Selekcja Zmiennych w języku Python

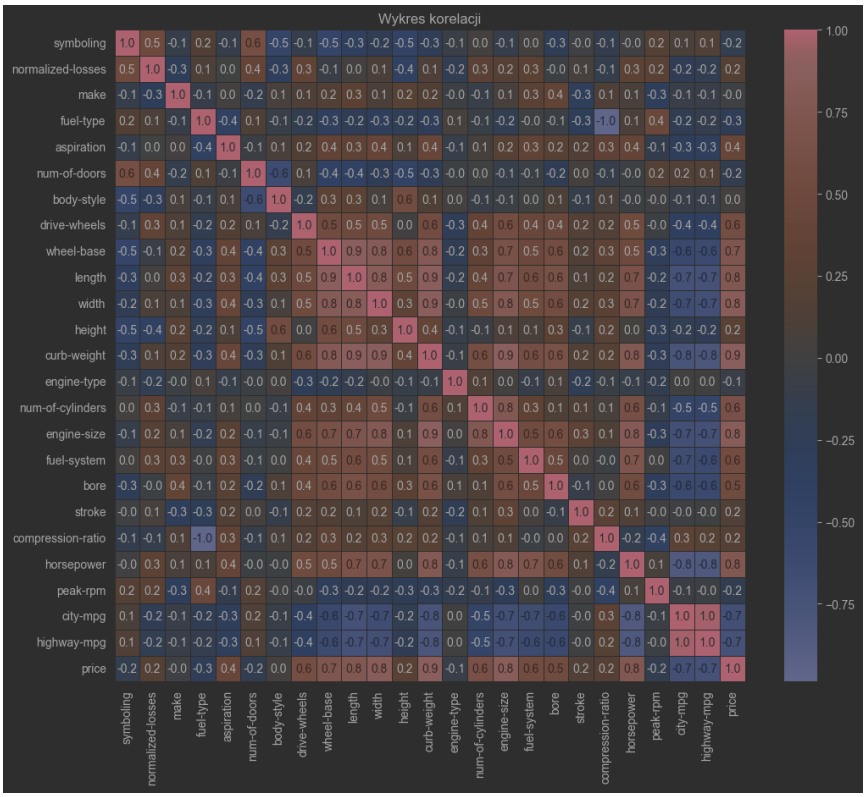
Aleksander Karpiuk 119229, Szymon Kietliński 119233

1. Opis zbiorów danych

Imports-85

Zbiór danych imports-85 opisuje auta z 1985 roku. Zawiera specyfikację dotyczącą cech charakterystycznych dla samochodów, wskaźnik ubezpieczenia takiego samochodu (odpowiada za to zmienna symobling i jest ona zmienną liczbową przyjmującą wartości <-3,3>, opisuje ryzyko związane z danym samochodem względem jego ceny), znormalizowaną stratę która służy do porównania wśród innych samochodów, odnosi się do średniej płatności za ubezpieczony samochód w ciągu roku. Zbiór posiada 26 cech przyjmujących wartości zarówno tekstowe jak i liczbowe.

W celu normalizacji danych w tym zbiorze zostały zamienione zmienne tekstowe na odpowiednie liczby. Rekordy z wartościami brakującymi zostały usunięte.



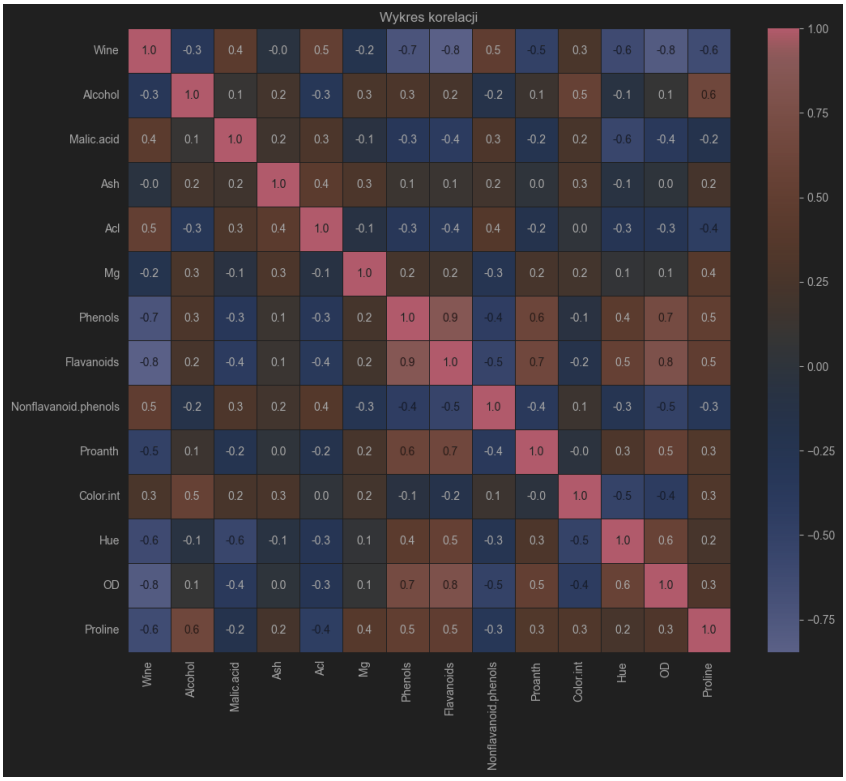
Wybór zmiennych dla tego zbioru danych został wykonany względem zmiennej *price*.

Wine

Zbiór danych Wine opisuje chemiczne parametry 178 win. Celem zbioru jest zmienna *wine* która odpowiada trzem różnym kategoriom win. Zbiór posiada 13 zmiennych, z czego wszystkie są numeryczne (całkowite oraz rzeczywiste). W zbiorze nie ma brakujących wartości.

Jest to najprostszy spośród użytych w tej pracy zbiorów.

Wartość *wine* jest silnie skorelowana ze zmiennymi, co było widać podczas selekcji zmiennych, tzn. nie udało się ich wiele wyeliminować.



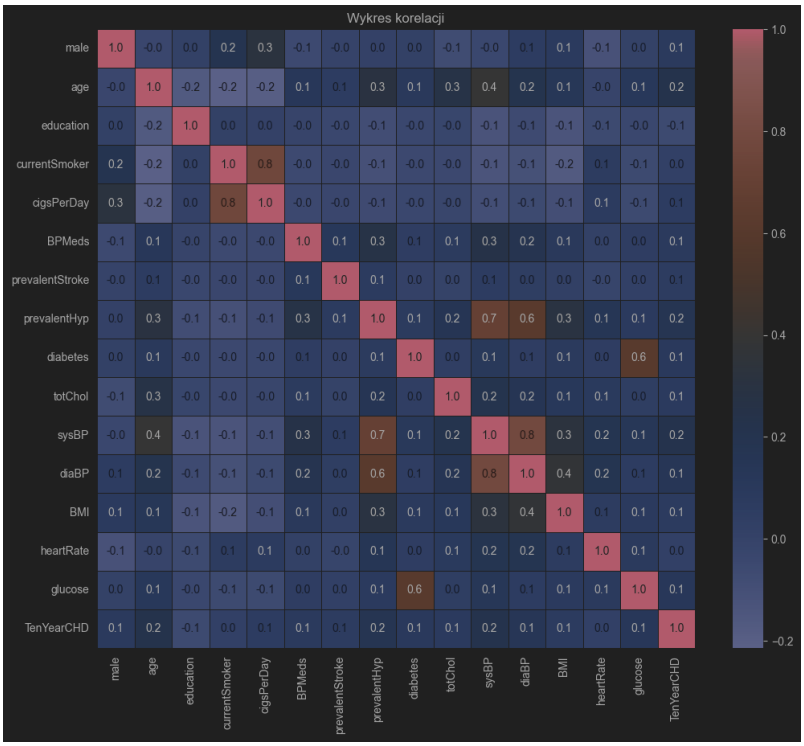
Framingham

Zbiór danych Framingham zajmuje się przewidywaniem na podstawie 15 zmiennych czy pacjent w przeciągu 10 lat ma szansę zachorować na chorobę serca (*TenYearCHD*).

Jest to największy użyty zbiór danych, zawierający 4238 rekordów.

Wszystkie zmienne w zbiorze są typu numerycznego (całkowite oraz rzeczywiste). Zbiór zawiera puste wartości, które zostały uzupełnione medianą danej zmiennej.

Jest to najsłabiej skorelowany zbiór spośród wykorzystanych w tej pracy.



2. Wykorzystane algorytmy

Metoda ręcznej selekcji zmiennych

Metoda polega na utworzeniu modelu względem wybranej kluczowej zmiennej. W kolejnych etapach usuwane są po kolei zmienne z największą wartością P-value, aż do momentu w którym żadna zmienna nie ma wartości P-value > 0.05. Ważne, żeby w każdym etapie była usuwana tylko jedna zmienna. Przy każdym kroku można śledzić czy wartość R-squared oraz Adjusted R-squared nie zmieniają się drastycznie.

Metoda zachłannej minimalizacji w przód oraz w tył

Metoda zachłannej minimalizacji w przód polega na wybieraniu kolejnej zmiennej dającej najlepsze rezultaty w połączeniu z poprzednio wybraną zmienną, proces ten trwa aż do osiągnięcia wprowadzonego kryterium (w naszym przypadku algorytm sam decyduje o najlepszej liczbie zmiennych. Parametr k_features='best'). W przypadku metody w tył, zaczyna się od modelu ze wszystkimi zmiennymi i każdorazowo usuwana jest najgorzej dopasowana zmienna. Metoda ta jest nazywana zachłanną, ponieważ ocenia wszystkie możliwe kombinacje zmiennych, co jest kosztowne obliczeniowo, szczególnie przy wysokiej wartości parametru walidacji krzyżowej. (3)

Metoda PCA

PCA (Principal Component Analysis) to metoda analizująca główne składowe celem wyznaczenia nowych zmiennych, które będą zależały od zmiennych pierwotnych. Kryterium redukcji wymiarów dla modelu zbioru danych to w naszym przypadku wykres osypiska. Pokazuje on zależność wartości własnych od kolejnych nowych zmiennych. Te o niskiej wartości własnej stanowią w głównej mierze szum. Wartość własna mówi o tym, jaka część całkowitej zmienności jest tłumaczona przez daną składową główną. Każda kolejna tłumaczy coraz mniejszy stopień wariancji. (4)

F-test

F-test (1) jest skróconą nazwą Testu Fishera, będącego testem statystycznym służącym do porównywania wariancji dwóch próbek. W naszym przypadku został on wykorzystany do analizy regresji modelu podstawowego oraz tego z zredukowaną liczbą zmiennych. F-test jest testem ANOVA dla dwóch próbek. F-test dostarcza wynik w postaci wartości P-value której interpretacja pozwala na odrzucenie lub nie hipotezy H0. Hipoteza H0 – modele są podobne, hipoteza H1 – istnieje istotna różnica między modelami. (2) W projekcie założyliśmy istotność na poziomie 5%, co oznacza że hipotezę H0 możemy odrzucić gdy uzyskamy p-value < 0.05.

3. Wyniki badań

Wyniki badań przedstawione są w formie poniższej tabeli. Zawiera ona wyniki F-testów w formie P-value względem modelu bazowego.

	Imports-85		Wine		Faringham	
	R-squared	Liczba zmiennych	R-squared	Liczba zmiennych	R-squared	Liczba zmiennych
Bazowy model	0.892	24	0.900	14	0.097	15
Ręczna selekcja	0.881	9	0.897	9	0.095	6
Zachłanna minimalizacja w przód	0.877	7	0.897	9	0.096	7
Zachłanna minimalizacja w tył	0.861	7	0.897	9	0.096	7
PCA	0.865	6	0.921	6	0.337	4

Podsumowanie

W projekcie użyliśmy 3 różnych od siebie zbiorów danych.
Zbiór imports-85 był od początku dobrze dopasowanym modelem, lecz posiadał dużo zmiennych. W jego przypadku najlepiej sprawdziła się ręczna selekcja. Zachłanna minimalizacja w tył oraz PCA były źle dopasowane i w związku z wynikami f-testu modele te zostały odrzucone.
Zbiór Wine był od początku bardzo dobrze dopasowanym modelem, najlepsze rezultaty, za równo pod względem liczby zmiennych, jak i wartości R-squared dała metoda PCA.
Zbiór Faringham na początku był modelem niezadawalającym, w jego przypadku metoda PCA zredukowała liczbę zmiennych ponad trzykrotnie i znacznie podniosła wartość R-squared, niemniej nadal nie był to dobrze dopasowany model.

Na podstawie uzyskanych rezultatów trudno ocenić która metoda selekcji zmiennych jest najlepsza. Każdy zbiór danych należy traktować indywidualnie i dobierać do niego konkretną metodę.

Bibliografia

https://mfiles.pl/pl/index.php/Test_Fishera

https://e.uksw.edu.pl/pluginfile.php/790360/mod_resource/content/0/UMSN_wyklad_2_2023.pdf

https://miroslawmamczur.pl/czym-jest-wybor-zmiennych-feature-selection-16-metod-ktore-warto-znac/

https://e.uksw.edu.pl/mod/resource/view.php?id=497418