndrome
- Congenital heart disease
- Coronary artery disease (narrowing of the arteries)
- Deep vein thrombosis and pulmonary embolism
- Heart attack
- Heart failure
- Heart muscle disease (cardiomyopathy)
- Heart valve disease
- Pericardial disease
- Peripheral vascular disease
- Rheumatic heart disease
- Stroke
- Vascular disease (blood vessel disease)

The goal is creating a machine learning algorithm that could predict a heart disease for further scheduling.

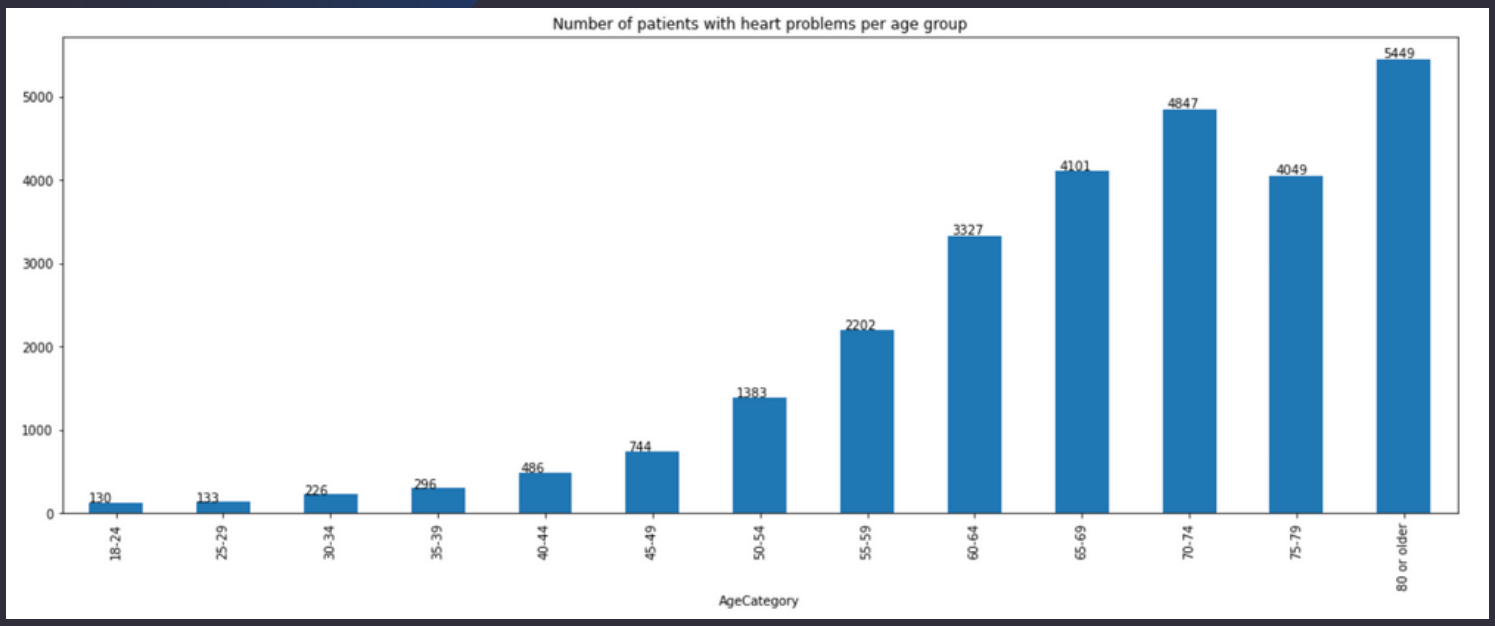
Data Description

- HeartDisease - target variable
- BMI - Body Mass Index
- Smoking - Have you smoked at least 100 cigarettes in your entire life?
- AlcoholDrinking - Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
- Stroke - Were u ever told you had a stroke?
- PhysicalHealth - For how many days during the past 30 have you experienced physical illness or injury?
- MentalHealth - for how many days during the past 30 days was your mental health not good?
- DiffWalking - Difficulty Walking
- Sex - Are you male or female? No other options.
- AgeCategory - Fourteen-level age category
- Race - Imputed race/ethnicity value
- Diabetic - Were u ever told you had diabetes?
- PhysicalActivity - Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
- GenHealth - Would you say that in general your health is... (5 unique choices)
- SleepTime - On average, how many hours of sleep do you get in a 24-hour period?
- Asthma - Were u ever told you had asthma?
- KidneyDisease - Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
- SkinCancer - Were u ever told you had skin cancer?

Data Description

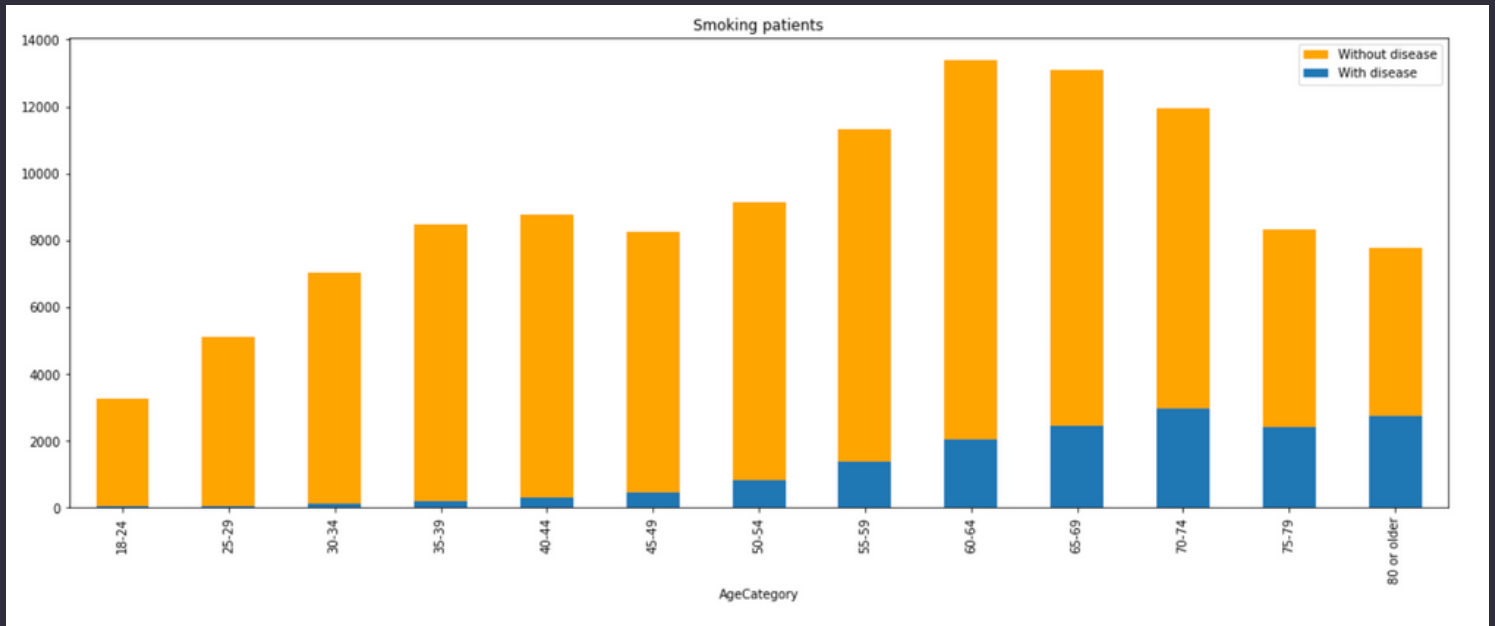
Age Category

During the exploratory data analysis there has been noticed a high growth in the number of patients who do have a heart disease as they get older. This indicated that this feature is suitable for future modelling and is also very easy to obtain.



Smoking

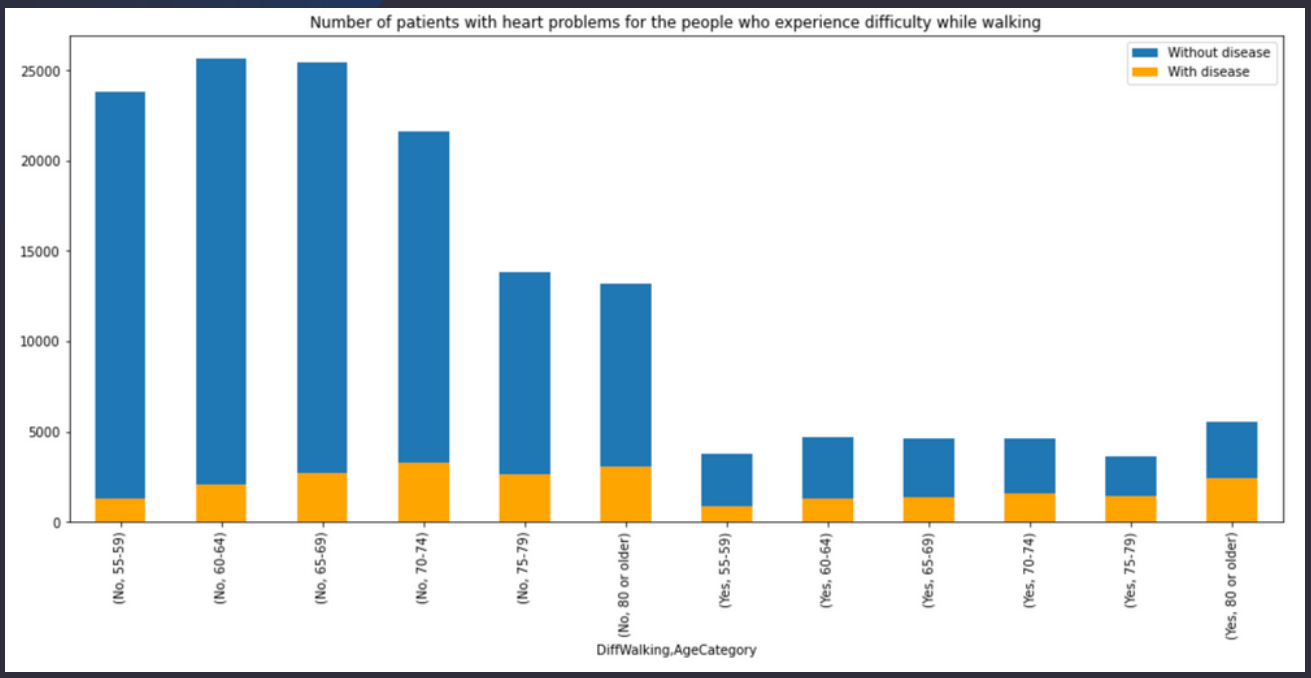
Smokers are also in the high risk of illnesses, as there are more smokers among sick people and the proportion of smoking and sick people also grows by age



Data Description

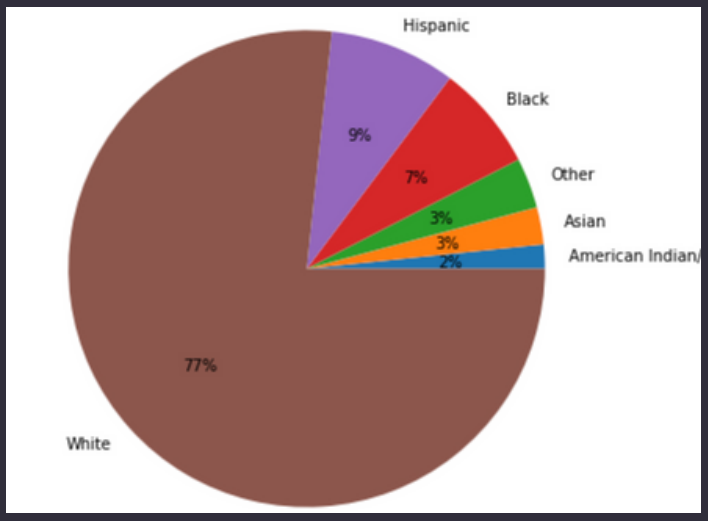
Difficulty Walking

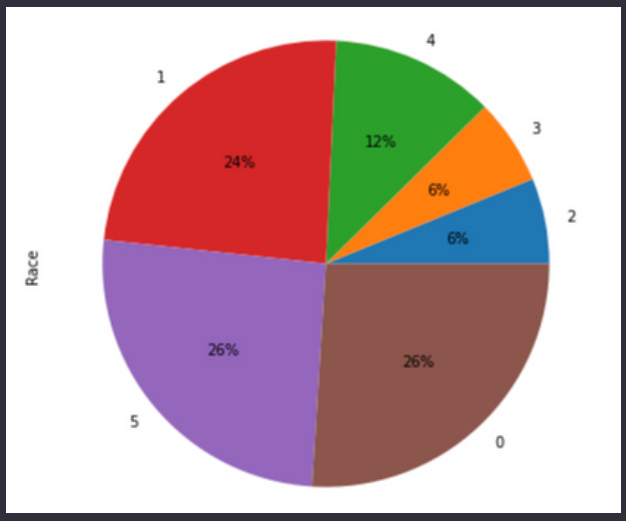
The higher proportion of the people who experience a difficulty walking also leans toward people who are sick, thus making it also a good feature for machine learning in the future.



Race

A problem I tried to prevent is the model getting biased towards a specific race. But, after I tried to balance it out, the race still did not show correlation with the rest of the features and I decided to not put more effort in this.





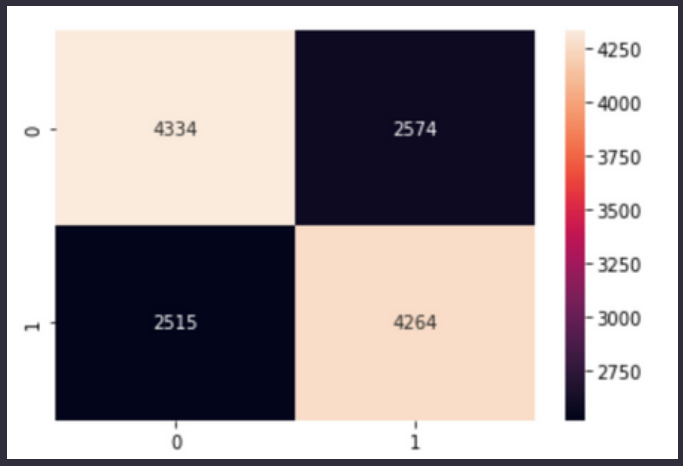
Model Algorithms

- kNN - Useful in development for being fast.
- SVM - Can be fine-tunable, precise and heavy.

Only the results of the SVM will be discussed since its more sophisticated, does not overfit and allows us to get the most out of a given dataset. This machine learning algorithm allows us to get the most out of given data.

Results

Iteration 1
Features: Smoking, PhysicalHealth, DiffWalking, Sex, AgeCategory, Diabetic, PhysicalActivity.
Evaluation: Not great results overall, the number of false positives and negatives is high (False negative - black cell on the left, False positive - black box on the right). The patient would be way safer if placed by a human being in the field.

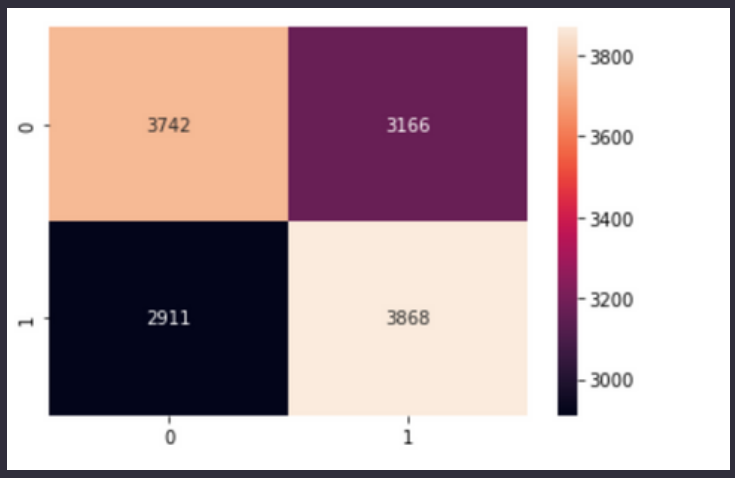


Model Algorithms

Iteration 2

Features: PhysicalHealth, DiffWalking, AgeCategory.

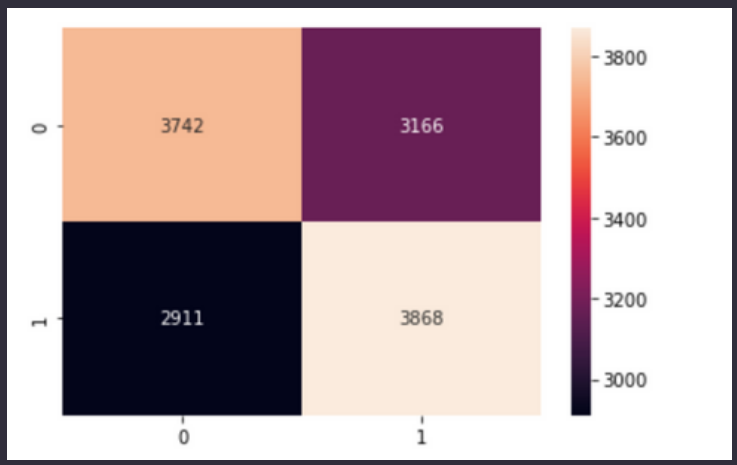
Results are not better than the iteration 1, but provided with better data, and trained with better data (which would raise the iteration 1's accuracy higher, I can see the model perform just a bit worse by having these features.



Iteration 3

Features: Stroke, DiffWalking, AgeCategory.

Also did not perform well considered we gave it a direct consequence of a heart disease. This once again asks for better quality data in the future research.



Impact Assessment

Among the projects that have the scope of replacing the specialist, due to its focus on creating a product that is more likely a handy business tool, it could help patients get help on time.

With this software a patient who needs more urgent healthcare might be scheduled too late due to the low recall scores. Also a healthier patient might be misplaced earlier making less room for the one that needs more attention.

Used in large scale we can see cases where a patient gets scheduled too late and dies before reaching his appointment.
Even though real medical workers are also prone to mistakes, giving wrong diagnoses and sometimes spreading misinformation, a real person making mistakes is way more socially accepted than a machine making mistakes.
A machine that keeps misjudging its patients time can result in it hitting the newsletters, getting complaints, getting the masses skeptical and evolve into governmental bans.

QUICKSCAN - CANVAS

AI

NAME: AI
DATE: June 23, 2022 2:46 PM
DESCRIPTION OF TECHNOLOGY
Heart disease prediction

HUMAN VALUES

The technology does not affect the identity since a patient is not supposed to know the result of the machine learning. It might be though that he is scheduled too late due to poor recall and get frustrated.

TRANSPARENCY

In this project, the notebook is explained in a clear way so that users can understand it well without further knowledge. The goals can be found in the document and the ideas as well. So everything is clear related to that

IMPACT ON SOCIETY

People get late schedules late on potentially serious problems. When a medical worker is known to be good, his schedule becomes tight and becomes harder to prioritise each patient correctly.

STAKEHOLDERS

- Client
- Patients

SUSTAINABILITY

We should take in account that there should be a technology change in the system and that the information should be saved of the patients.

HATEFUL AND CRIMINAL ACTORS

This technology cannot be used to break the law.

DATA

This dataset is clean and therefore doesn't need much of adjustments. I am aware of how well the data is and therefore we can conclude that it wont give any problems with proceeding.

FUTURE

We could save lives

PRIVACY

The technology does use personal data, data about the lifestyle, harmful habits and current diseases.

INCLUSIVITY

Yes, it might get biased by race but of course I will look into it and see if i can find more information on this and try to overcome this problem

FIND US ON WWW.TICT.IO

THIS CANVAS IS PART OF THE TECHNOLOGY IMPACT CYCLE TOOL. THIS CANVAS IS THE RESULT OF A QUICKSCAN. YOU CAN FILL OUT THE FULL TICT ON WWW.TICT.IO

Deployment recommendation

Due to its accuracy score of roughly 63%, It is not advisable to use it in healthcare. The algorithm creates many false negatives which is unsuitable for healthcare.

More on that you can read in the jupyter notebook.

Deployment of this product in a business is not advisable at this stage.

Conclusion

This project should be under further research and development. Healthcare requires sophisticated products, it is better to spend more time and resources on getting higher prediction scores than now, unlike the IT companies that release underdeveloped products, and make the users wait for future updates to match with the deadline set by the product owners to furthermore raise the return on investment of the company. A better quality dataset would facilitate future research, as the algorithms I've tried to used already got to the potential of the current dataset.