Raport – Uczenie Maszynowe Predykcja Raka

Tomasz Wnykowicz, Marek Varisella, Oliwier Świtek

1. Zaimportowanie potrzebnych bibliotek takich jak numpy, sklearn czy pandas oraz zdefiniowanie funkcji w klasie Show Data do wyświetlania specyficznych danych i operacji na nich:Obraz zawierający tekst, zrzut ekranu, oprogramowanie, numer

   Opis wygenerowany automatycznie
2. Wyświetlenie pierwszych 5-ciu pozycji za pomocą funkcji CurrentData z datasetu w celu przeanalizowania zbioru danych pod kątem potencjalnych kolumn do normalizacji danych:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Nasz zbiór danych zawiera łącznie 26 kolumn i 1000 wierszy.

1. Analiza danych pod kątem dystrybucji zmiennej Level w datasecie, która opisuje poziom ryzyka zachorowania danego pacjenta na raka płuc:

Obraz zawierający zrzut ekranu, tekst, oprogramowanie, Ikona komputerowa

Opis wygenerowany automatycznie

1. Analiza danych pod kątem dystrybucji zmiennej Age, za pomocą funkcji AVGData z klasy ShowData, która pokazuje średnią wieku w naszym datasecie:

Obraz zawierający zrzut ekranu, tekst, diagram, Prostokąt

Opis wygenerowany automatycznie

Mediana wieku badanych osób wynosiła 36 lat. Górny kwartyl wieku wynosił 45 lat, a dolny 28.

1. Normalizacja danych polegająca na usunięciu niepotrzebnych czynników przed trenowaniem takich jak ID Pacjenta, Płeć czy Alergia na Pyłki. Dodatkowa zmiana wartości typu „string” na numeryczne odpowiedniki dla zmiennej Level. Te operacje są konieczne do trenowania naszego modelu (finalnie zostało nam 1000 wierszy i 7 kolumn) :

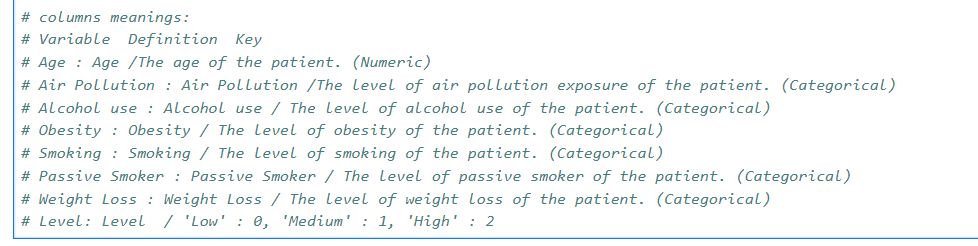
Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Po analizie do naszego finalnego modelu trafiają kolumny: “Age“, “Air Pollution“, “Alcohol use“,

“Obesity“, “Smoking“, “Passive Smoker“, “Weight Loss“, “Level“. Wszystkie wartości poza “Age”,

odnoszą się do skali natomiast “Age” to wartość numeryczna.



1. Wyświetlenie macierzy korelacji dla czynników, które postanowiliśmy, że będziemy przeznaczać do uczenia naszego modelu treningowego z dokładnością do .2f czyli dwóch miejsc po przecinku.:

Obraz zawierający zrzut ekranu, tekst, kwadrat, Prostokąt

Opis wygenerowany automatycznie

1. Zdefiniowanie klasy MainTrainingModel, która ułatwia nam wprowadzanie danych i użycie 3 modeli tj. Logistic Regression, Support Vector Machine oraz Neural Network do testów jednostkowych na podstawie 20% naszych danych pozostałe 80% użyliśmy do trenowania:

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

1. Przeprowadzenie dwóch treningów na modeli na danych z data.CurrentData() dla etykiety "Level". Pierwszy model jest tworzony bez żadnej transformacji, natomiast drugi model już ma włączoną transformację danych. Następnie stworzyliśmy dwa obiekty progression i progression2, które reprezentują obydwa modele regresji logistycznej, a następnie dla obu modeli są obliczane wyniki prognoz na podstawie danych dotyczących wieku i tego czy dana osoba jest biernym palaczem.



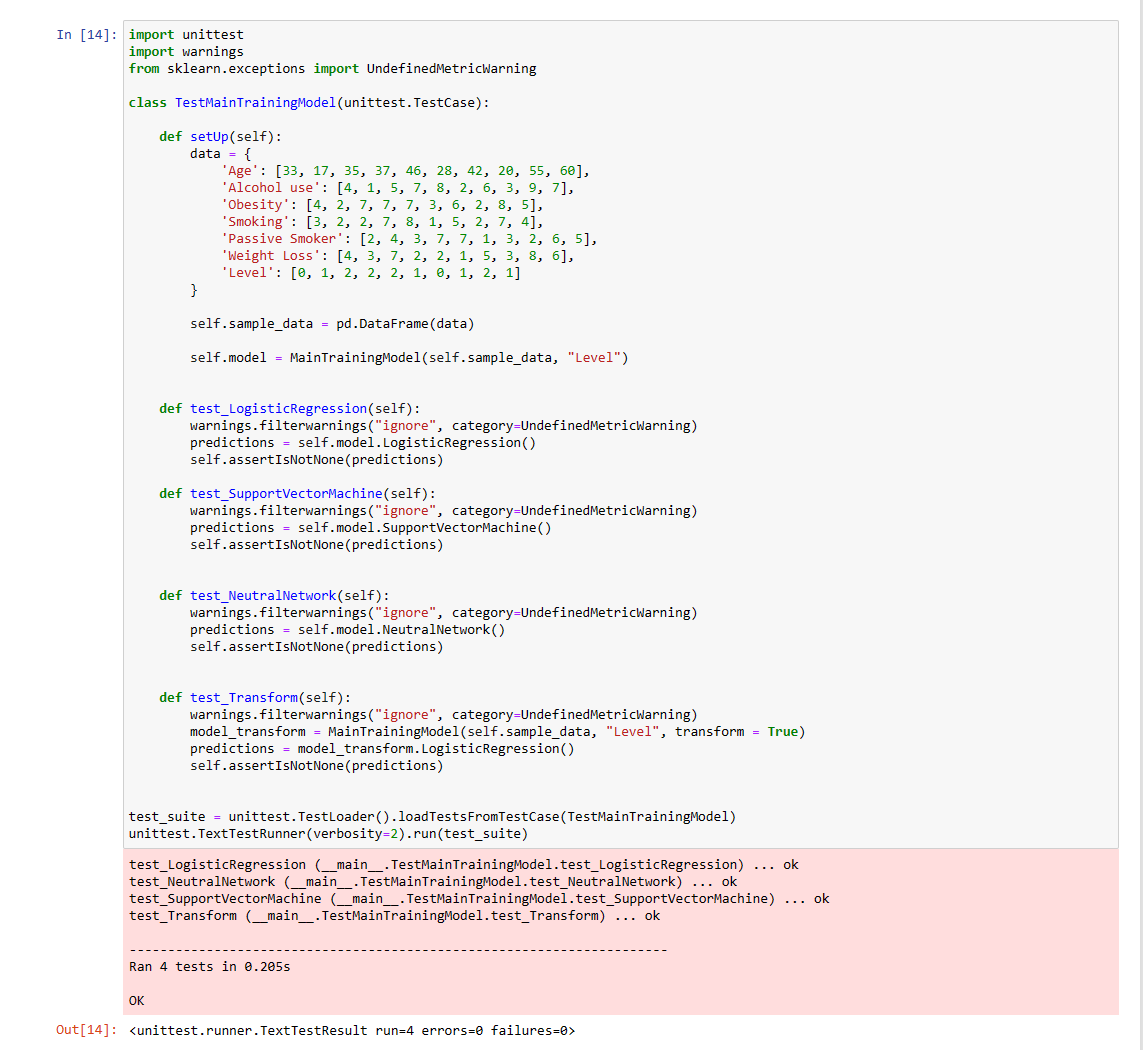


Możliwość porównania wybranych przez nas kolumn (czynników) wpływających na stopień

ryzyka zachorowania na raka. Dając nam przy tym możliwość weryfikacji każdego z

osobna, z naszymi testami jednostkowymi.

1. Testy jednostkowe naszych trenowanych modeli na podstawie przykładowych danych nie wykorzystanych podczas procesu trenowania tj. 20% z całej puli dostępnych danych:

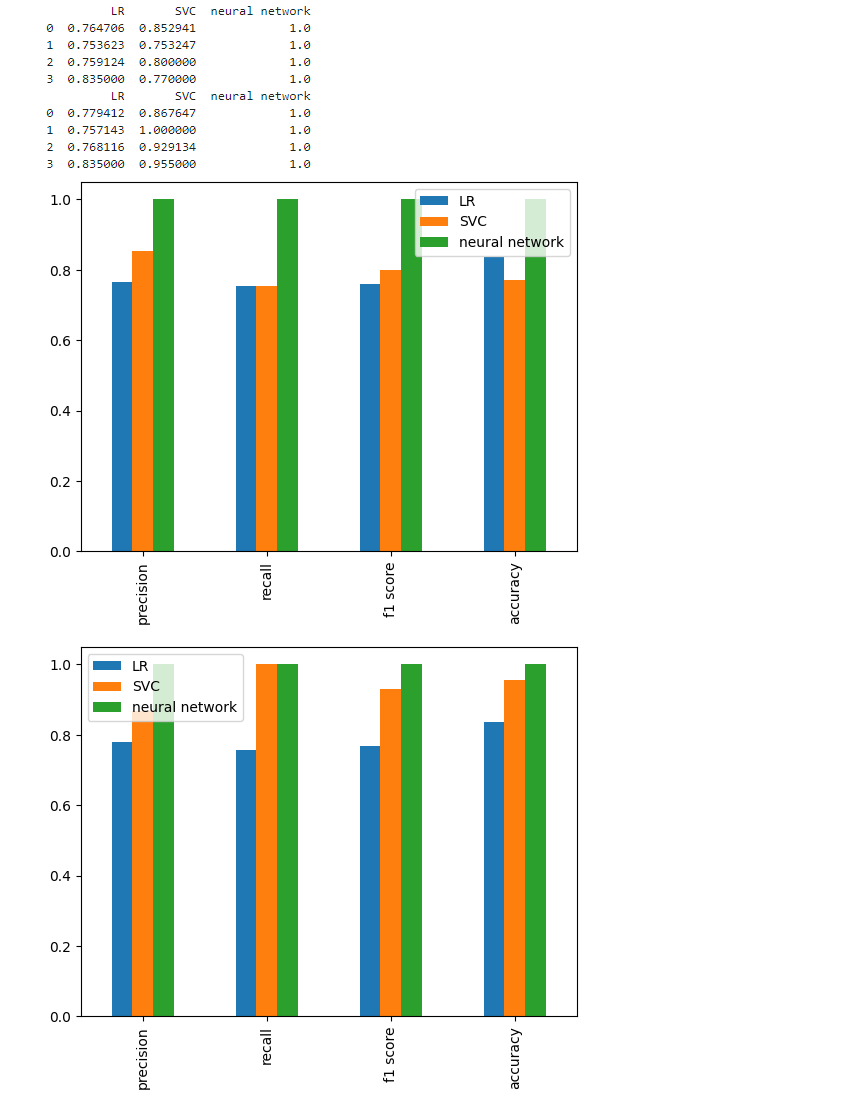


Analiza wyników.

Obraz zawierający zrzut ekranu, tekst, kwadrat, Prostokąt

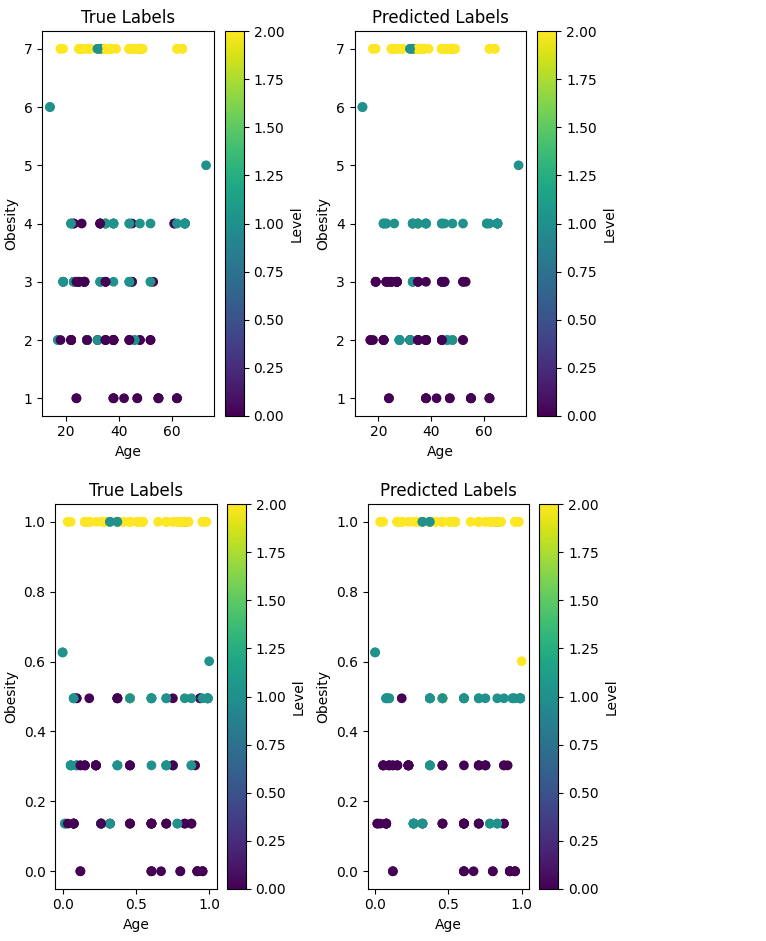
Opis wygenerowany automatycznie

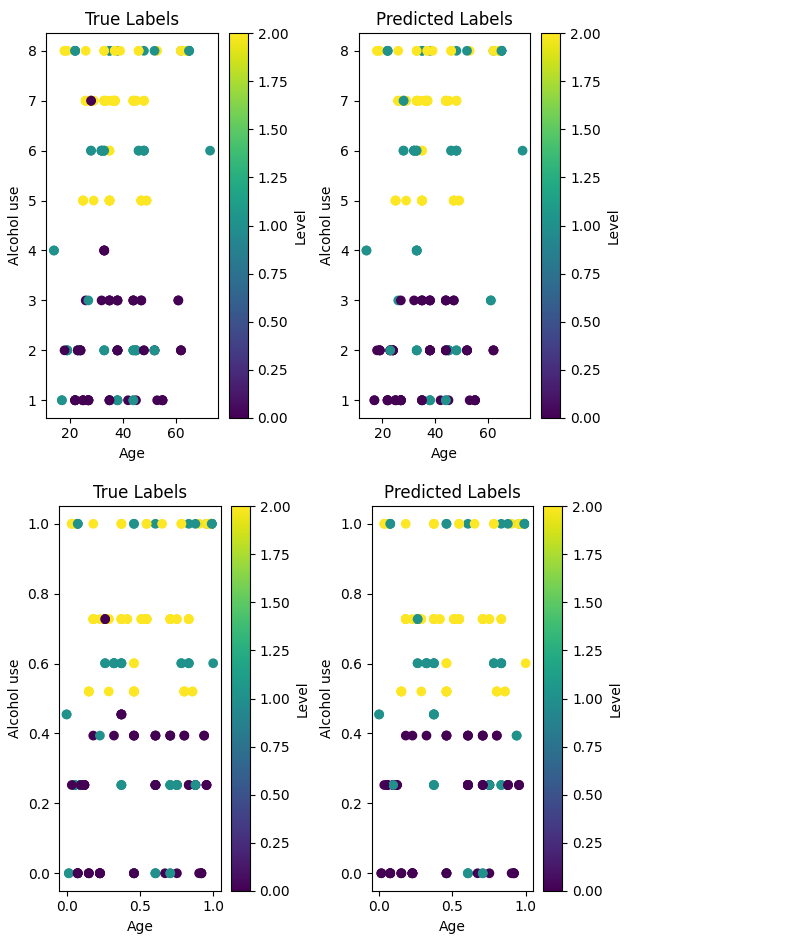
Macierz naszego modelu wskazuje, że najsilniejsza korelacja z poziomem ryzyka zachorowania na raka płuc występuje z otyłością. Dalej nadal bardzo silna korelacja występuje z ilością spożywanego alkoholu(im więcej tym szansa jest większa) i bycia biernym palaczem. Kolejna znacznie słabsza korelacja występuje z samym paleniem co pokazuje, że paląc jesteśmy mniej narażeni na raka płuc, niż osoby niepalące przebywające w tym samym otoczeniu. Korelacja utraty wagi ma o wiele najmniejszy wpływ na ryzyko zachorowania. Dodatkowo po przyjrzeniu się dokładniej korelacją “Weight Loss” możemy powiedzieć, że osoby tracące wagę, były najczęściej niepalącymi, nie były otyłe, nie nadużywały alkoholu oraz przeważnie nie były biernymi palaczami. Ostatnią korelacją jest wiek, który koreluje najsłabiej ze wszystkimi czynnikami.

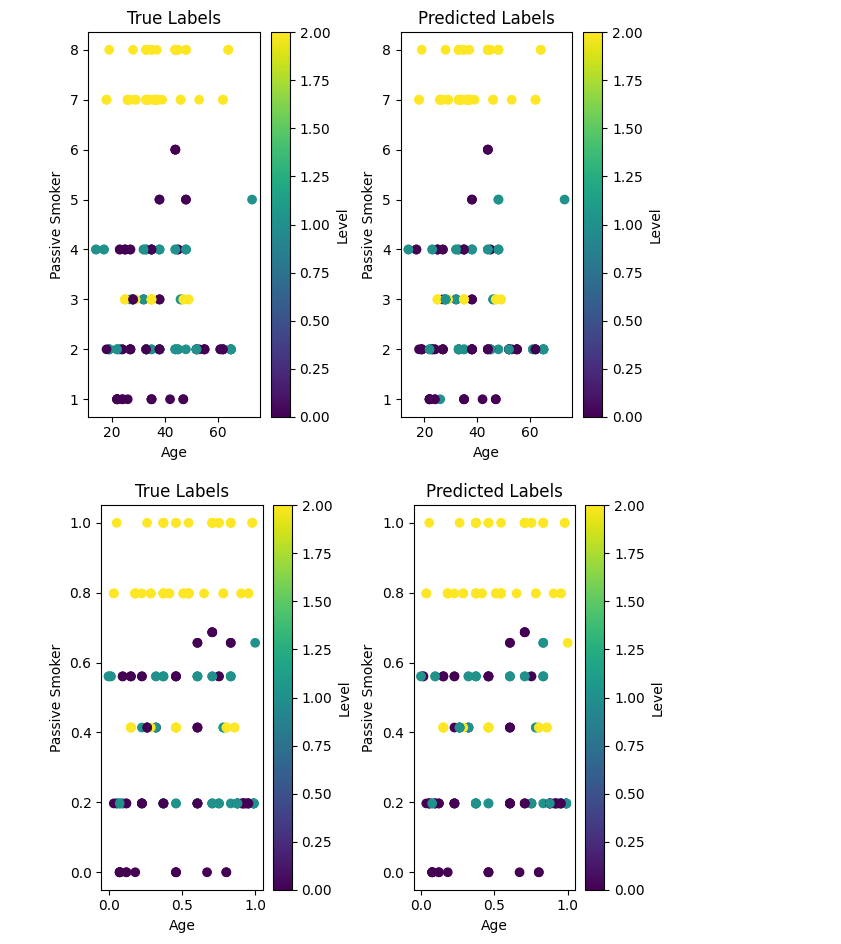


Nasz 1 wykres przedstawia wytrenowany model, a poniżej znajduje się wynik naszych testów jednostkowych. Sieć neuronowa, może wskazywać, że model ten jest przeuczony(nadmiernie dopasowany do danych treningowych), co może prowadzić do problemów podczas dostarczania nowych nieznanych danych. Kiedy jednak spojrzymy do testów jednostkowych okazuje się, że nasza wartość nadal wynosi 1.0.

Wyniki SVC i LR wydają się być bardziej realistyczne, lecz przedstawiają mniej precyzyjne wyniki. Nadają się bardziej do praktycznego zastosowania.







Wyżej znajdziemy porównanie 3 kolumn z najmocniejszą korelacją z "Level" naszego modelu z testami jednostkowymi. Kolejno: otyłość, korzystanie z alkoholu, bierne palenie. Dla otyłości możemy zaobserwować nadmierne przypisywanie osób średnio otyłych do większego ryzyka zachorowania na raka niż jest w, rzeczywistości. Przy bardzo niskim i wysokim poziomie otyłości model radził sobie znacznie lepiej. Przy korzystaniu z alkoholu błędne przewidywania występują w prawie całym zakresie lecz są one pojedyncze, brak zauważalnej tendencji. Bierne palenie przewiduje poprawnie wynik dal osób najmocniej narażonych na takie środowisko a w przypadku pacjentów, u których takie ryzyko występuje w mniejszym stopniu model popełnia nieregularne błędy.