

**Dokumentacja projektu**

Przedmiot: **Zaawansowane technologie programowania**

Tytuł projektu:

Budowa programu wykorzystującego QDA do predykcji prawdopodobieństwa zagrożenia zdrowia.

 Prowadzący: Wykonawca:

*Dr inż. Arkadiusz Lewicki Jan Drabek w61064*

Semestr i symbol kierunku:  7 IID-P

Grupa: 2018-SPL01

Rzeszów, 2022

Spis treści

[Charakterystyka problemu i cel projektu 3](#_Toc92837199)

[Charakterystyka problemu 3](#_Toc92837200)

[Cel projektu 3](#_Toc92837201)

[Źródło danych 3](#_Toc92837202)

[Informacje o rozpatrywanych atrybutach 3](#_Toc92837203)

[Implementacja 5](#_Toc92837204)

[Biblioteki i moduły wykorzystywane w realizowanym rozwiązaniu 5](#_Toc92837205)

[Struktura projektu 5](#_Toc92837206)

[Przedstawienie najważniejszych elementów 5](#_Toc92837207)

[Wyniki 7](#_Toc92837208)

[Wnioski 13](#_Toc92837209)

# Charakterystyka problemu i cel projektu

## Charakterystyka problemu

Choroby o charakterze sercowo-naczyniowym są jedną z najczęstszych przyczyn zgonów występujących nie tylko w Polsce, ale i na świecie. Na podstawie badań przeprowadzonych w latach 2009-2012 w Polsce, co roku ponad 80 tys. osób umierało z powodu zawału serca. Jak podają dane 60% osób zapadających na tę chorobę to mężczyźni. W ramach raportu przeprowadzonego w 2014 roku na podstawie Narodowej Bazy Danych zawałów serca można dowiedzieć się, iż główną przyczyną prowadzącą do zawału serca jest nieodpowiednio prowadzona profilaktyka przedwczesna. Koniecznym jest więc regularne poddawanie się wszelkim badaniom, które mogą ukazać ewentualne problemy w funkcjonowaniu i odpowiednia reakcja przeciwdziałająca im. Jak pokazują dane, śmiertelność u osób, które na wczesnym etapie wykryły chorobę jest niska. Pokazuje to, jak ważnym czynnikiem jest odpowiednie podejście do tego problemu i przedwczesne działanie prewencyjne.

Źródło badań: <http://www.pro-plus.pl/sites/default/files/tele/2014-05-11-Raport-ZS-w-Pl-ost.pdf>

## Cel projektu

Celem tworzonego projektu było stworzenie modelu, który będzie w stanie obliczyć prawdopodobieństwo, czy osoba o konkretnych danych zdrowotnych jest narażona na zachorowanie na chorobę układu sercowego. Klasyfikator zbudowany został na podstawie przeanalizowania kilkuset przypadków tj. w ramach 13 cech, które zostały uznane za najbardziej istotne ze zbioru 75 atrybutów oraz decyzji w postaci wyniku określającego czy dana osoba zachorowała na atak układu sercowego, czy też nie. W ramach projektu rozważonych zostało 270 obserwacji .

## Źródło danych

<https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29>

## Informacje o rozpatrywanych atrybutach

1. Wiek pacjenta.
2. Płeć pacjenta (1.0: Mężczyzna, 0.0: Kobieta).
3. Rodzaj bólu w klatce piersiowej (1.0 -Typowa dusznica bolesna , 2.0-Nietypowa dusznica bolesna, 3.0-Ból niezwiązany z dusznicą bolesną, 4.0-Bezobjawowy).
4. Ciśnienie w spoczynku.
5. Cholesterol.
6. Cukier w krwi na czczo (1.0> 120 mg/dl, 0- w przeciwnym wypadku).
7. Wartość ECG w spoczynku (0.0-Normalny, 1.0-Nieprawidłowości załamka ST-T, 2.0- LVH.
8. Maksymalne tętno (Wartość numeryczna pomiędzy 60 a 202).
9. Dławica przy wysiłku(0.0-nie, 1.0-tak).
10. Obniżenie ST wywołane wysiłkiem fizycznym w stosunku do odpoczynku (Wartość numeryczna mierzona w depresji).
11. Nachylenie szczytowego odcinka ST wysiłkowego(1.0-wznoszące, skierowane w górę, 2.0-płaske, 3.0-spadkowe,nachylone w dół).
12. liczba głównych naczyń zabarwionych fluoroskopowo (wartość numeryczna pomiędzy 0 a 3) .
13. Thal (3.0- prawidłowy, 6.0-wada naprawiona, 7.0- wada możliwa do naprawy).
14. Klasa (2-choroba serca, 1-brak choroby serca)

# Implementacja

## Biblioteki i moduły wykorzystywane w realizowanym rozwiązaniu

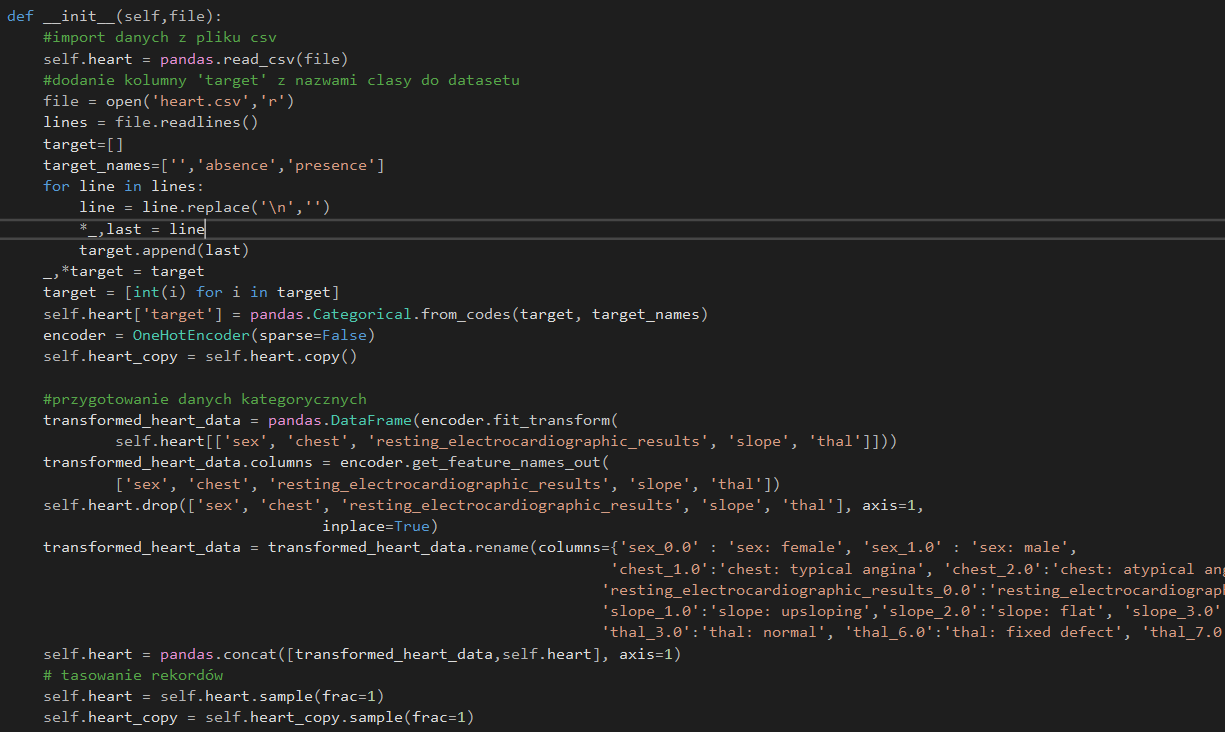
* Scikit-learn (1.0.1) biblioteka wykorzystywana do uczenia maszynowego przy wykorzystaniu języka Python. Biblioteka ta zawiera różne algorytmy służące do budowania klasyfikatorów, w tym projekcie wykorzystany zostanie algorytm kwadratowej analizy dyskryminacyjnej.
* Pandas (1.3.4) jest biblioteką napisaną w języku Python i wykorzystywana jest do manipulacji oraz analizy danych.
* Matplotlib (3.2.1) jest biblioteką która pozwala na tworzenie wykresów dla języka Python.
* Seaborn (0.11.2) rozszerzenie biblioteki Matplotlib
* Numpy (1.18.2) prosta, lecz potężna biblioteka do operacji na tablicach.

## Struktura projektu

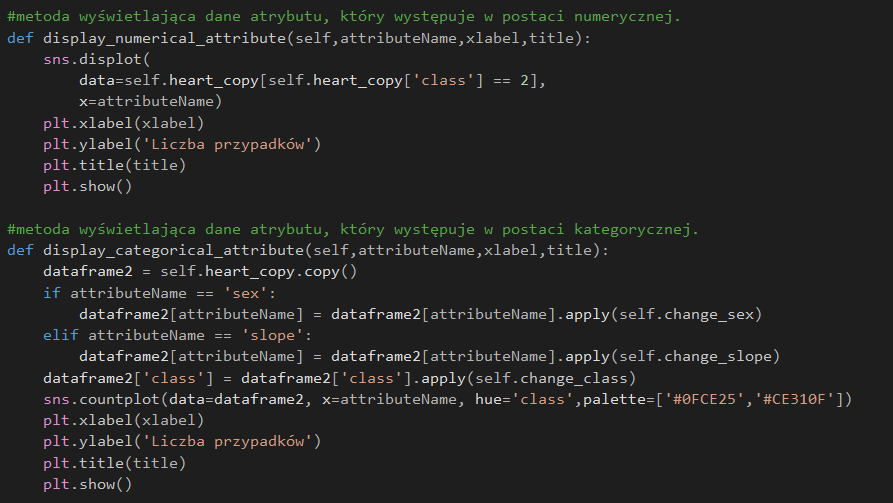
* Moduł dataset odpowiedzialny jest za inicjację danych oraz wizualizację informacji dotyczących rozpatrywanych obiektów.
* Moduł model odpowiada za tworzenie modelu kwadratowej analizy dyskryminacyjnej oraz wyświetlanie informacji związanych z procesem tworzenia jak i otrzymanymi wynikami.
* Moduł main pełni rolę wywoławczą dla pozostałych modułów.

## Przedstawienie najważniejszych elementów

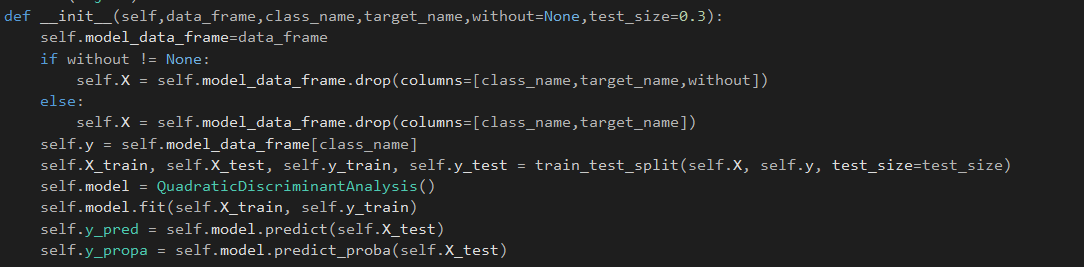
Zaimportowanie danych metodą pandas.read\_csv, oraz przygotowanie danych kategorycznych przy pomocy OneHotEncodera:



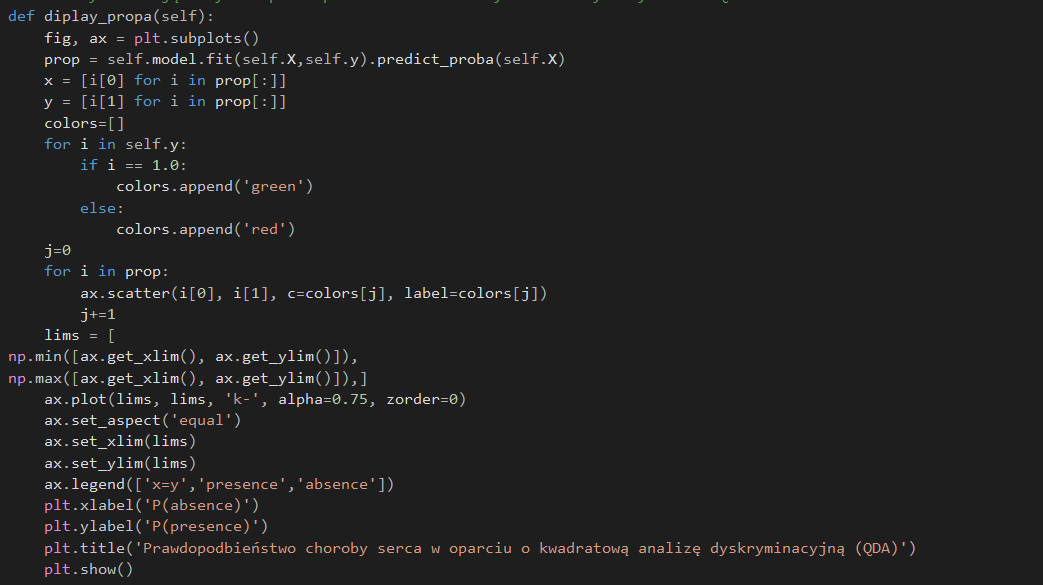
Metody użyte do wyświetlania wykresów dla atrybutów numerycznych oraz kategorycznych:



Inicjowanie modelu QDA:



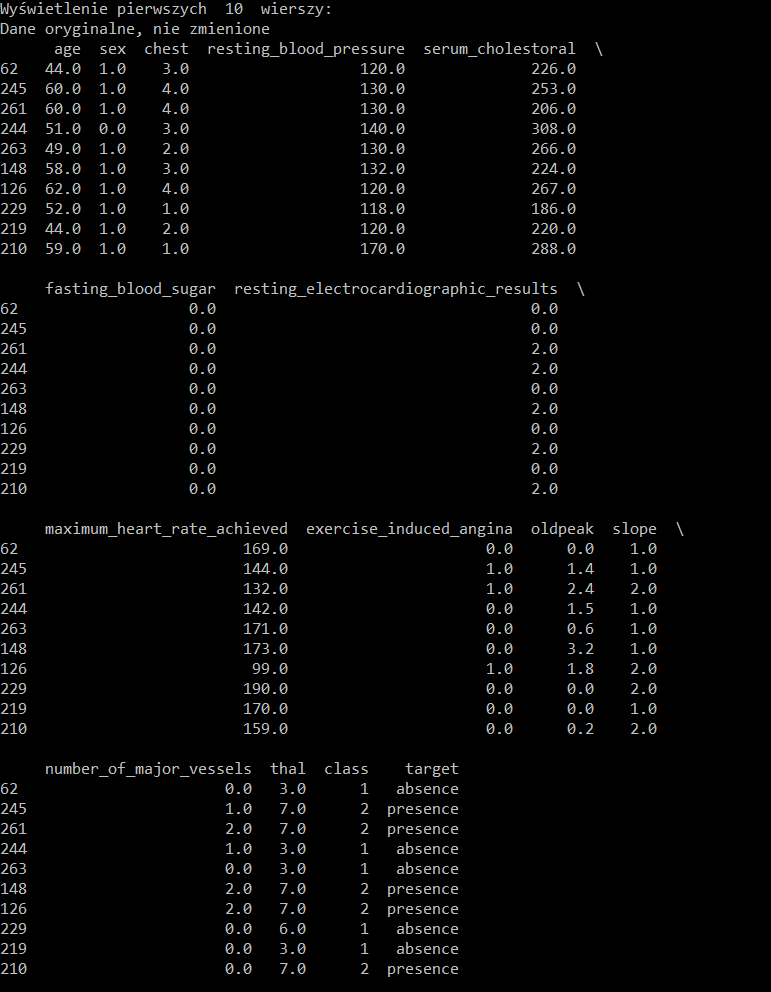
Metoda wyświetlania wykresu prawdopodobieństwa choroby:



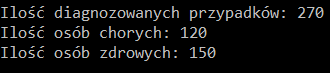
# Wyniki

Wszystkie poniżej przedstawione zrzuty ekranu pochodzą z okna konsoli, wykresu matplotlib lub ze środowiska visual studio

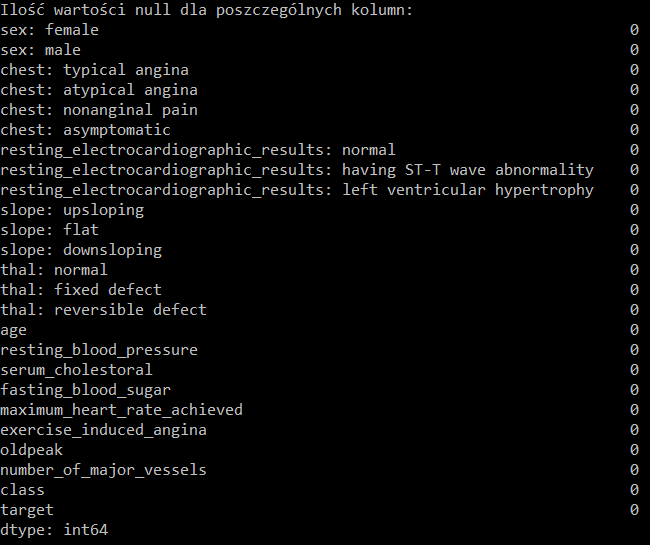
Przedstawienie 10 pierwszych rekordów:



Wykaz ilości wszystkich obiektów z podziałem na chorych i zdrowych:

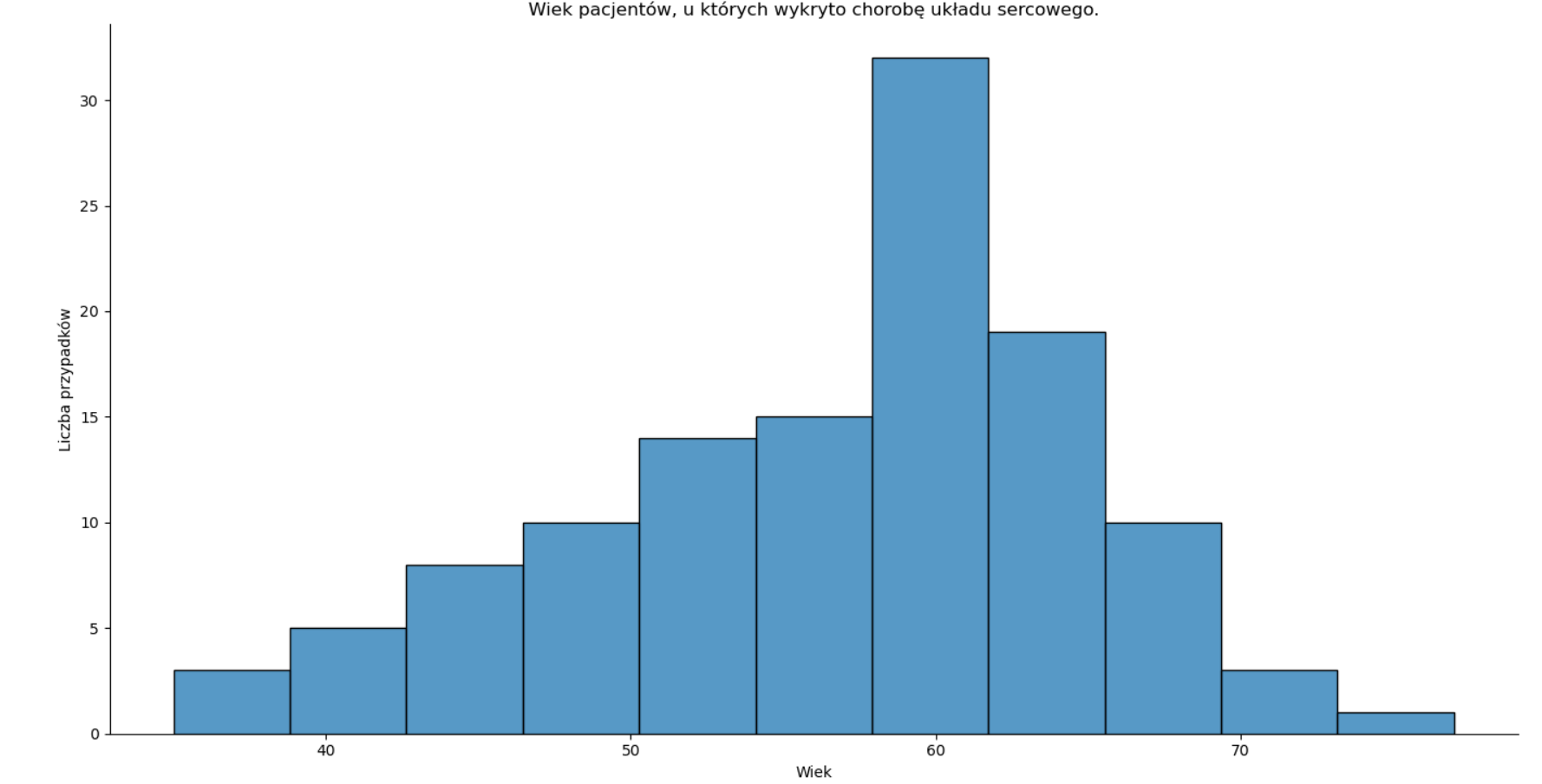


Sprawdzenie czy w naszych danych nie ma żadnego nulla:

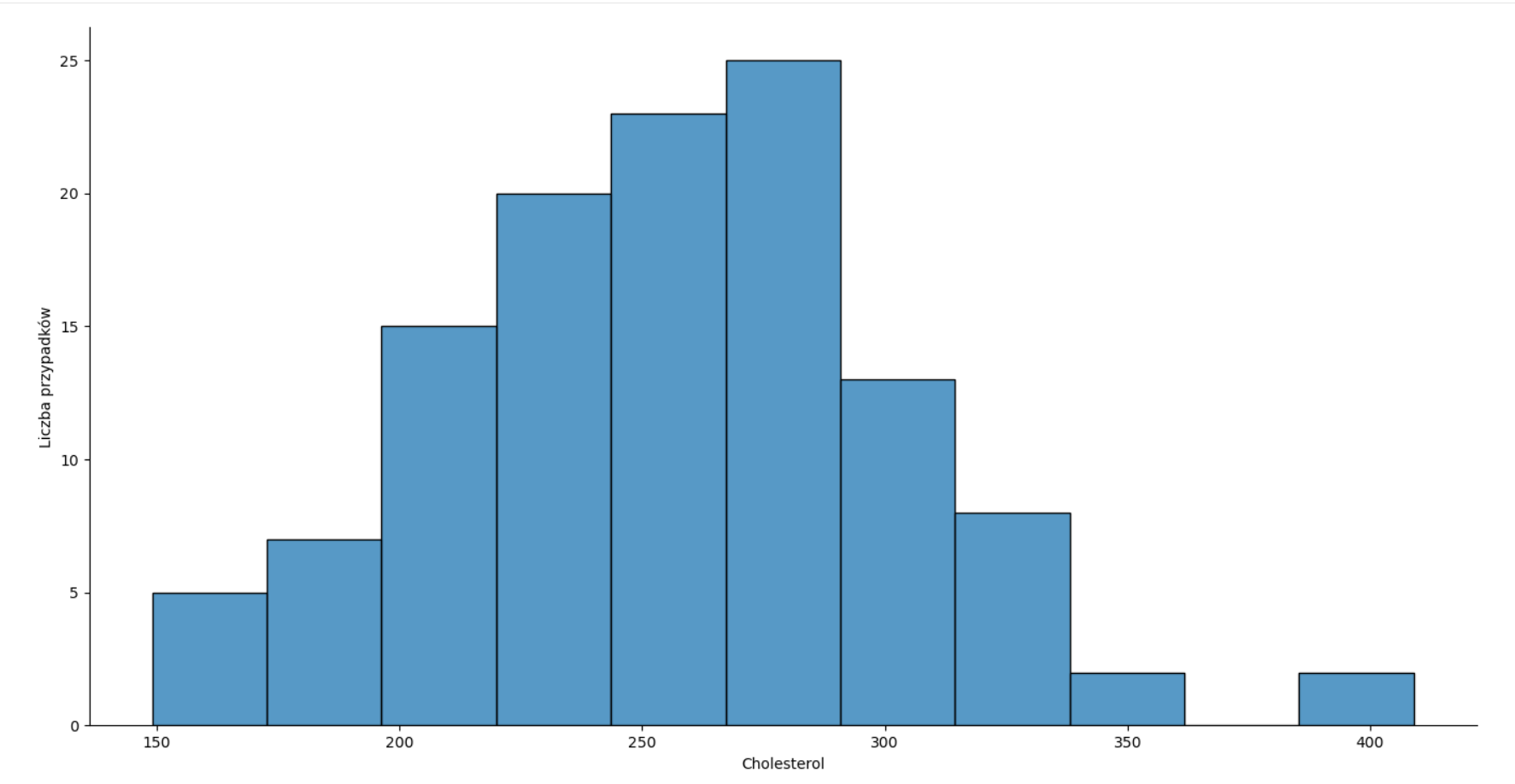


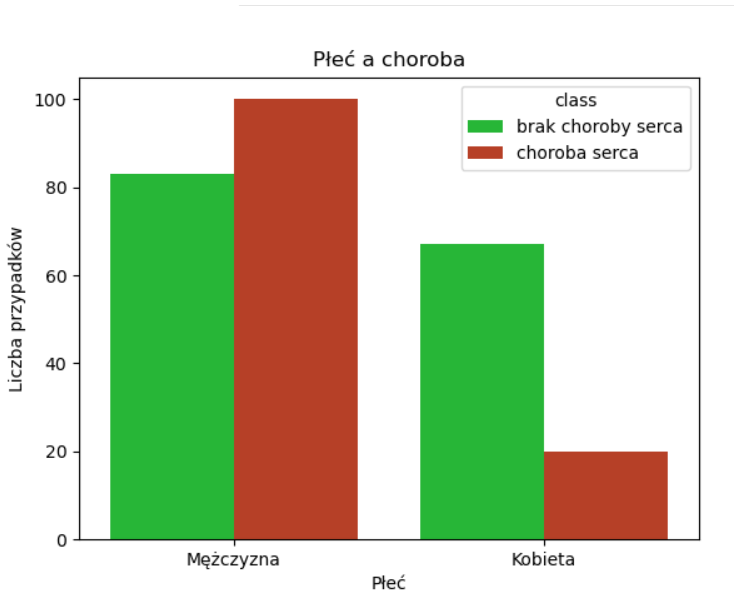
Poniżej został zaprezentowany wpływ wybranych atrybutów na choroby serca.

Wykres wieku pacjentów, u których wystąpiła choroba serca:

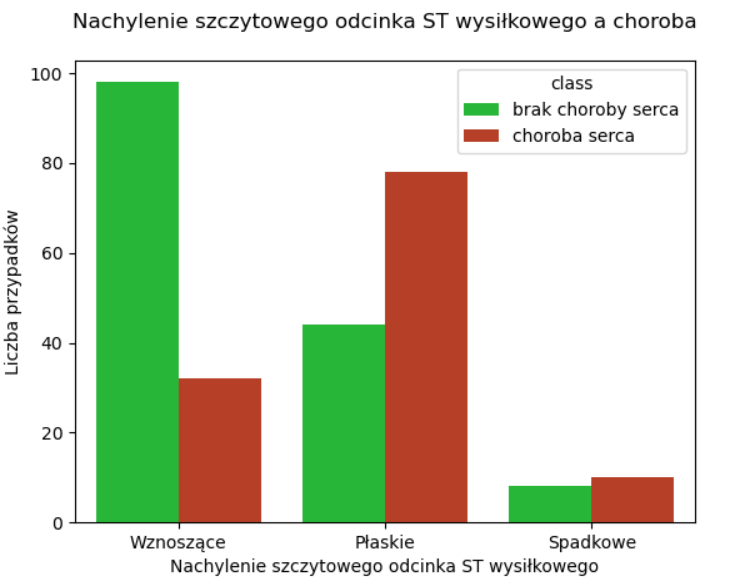


Wykres cholesterolu we krwi u pacjentów, którzy są chorzy:

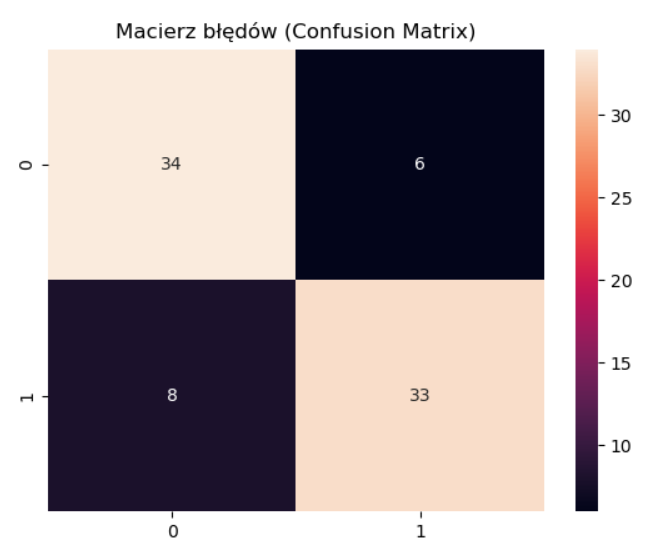


Następny wykres pokazuje jak płeć wpływa na wystąpienie problemów z sercem: 

wykres, który przedstawia stosunek nachylenia szczytowego odcinka ST do problemów z sercem:



Macierz błędu klasyfikatora:

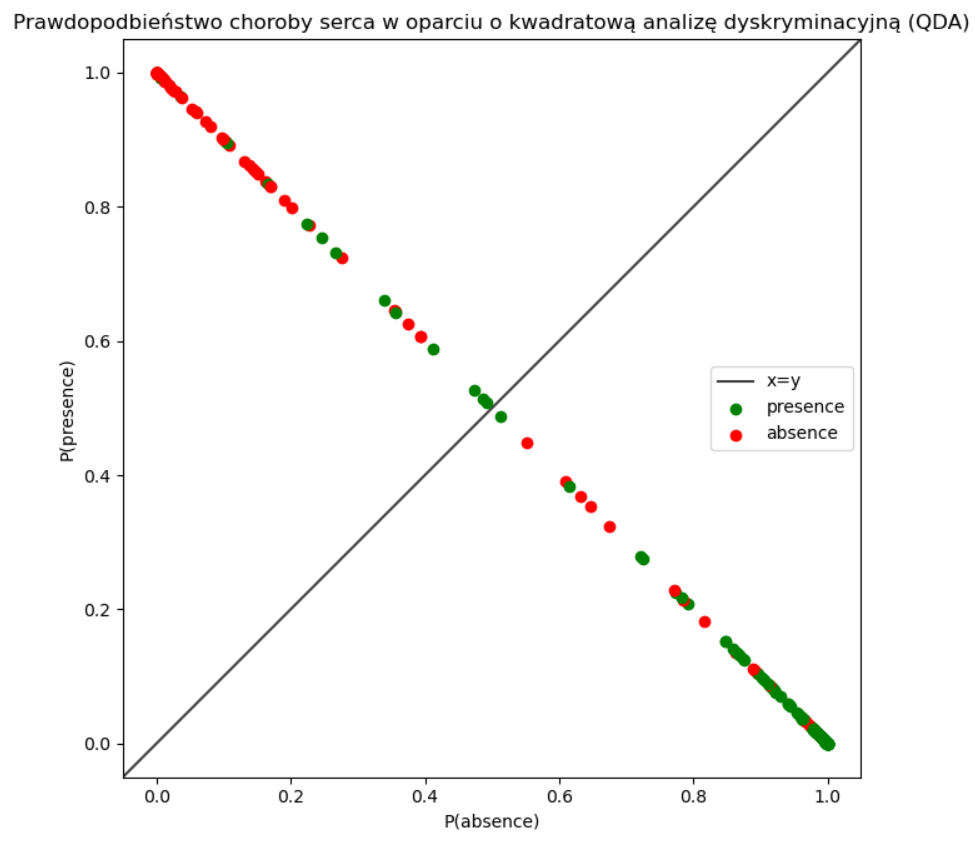


Jak widać stworzony klasyfikator sklasyfikował poprawnie 67 przypadków, błędnie tyko 14 co daje 83% poprawnie sklasyfikowanych obiektów.

Wynik predykcji po 100 testach:

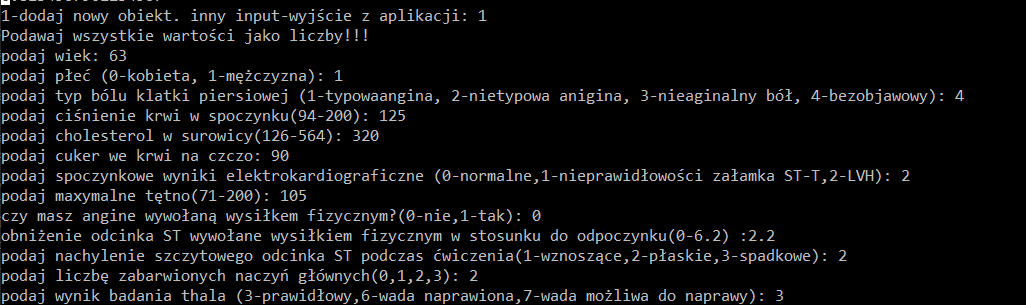


Prawdopodobieństwo choroby w oparciu o model kwadratowej analizy dyskryminacyjnej(QDA):



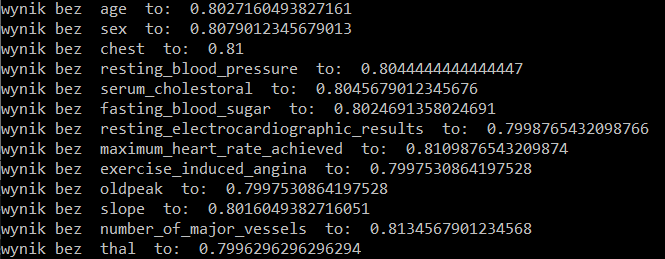
Aby obiekt klasy „presence” był sklasyfikowany poprawnie powinien być poniżej linii x=y, natomiast dla klasy „absence” powinien być powyżej x=y. Jak widać skonstruowanemu modelowi zdarzają się pomyłki w obie strony.

Obliczanie prawdopodobieństwa dla nowych obiektów:





Poniżej zaprezentowano skuteczność klasyfikatora po 100 testach po odłączeniu jednego atrybutu:



Jak widać wyniki bez jednego atrybutu jest porównywalne do tych ze wszystkimi atrybutami.

Jedna można stwierdzić, że najmniejszy wpływ ma liczba głównych naczyń zabarwionych fluoroskopowo, ponieważ bez tego atrybutu udało się uzyskać nawet wyższy wynik, niż ze wszystkimi. Natomiast największy wpływ na wynik mają: Wartość ECG w spoczynku, thal, Dławica przy wysiłku, Obniżenie ST wywołane wysiłkiem fizycznym w stosunku do odpoczynku. Tylko wyniki bez tych wartości uzyskały poniżej 80%.

# Wnioski

Zbudowanie klasyfikatora pozwoliło na otrzymanie ok. 81 procentowego prawdopodobieństwa wykrycia choroby serca na podstawie przedstawionych danych, zarówno rozpatrując wszystkie atrybuty, jak i bez jednego dowolnego atrybutu. Kilka razy zdarzyło się, że konstruowany model mimo ponad 90% prawdopodobieństwa na wystąpienie choroby lub 90% na jej brak, błędne klasyfikował dany obiekt. Zaobserwowane zjawisko jest bardzo niepokojące, ponieważ ktoś po użyciu modelu może stwierdzić, że nie jest chory, a rzeczywistość jest zupełnie inna. Jednakże taka sytuacja nie zdarzała się często. Jest to jedynie prosty program, a jednak pozwala z dużym prawdopodobieństwem przewidzieć narażenie na wystąpienie choroby.