



Wydział Matematyki i Nauk Informacyjnych

POLITECHNIKA WARSZAWSKA

Zaawansowane Metody Uczenia Maszynowego

Zastosowanie i Analiza Metod Wyjaśnialnego Uczenia Maszynowego na Zbiorze Adult

Jan Cichowlas

Ziemowit Głowaczewski

Szymon Pawlonka

Alicja Pruszyńska

Krzysztof Stolarski

WARSZAWA, 2025



1 Wstęp

Celem niniejszego projektu jest stworzenie klasyfikatora przewidującego poziom dochodów osób na podstawie danych demograficznych i zawodowych. Głównym kryterium oceny modelu będzie wysoka wartość wskaźników **Accuracy**, **Precision**, **Recall** na zbiorze walidacyjnym.

Kluczowym elementem projektu jest zastosowanie metod wyjaśnialnej sztucznej inteligencji (XAI, ang. *eXplainable AI*). W projekcie zostaną wykorzystane różne klasyfikatory reprezentujące odmienne podejścia modelowania, takie jak:

- naiwny klasyfikator Bayesa,
- Gradient Boosting,
- Random Forest,
- maszyna wektorów nośnych (SVC),
- sieć neuronowa.
- k-najbliższych sąsiadów.

W celu lepszego zrozumienia wpływu cech na wynik klasyfikatora oraz identyfikacji potencjalnych błędów modelu, zastosowane zostaną różne techniki XAI, w szczególności:

- wartości SHAP (ang. *SHapley Additive exPlanations*),
- permutacyjna istotność zmiennych (ang. *permuation feature importance*),
- wykresy ICE (ang. *Individual Conditional Expectation*) oraz wykresy PDP (ang. *Partial Dependence Plots*),
- ALE (ang. *Accumulated Local Effects*).

Dodatkowo, porównane zostaną podejścia PDP i ALE, z uwzględnieniem ich różnic, m.in. tego, że ALE uwzględnia rozkład brzegowy zmiennych niezależnych, co czyni go mniej podatnym na błędne wnioskowanie w przypadku skorelowanych cech.

Końcowym celem projektu jest głównie wyciągnięcie wniosków dotyczących interpretowalności poszczególnych modeli oraz efektywności wybranych technik XAI w kontekście danych tablicowych.



2 Krótki opis wybranych metod

SHAP Przez f oznaczmy funkcję decyzyjną $f : P(X^p) \rightarrow Y$ analizowanego modelu. Metoda SHAP nadaje każdej z uwzględnionych zmiennych x_1, \dots, x_p odpowiednie wartości ϕ_1, \dots, ϕ_p , odpowiadające wpływowi danych zmiennych na predykcję modelu.

W iteracyjny sposób definiujemy:

- $\phi_1 := \mathbb{E}[f(x) | x_1] - \mathbb{E}[f(x)]$
- $\phi_k := \mathbb{E}[f(x) | x_1, \dots, x_k] - \mathbb{E}[f(x) | x_1, \dots, x_{k-1}]$

Poszczególne wartości ϕ_i wskazują nam wpływ danej zmiennej na predykcję. Największą zaletą metody jest jej uniwersalność - SHAP działa dla dowolnego modelu. Należy jednak zwrócić uwagę na dwa problemy:

- Kolejność wyznaczania ϕ_i ma znaczenie dla zmiennych zależnych.
- Proces wyznaczania wartości Shapleya jest złożony obliczeniowo (dla modeli nielinowych wymaga obliczeń permutacyjnych).

SHAP pozwala na określenie globalnych zależności, jak i na sprawdzenie wpływu zmiennych dla pojedynczych obserwacji. W sekcji z wykresami prezentujemy wykresy SHAP dla próby 100 obserwacji (wykresy wiolinowe) oraz dla jednej ustalonej obserwacji - wykresy prezentujące poszczególne ϕ_i .

Permutacyjna istotność zmiennych to metoda oceny wpływu poszczególnych zmiennych objaśniających na jakość predykcji modelu. Polega ona na losowym permutowaniu wartości wybranej zmiennej w zbiorze testowym, przy jednoczesnym pozostawieniu pozostałych zmiennych bez zmian. Następnie oblicza się spadek jakości modelu, np. wzrost błędu (MSE, MAE) lub spadek miary dopasowania (R^2 , accuracy).

Jeżeli permutacja danej zmiennej powoduje wyraźne pogorszenie wyników modelu, oznacza to, że zmienna ta odgrywa istotną rolę w procesie predykcji. Z kolei niewielka zmiana w jakości modelu po permutacji sugeruje, że zmienna może być mniej istotna lub nadmiarowa (np. silnie skorelowana z inną zmienną).

Zaletą tej metody jest jej uniwersalność – może być stosowana do dowolnego modelu predykcyjnego, niezależnie od jego wewnętrznej struktury. Ponadto jest łatwa do interpretacji i pozwala w przejrzysty sposób porównać wpływ różnych zmiennych na wynik modelu.



PDP Niech f będzie funkcją decyzyjną $P(X^p) \rightarrow Y$ modelowaną przez \hat{f} . Oznaczmy przez $S \subset \{1, \dots, p\}$ zbiór wybranych cech oraz przez C jego zbiór dopełniający. Przez \mathbf{x}_S rozumiemy obserwacje powstałe z $\mathbf{x} \in X^p$ poprzez wybór cech S . Wówczas krzywą PDP określamy wzorem

$f_S = \mathbb{E}_{\mathbf{x}_C} [f(\mathbf{x}_S, \mathbf{x}_C)] = \int f(\mathbf{x}_S, \mathbf{x}_C) dP(\mathbf{x}_C)$, gdzie \mathbf{x}_S jest ustalone, a \mathbf{x}_C przyjmuje wartości z jego rozkładu brzegowego $dP(\mathbf{x}_C)$. Nie znamy f i $dP(\mathbf{x}_C)$, dlatego estymujemy to wyrażenie za pomocą średniej

$$\hat{f}_S = \frac{1}{N} \sum_{i=1}^N \hat{f}(\mathbf{x}_S, \mathbf{x}_{Ci}),$$

gdzie X_{Ci} to próbki ze zbioru treningowego z wybranymi cechami C .

ICE Wada metody PDP leży właśnie w tym uśrednieniu, które stosuje - przeciwne interakcje mogą się ze sobą znosić. Po więcej szczegółów odsyłamy do artykułu źródłowego^[1]. To stanowiło punkt wyjściowy do opracowania metody ICE. Rozważmy zbiór wszystkich próbek $\{(x_{Si}, \mathbf{x}_{Ci})\}_{i=1}^N$. Krzywa ICE $\hat{f}_S^{(i)}$ przyjmuje wartości dla ustalonego \mathbf{x}_C przy zmiennym \mathbf{x}_S osiągającym wszystkie wartości ze zbioru treningowego. Ponieważ patrzymy na każdą realizację \mathbf{x}_C z osobna, w metodzie uwzględniamy nawet te lokalne interakcje.

ALE to technika mająca na celu analizę i wizualizację wpływu poszczególnych cech wejściowych na predykcję modelu. Jest niejako ulepszeniem metody PDP, która radzi sobie nie najlepiej wtedy, gdy cechy są skorelowane.

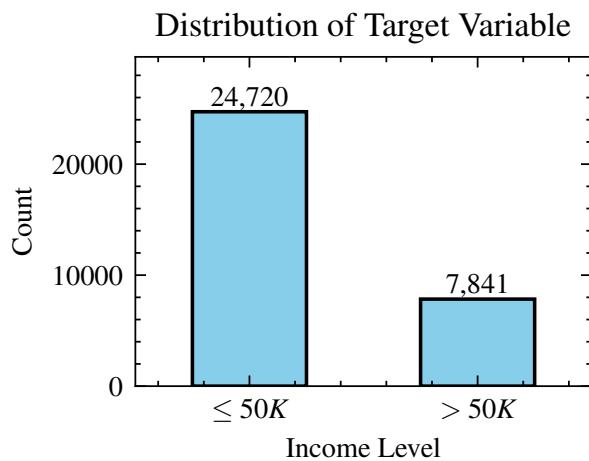
Załóżmy, że dysponujemy modelem predykcyjnym f i chcemy oszacować wpływ zmiennej x_j na predykcję. Niech x_{-j} oznacza wektor pozostałych zmiennych. W metodzie ALE dzielimy najpierw dziedzinę wartości, jakie x_j może przyjmować, na K przedziałów $[z_k, z_{k+1}]$, gdzie $k = 1, 2, \dots, K$. Wówczas definiujemy efekt lokalny

$$\Delta f_j^{(k)} = \mathbb{E}_{x_{-j}|x_j \in [z_k, z_{k+1}]} [f(z_{k+1}, x_{-j}) - f(z_k, x_{-j})].$$

Efekt lokalny interpretowany jest jako średnia zmiana predykcji, gdy zmieniamy x_j z z_k do z_{k+1} , pozostawiając pozostałe cechy na wartościach występujących w danym zakresie $x_j \in [z_k, z_{k+1}]$.

Efekt skumulowany to suma efektów lokalnych z wszystkich przedziałów:

$$\hat{f}_{ALE_j} = \sum_{l=1}^K \Delta f_j^{(l)}.$$



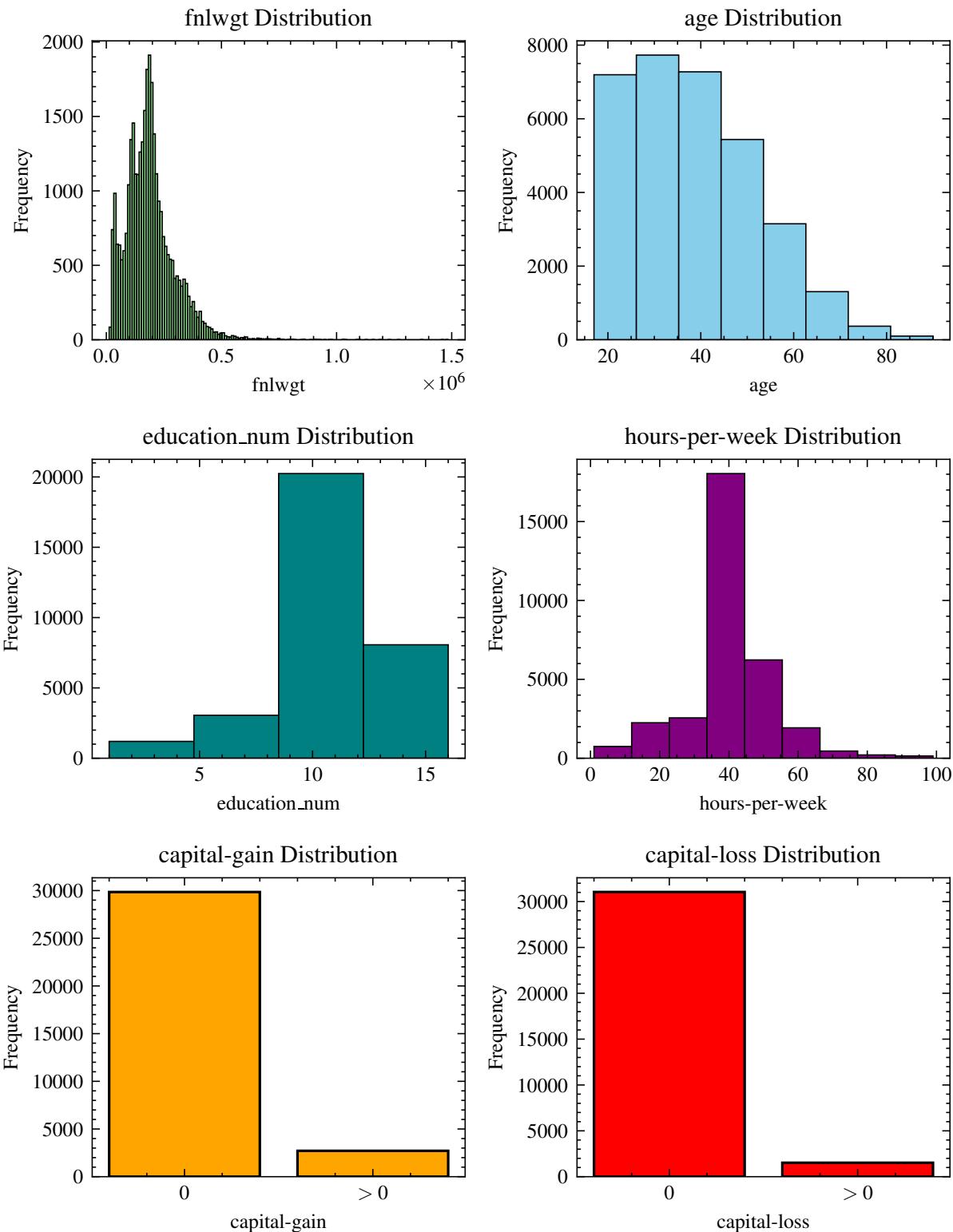
Rysunek 1: **Rozkład zmiennej objaśnianej.** Jak obrazuje wykres, zbiór nie jest zbalansowany.

3 Zbiór danych

Do analizy metod wyjaśnialnego uczenia maszynowego użyto zbioru Adult udostępnionego na UC Irvine^[2]. Zawiera on niespełna 50 tysięcy instancji opisanych przez 14 zmiennych objaśniających (m.in. dochód, poziom edukacji, stan cywilny) i jedną binarną zmienną objaśnianą określającą, czy dochód danej osoby jest większy od 50 000 dolarów.

Wykres 1 przedstawia rozkład zmiennej objaśnianej. Zbiór jest niezbalansowany. Obserwacji, dla których poziom zarobków jest niższy od 50 000 dolarów jest ponad trzy razy więcej niż obserwacji o większym poziomie dochodów. Ponadto znaczna ilość obserwacji pochodzi od osób, które nie zanotowały wzrostu lub straty kapitału (Rysunek 2. Dane pochodzą od osób młodych i w średnim wieku, dobrze wykształconych. Zdecydowana większość pracuje na pełen etat.

Podczas preprocessingu zakodowano dane kategoryzowane metodą 'one hot'. Brakujące dane uzupełniono za pomocą drzewa decyzyjnego. Do treningu nie użyto zmiennej *education*, ponieważ powielała zmienną *education_num* oraz *fnlwgt*, bo stanowiła miarę reprezentatywności danej próbki.



Rysunek 2: **Rozkład wybranych zmiennych objaśniających.** Jest to sześć zmiennych numerycznych, które miały największe w dopasowanym modelu Random Forest.



Tabela 1: Wyniki wybranych klasyfikatorów na zbiorze testowym przy zastosowaniu domyślnych parametrów.

Model	Accuracy	Precision	Recall
GradientBoosting	0.87	0.83	0.78
MLP	0.85	0.80	0.78
RandomForest	0.85	0.80	0.77
KNN	0.84	0.79	0.78
GaussianNB	0.82	0.76	0.80
SVC	0.80	0.77	0.62

4 Eksperymenty

4.1 Wyniki Modeli

Podczas wszystkich eksperymentów ziarno generatora liczb losowych było ustawione na 42. Tabela 1 przedstawia wyniki uzyskane przez wybrane modele dla domyślnych parametrów przy 5-krotnej walidacji krzyżowej. Dodatkowo przy MLP skorzystano ze skalowania min-max, co pozwoliło na znaczną poprawę wyników - wzrost accuracy o 6 punktów procentowych oraz wzrost TPR z 0.13 do 0.61. W przypadku SVC i RandomForest wyniki pokrywają się z punktami odniesienia umieszczonymi na stronie ze zbiorem danych [2]. W przypadku MLP udało się uzyskać wyższy wynik dzięki zastosowaniu skalowania cech numerycznych.



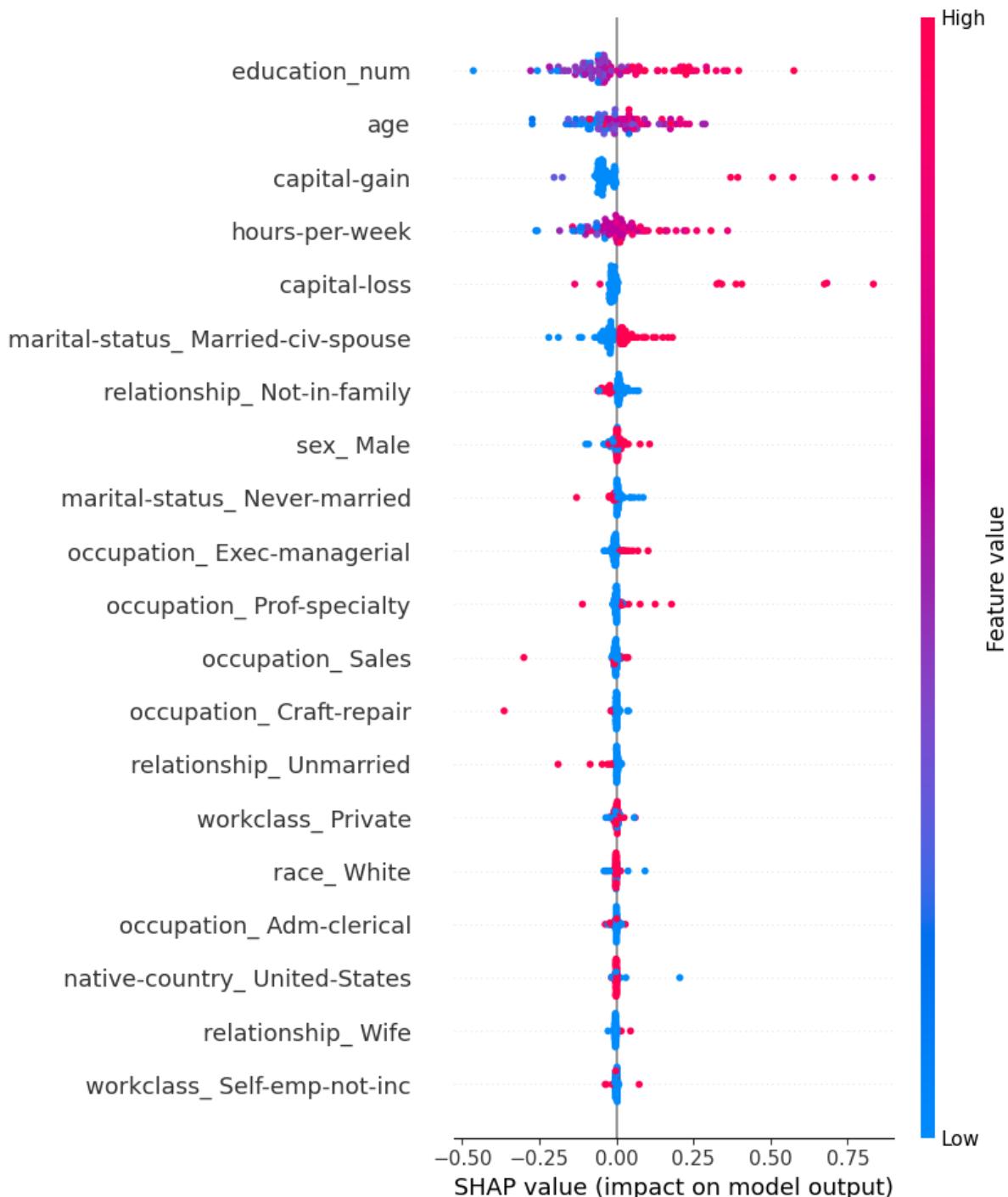
4.2 SHAP

Na powyższych wykresach przedstawiamy wartości Shapley'a dla wybranych modeli. Wykresy 3 - 8 przedstawiają globalny wpływ zmiennych o największej mocy dyskryminującej (pod względem ϕ) dla poszczególnych modeli. Jak widzimy, niezależnie od modelu istotny wpływ na predykcję miała zmienna *capital-gain*, często decydująca była również *capital-loss*. SHAP wykazuje swój potencjał w selekcji zmiennych, pozwalając na uchwycenie globalnych zależności, przy jednoczesnym zaprezentowaniu wartości obserwacja-poobserwacji. Szczególnie interesującym wykresem jest 8, na którym bardzo wyraźnie widać granicę dyskryminacyjną modelu SVC.

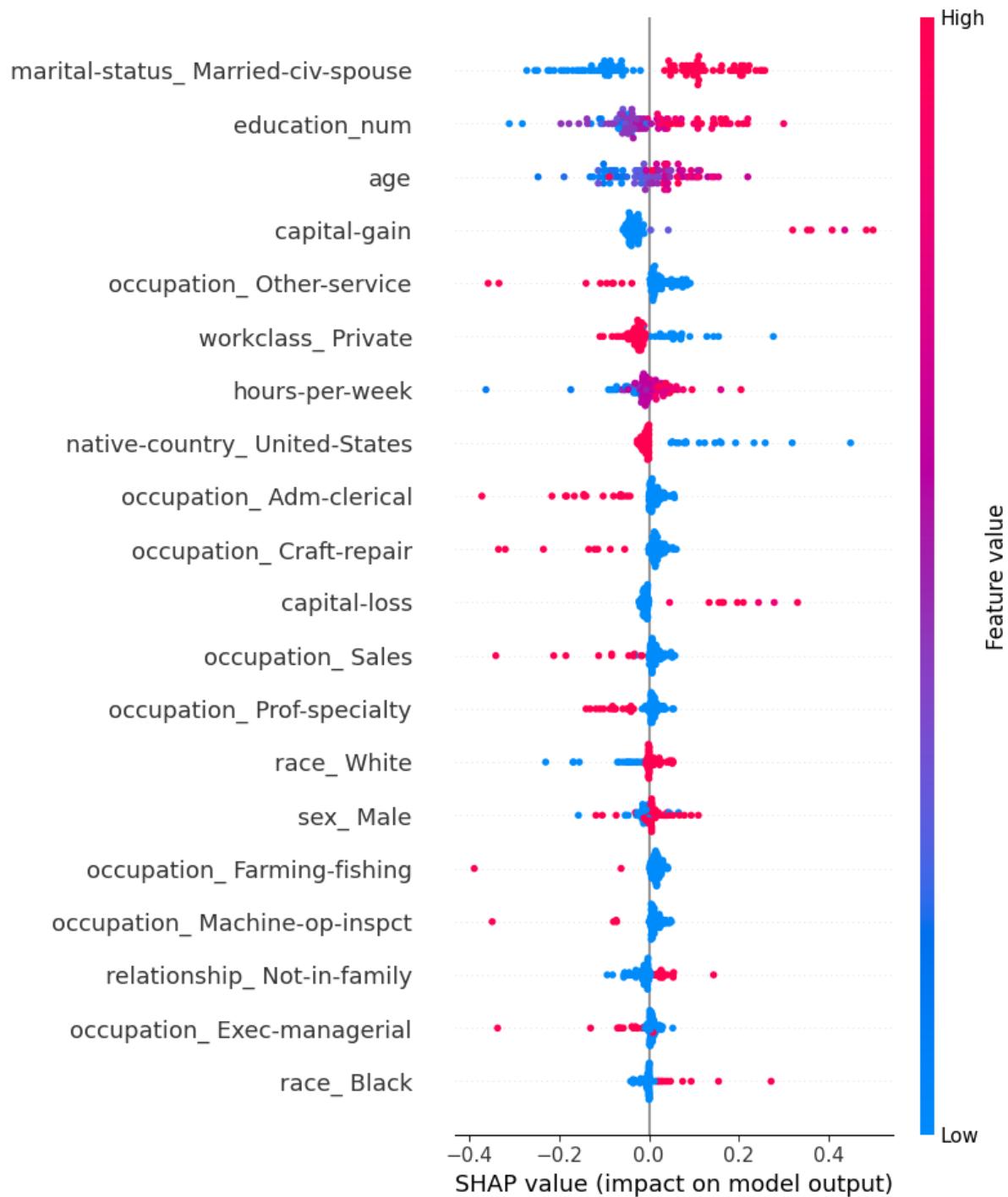
Na wykresach 9-14 przedstawione zostały wykresy wpływu wartości poszczególnych zmiennych na predykcję dla ustalonej obserwacji ze zbioru testowego (ta sama obserwacja na każdym wykresie). W zadanym przykładzie dominują zmienne związane z dostępnym kapitałem oraz wykształceniem. Modele są "zgodne" co do doboru klasy (klasy zero), za wyjątkiem metody Gaussian Naive Bayes - duża wartość *capital-loss* oraz wartość *workclass-Self-emp-inc* zdecydowały o przydzieleniu obserwacji do klasy pierwszej.



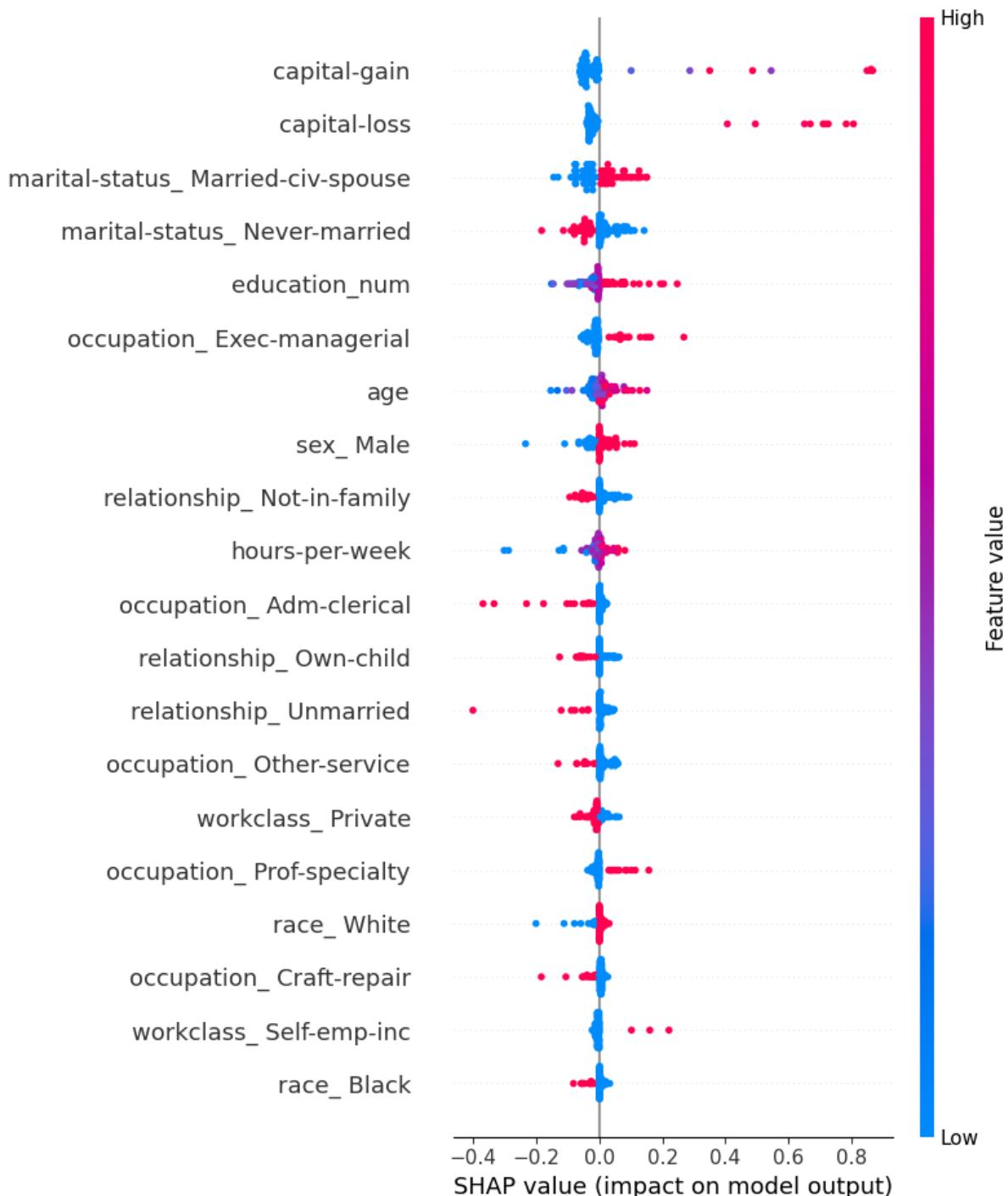
Rysunek 3: SHAP dla metody Gradient Boosting



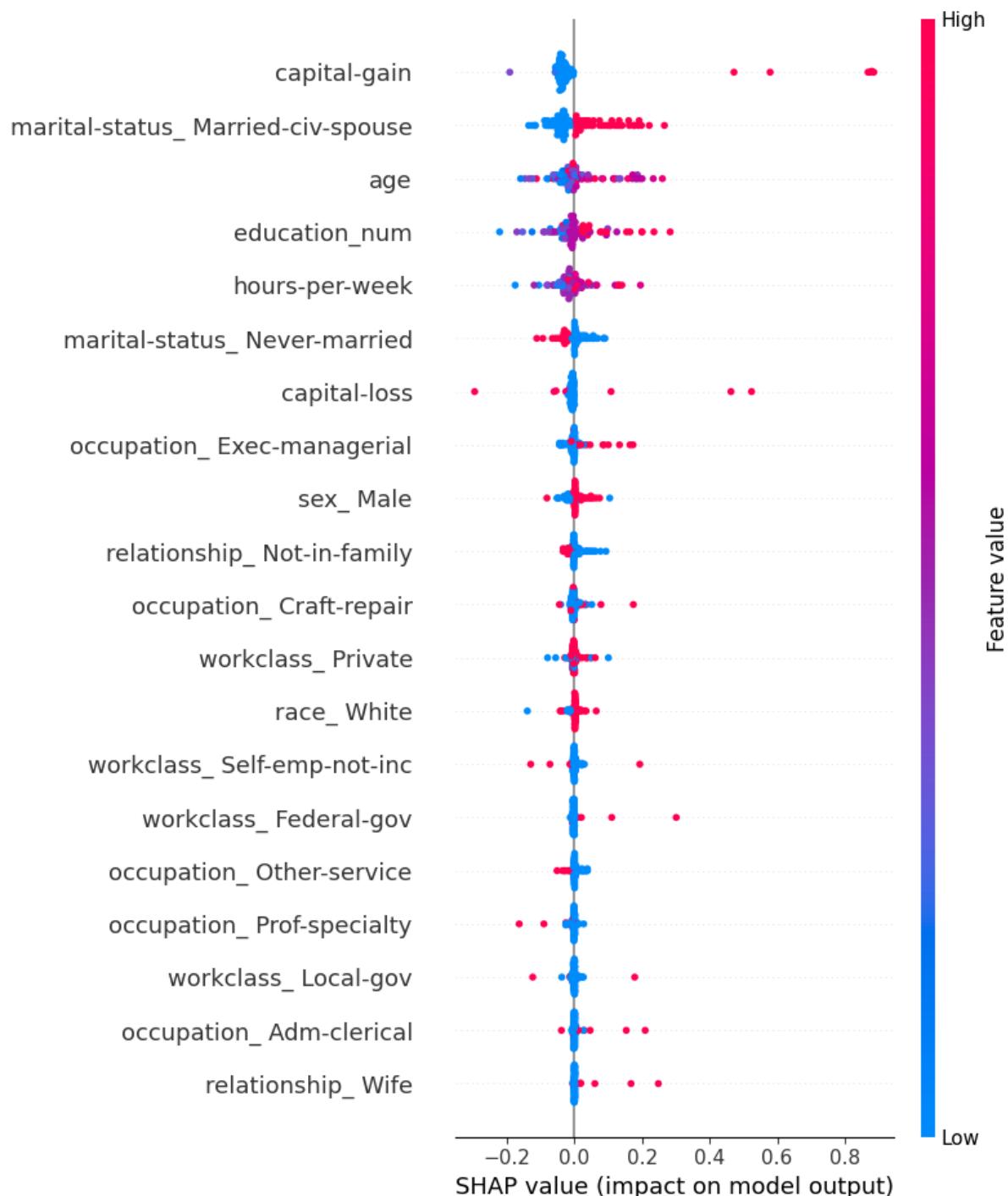
Rysunek 4: SHAP dla KNN



Rysunek 5: SHAP dla prostej sieci neuronowej



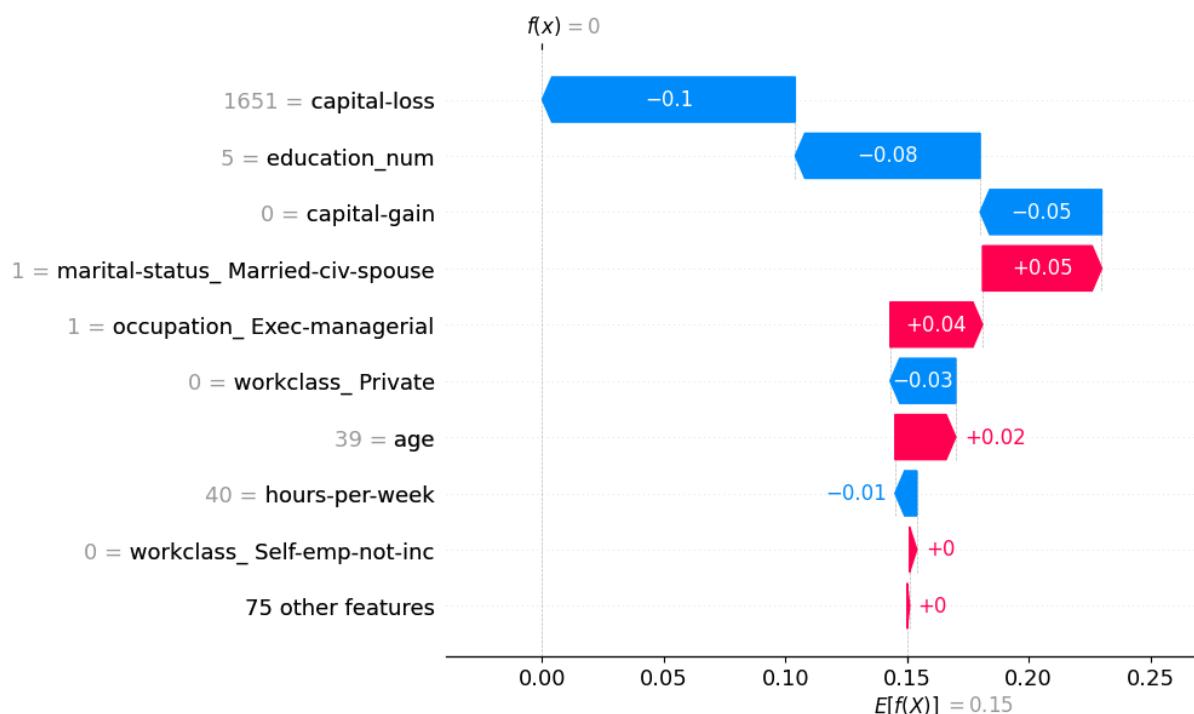
Rysunek 6: SHAP dla metody Gaussian Naive Bayes



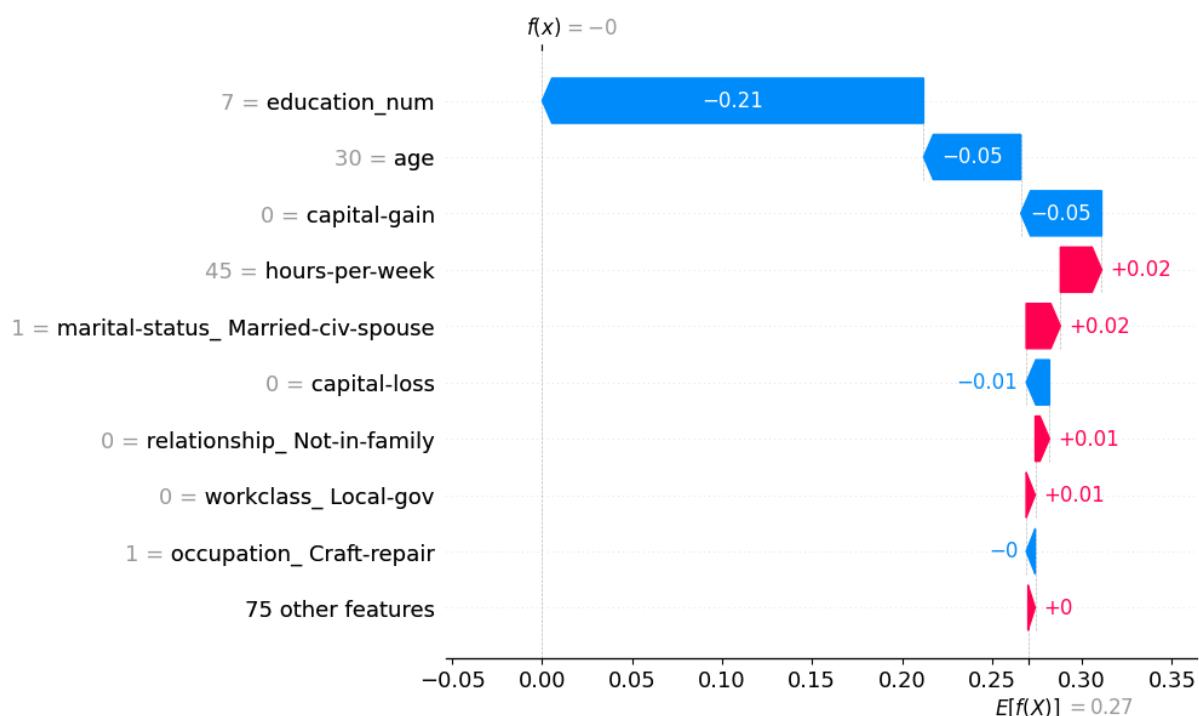
Rysunek 7: SHAP dla metody Random Forest



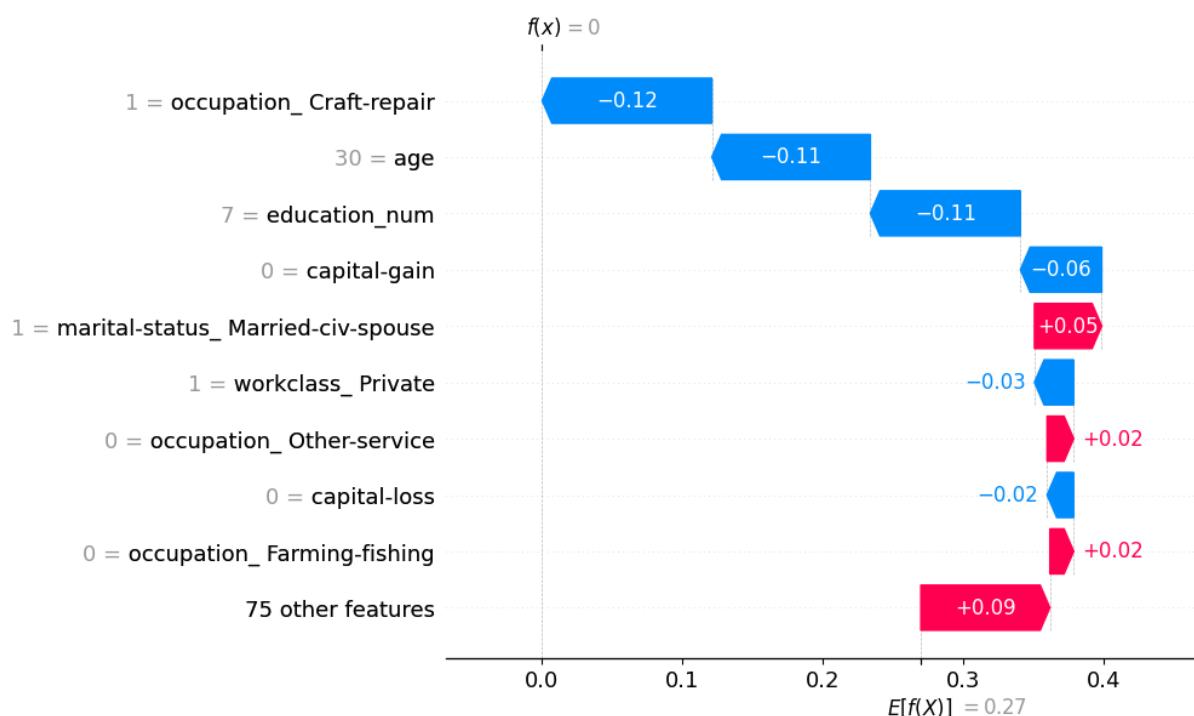
Rysunek 8: SHAP dla metody SVC



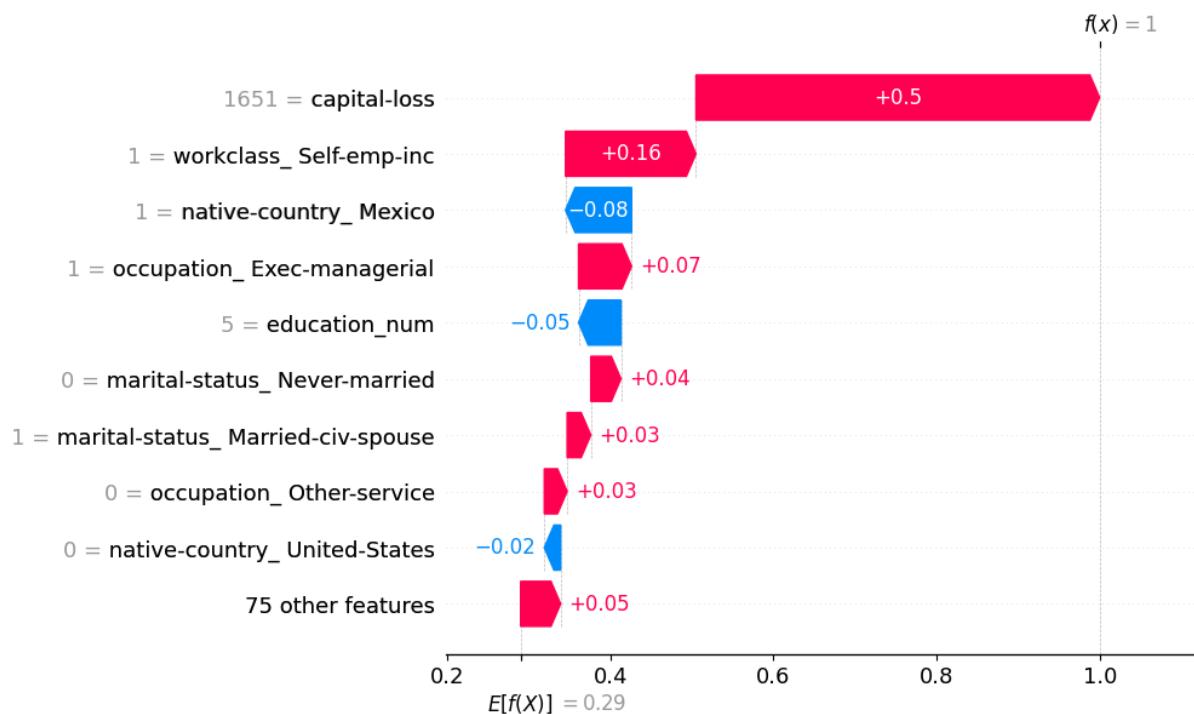
Rysunek 9: Wykres wartości SHAP modelu Gradient Boosting dla przykładowej obserwacji



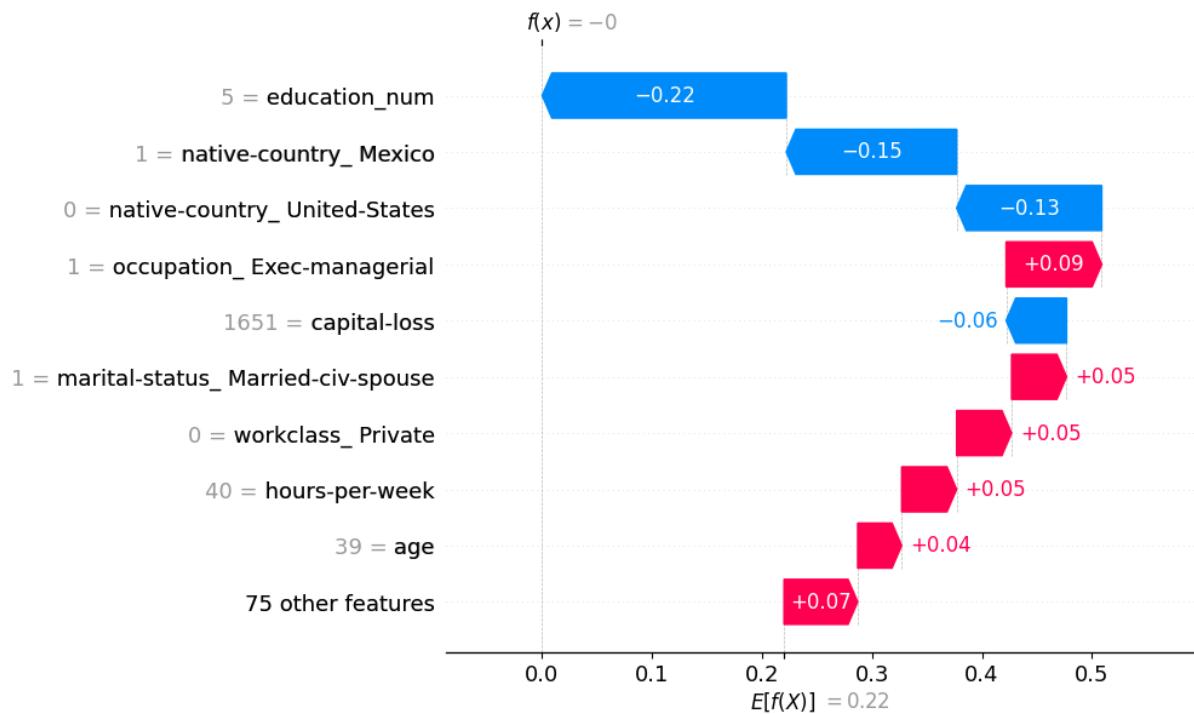
Rysunek 10: Wykres wartości SHAP modelu KNN dla przykładowej obserwacji



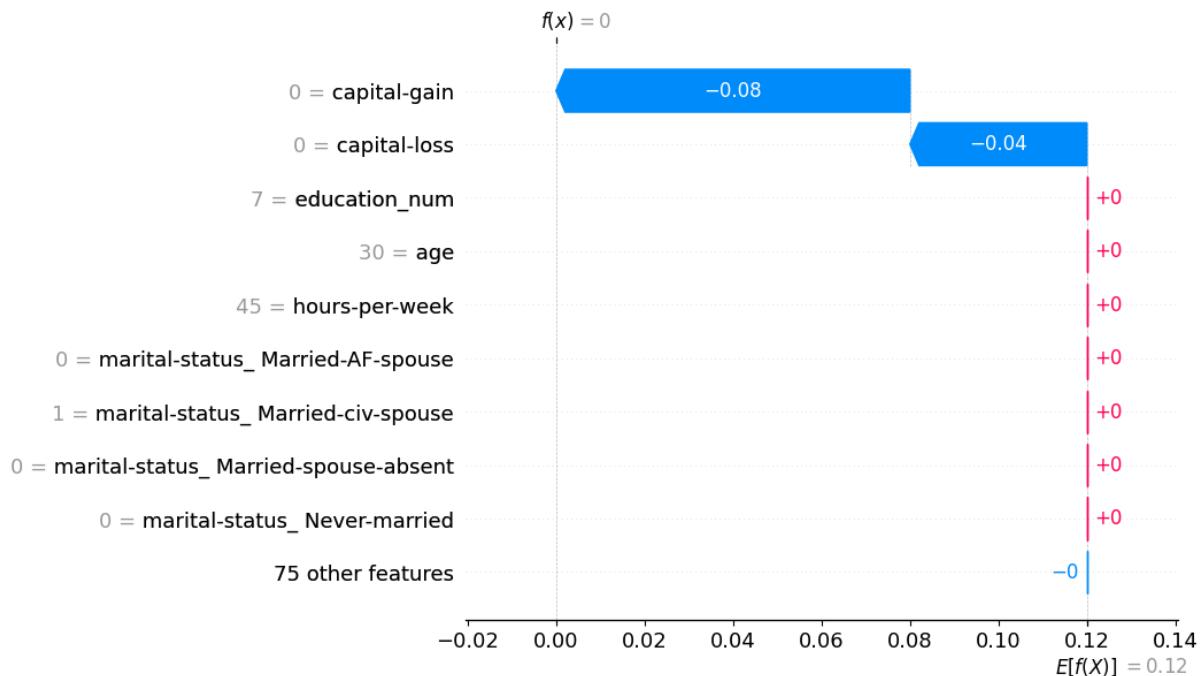
Rysunek 11: Wykres wartości SHAP prostej sieci neuronowej dla przykładowej obserwacji



Rysunek 12: Wykres wartości SHAP modelu Gaussian Naive Bayes dla przykładowej obserwacji



Rysunek 13: Wykres wartości SHAP modelu Random Forest dla przykładowej obserwacji



Rysunek 14: Wykres wartości SHAP modelu SVC dla przykładowej obserwacji



4.3 ICE i PDP

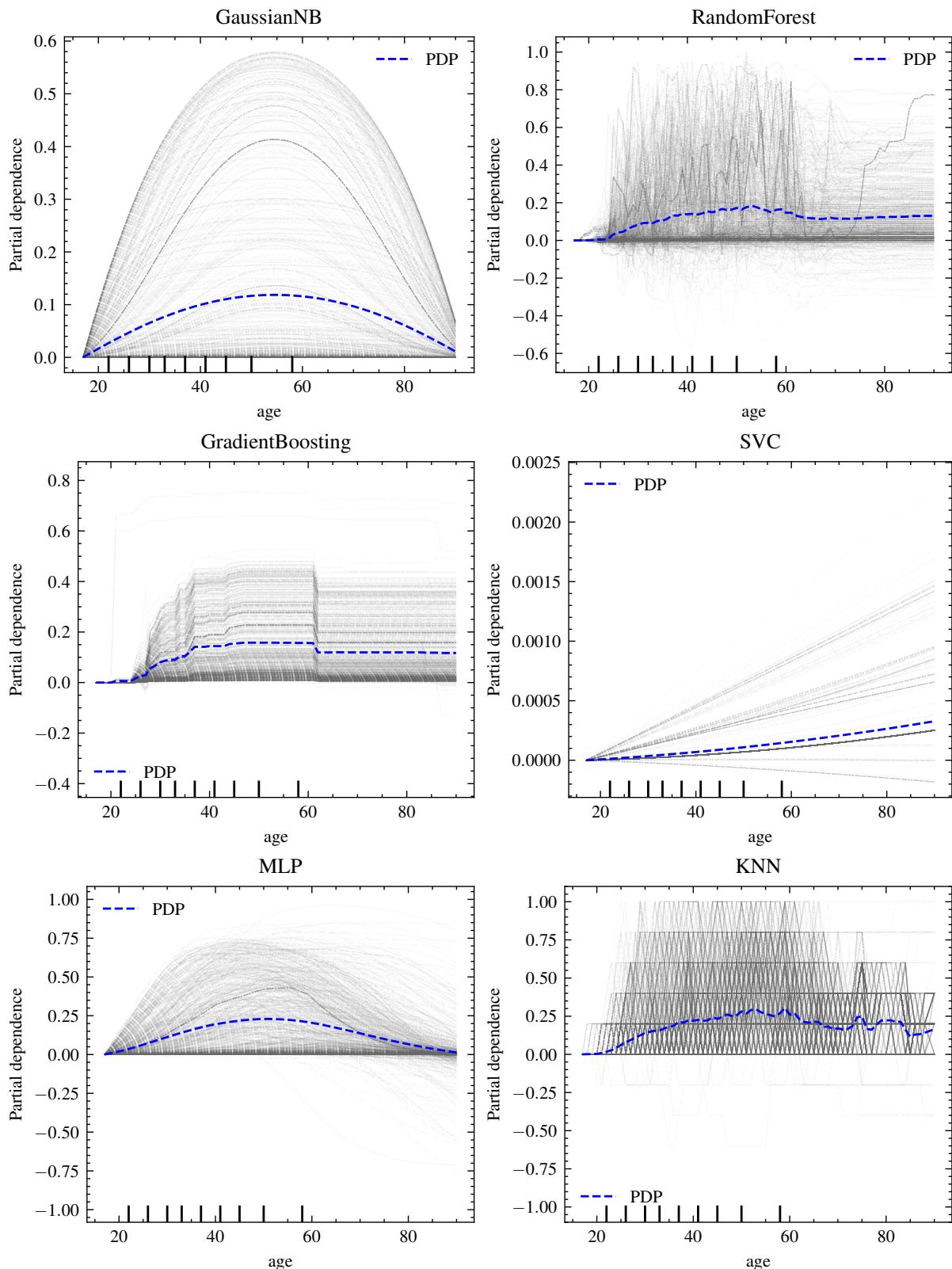
Krzywe ICE dla poszczególnych cech zostały zaprezentowane na Rysunkach 15 - 19. W przypadku zmiennej *age* dla większości modeli istotność zmiennej osiąga maksimum między 40 a 60 rokiem życia, a następnie maleje. Odstaje jedynie model SVM, dla którego istotność nieprzerwanie rośnie. Warto zaznaczyć, że w przypadku modelu RandomForest widzimy zestaw linii, który wyróżnia się szybkim wzrostem w okolicach 70 roku życia.

Wykresy dla zmiennej *capital-gain* nie są już tak jednoznaczne. W szczególności dla modeli GaussianNB oraz MLP krzywe pokazują rosnący wpływ zmiennej wraz ze wzrostem przyjmowanych przez nią wartości. Dla SVC do 2000 mamy predykcję przechyloną w stronę klasy 0, po czym następuje wzrost w kierunku klasy 1. Dla modeli RandomForest, GradientBoosting oraz KNN występują lokalne piki w okolicach wartości 3000, 4200 i 5000.

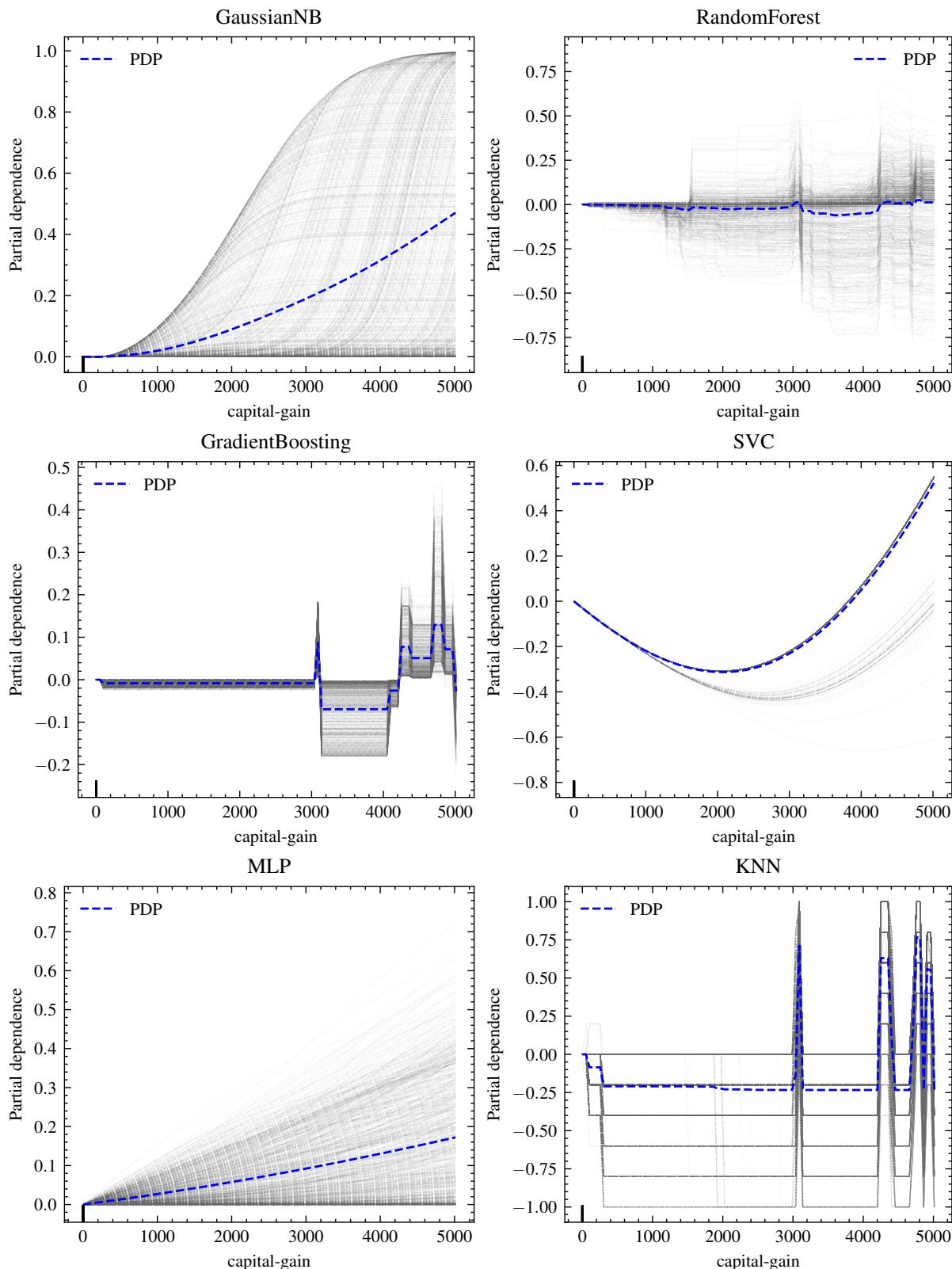
Przechodząc do *capital-loss*, dla tej zmiennej wszystkie wykresy ICE pokazują zwiększający się wpływ na predykcję klasy 1 przy zwiększaniu wartości zmiennej. W przypadku RandomForest, GradientBoosting oraz KNN wykres jest nieco bardziej poszarpany, z lokalnym pikiem przy 2000. Dodatkowo dla KNN widzimy zmianę wpływu na predykcję dla dużych wartości, co odbiega względem pozostałych modeli.

W przypadku edukacji dla wszystkich modeli można zauważać tendencję wzrostową, jednakże tylko dla SVC jest ona liniowa. Wykres ICE dla RandomForest pokazuje, że szczególnie istotne są dla niego *education_num* powyżej 12.5, ponieważ przy tej wartości widoczny jest nagły skok wartości dużej liczby krzywych.

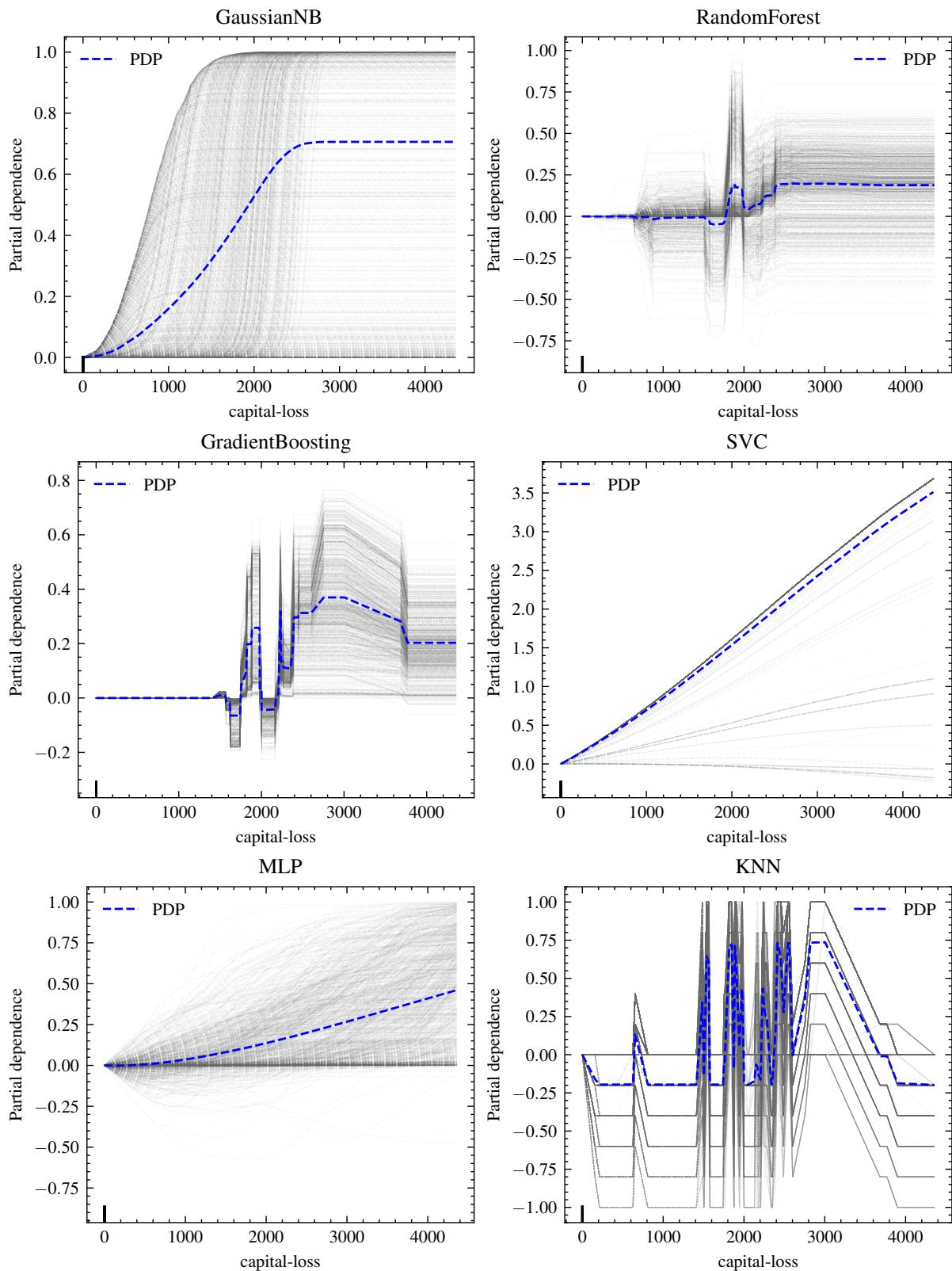
Wykresy ICE dla zmiennej *hours-per-week* pokazują, że wszystkie modele oprócz GradientBoosting i SVC rozpoznały spadek istotności dla dużych wartości tej zmiennej, co jest znacznie lepiej widoczne niż dla krzywej PDP.



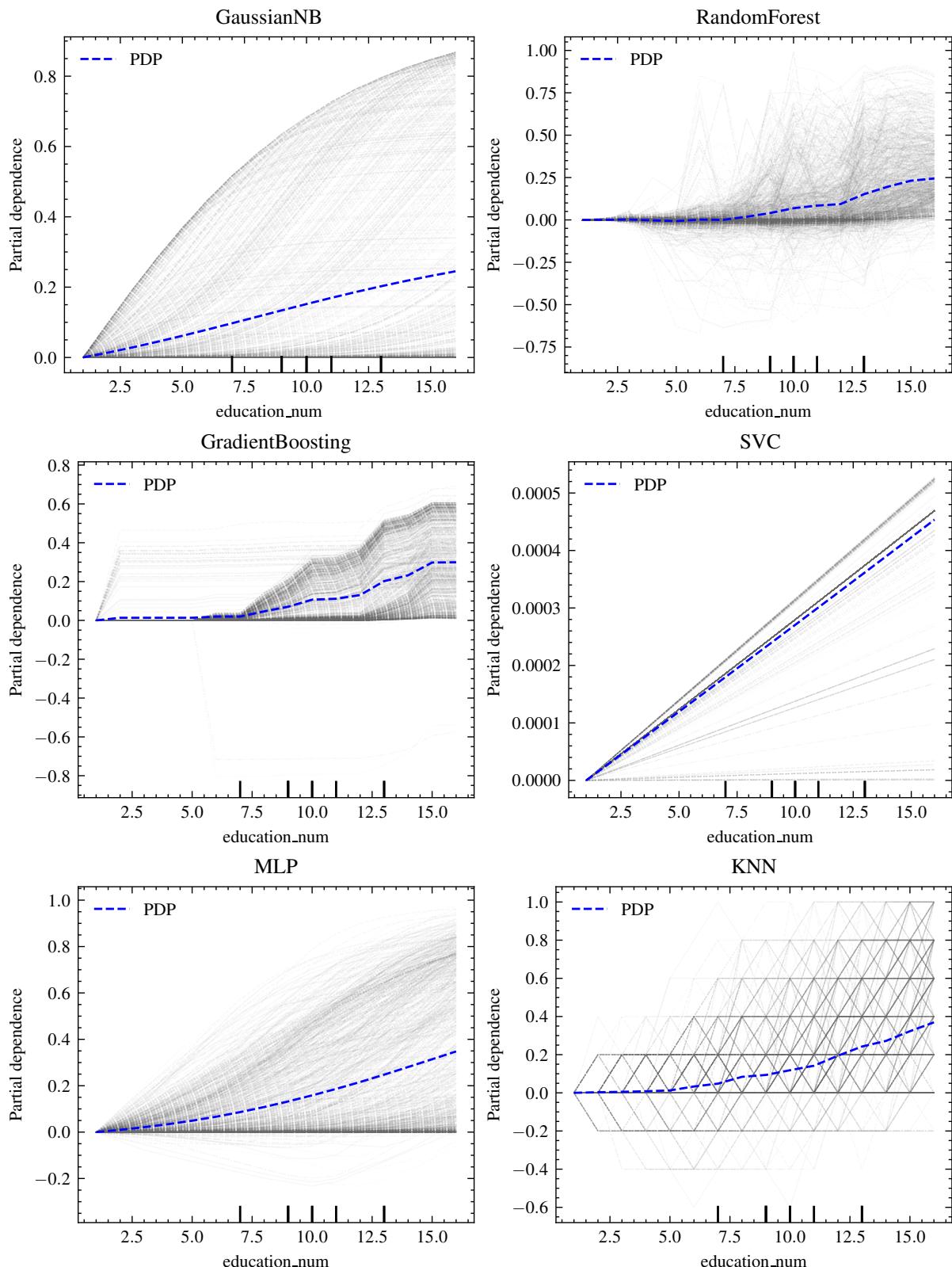
Rysunek 15: Krzywe ICE dla zmiennej age.



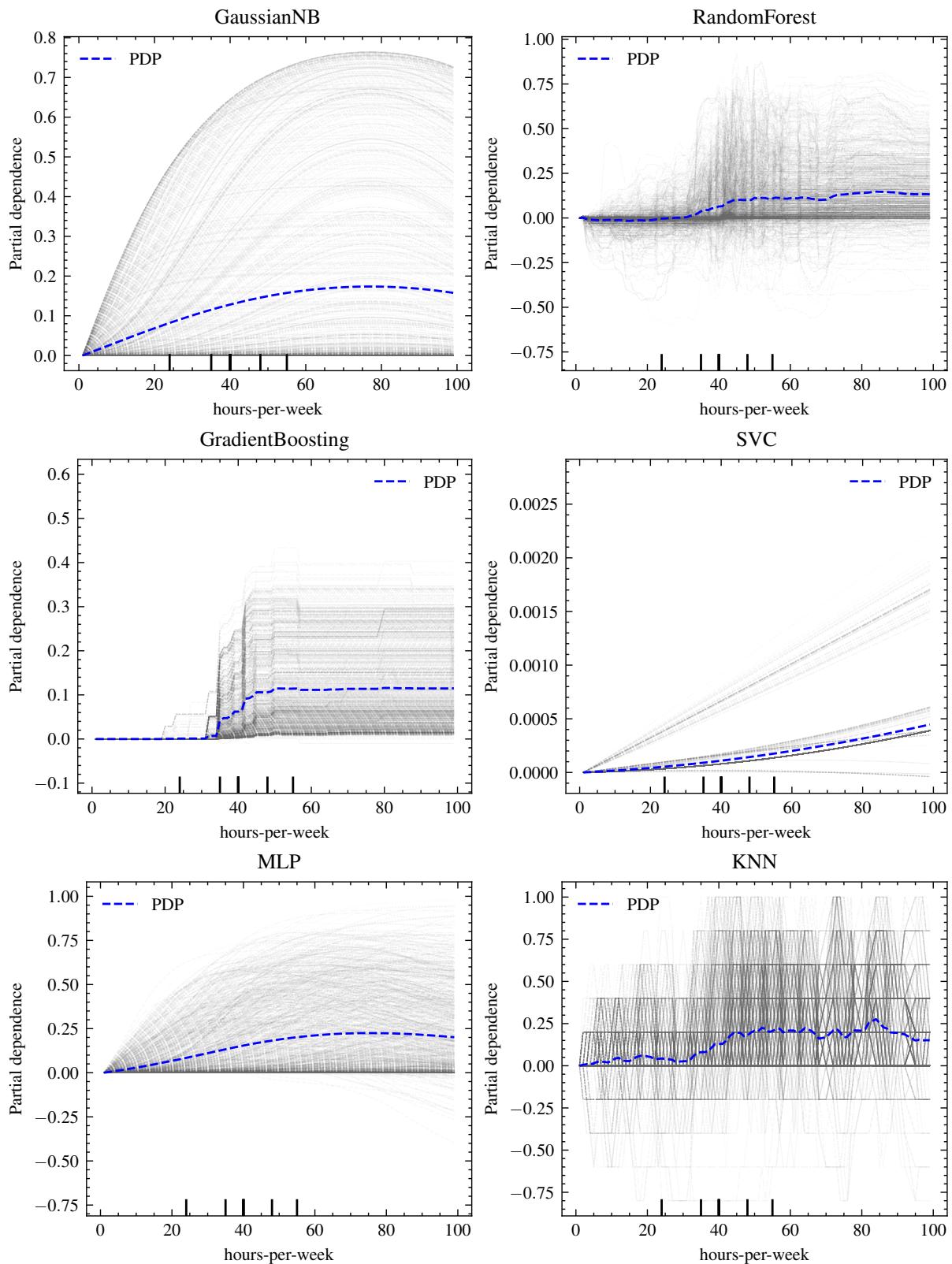
Rysunek 16: Krzywe ICE dla zmiennej `capital-gain`.



Rysunek 17: Krzywe ICE dla zmiennej capital-loss.



Rysunek 18: Krzywe ICE dla zmiennej `education-num`.



Rysunek 19: Krzywe ICE dla zmiennej `hours-per-week`.



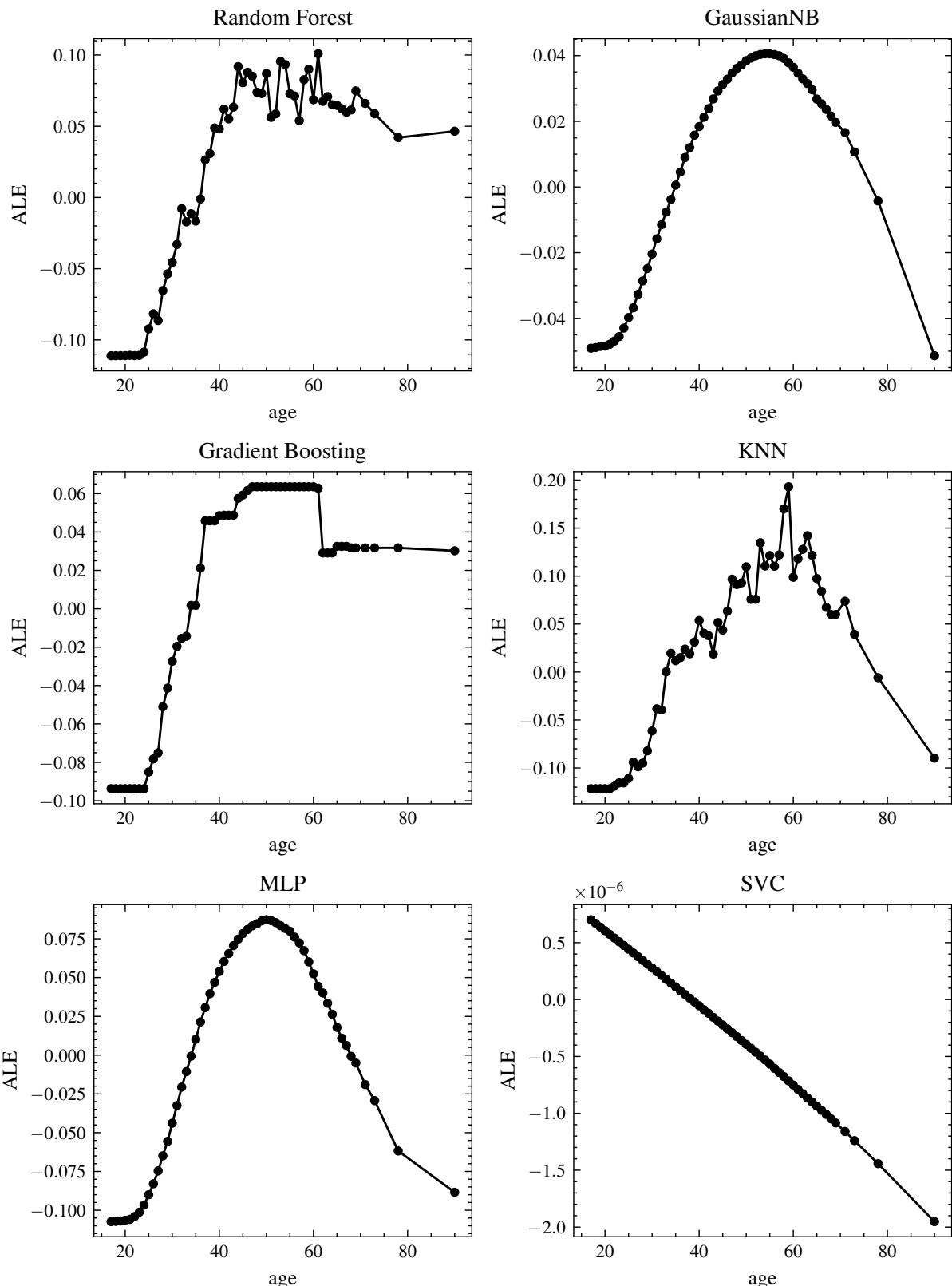
4.4 ALE

Przedstawione wykresy krzywych ALE pozwalają dostrzec podobieństwa, jak i różnice między modelami. Dają one nam również możliwość zrozumienia prawidłowości, według których model przewiduje dochody badanej osoby (zmienna objaśniana).

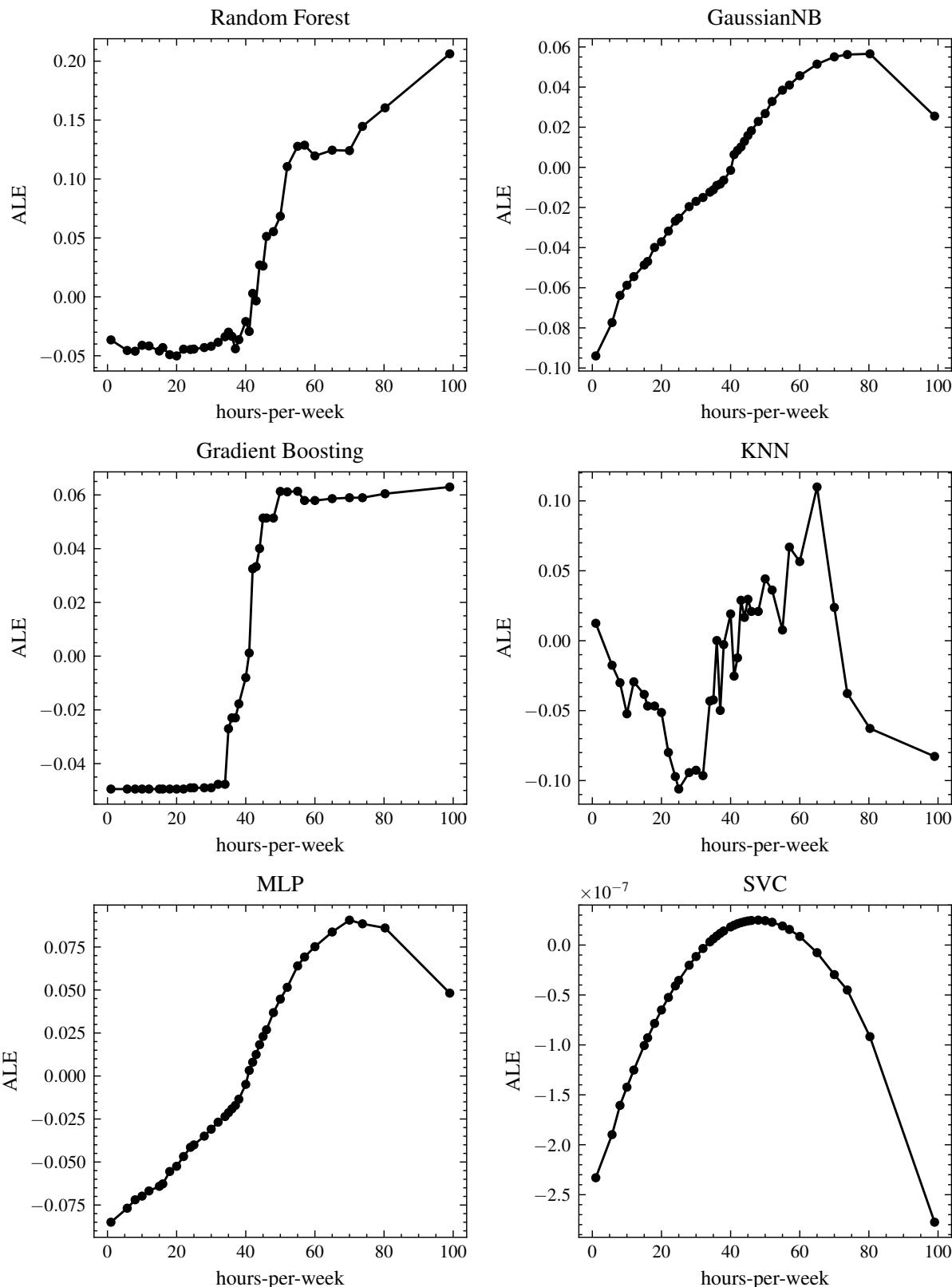
Dzięki wykresom, w łatwy sposób można dostrzec ogromny wpływ zmiennej *capital-gain* na predykcję wszystkich modeli.

Dość duży i zgodny z intuicją wpływ na predykcję ma zmienna *education-num*, gdzie obserwujemy, że im większy poziom edukacji, tym większy dodatni wpływ na zmienną zarobków.

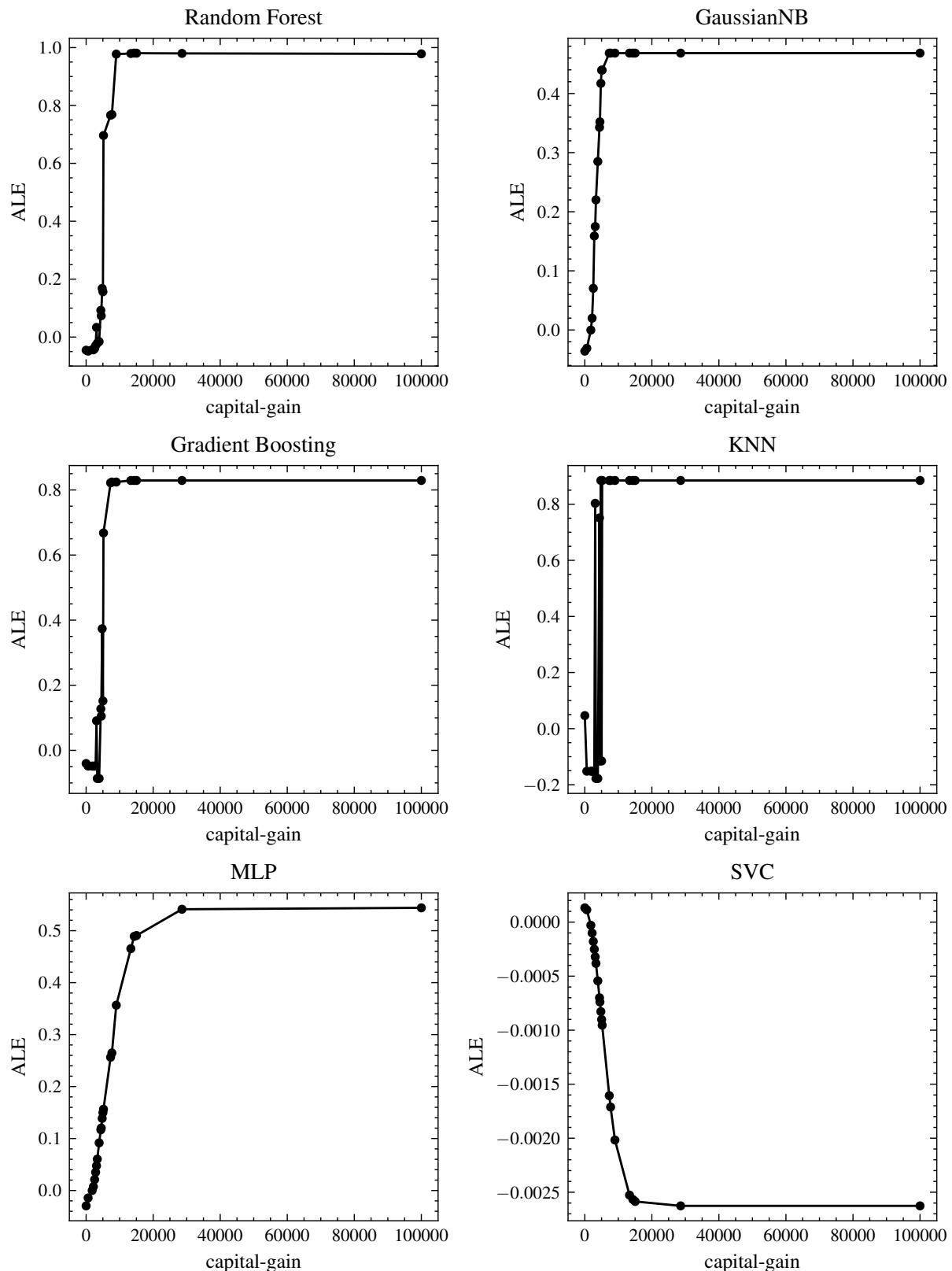
Najciekawszym przykładem zdaje się być jednak zmienna *age*, która choć nie ma relatywnie dużego wpływu na predykcję, to w wielu modelach oddaje kształt odwróconej paraboli, wskazując na to, iż najwięcej zarabiamy w średnim wieku, mniej zaś jako młodzi dorosli czy osoby starsze.



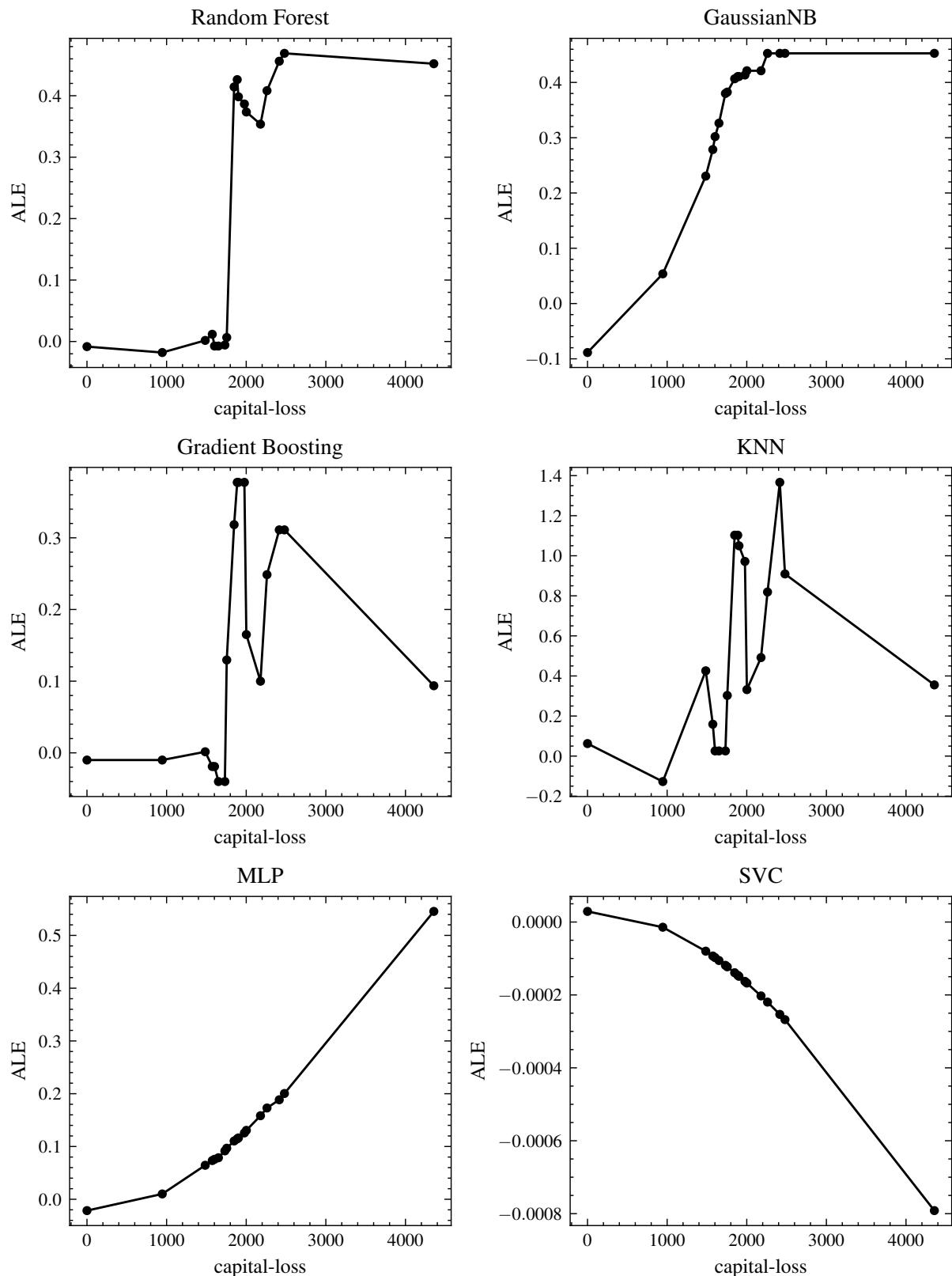
Rysunek 20: Krzywe ALE dla zmiennej *age*.



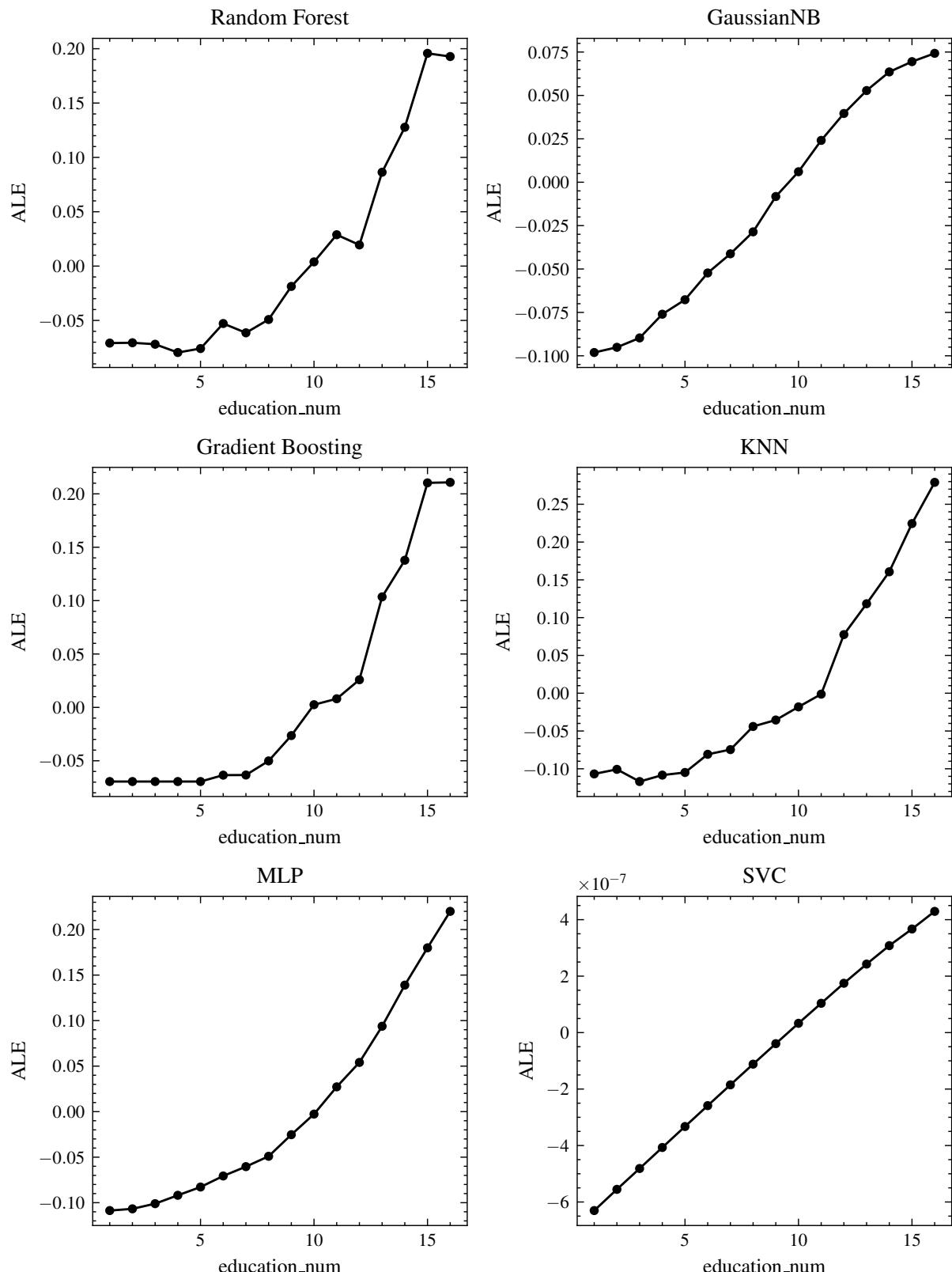
Rysunek 24: Krzywe ALE dla zmiennej *hours-per-week*



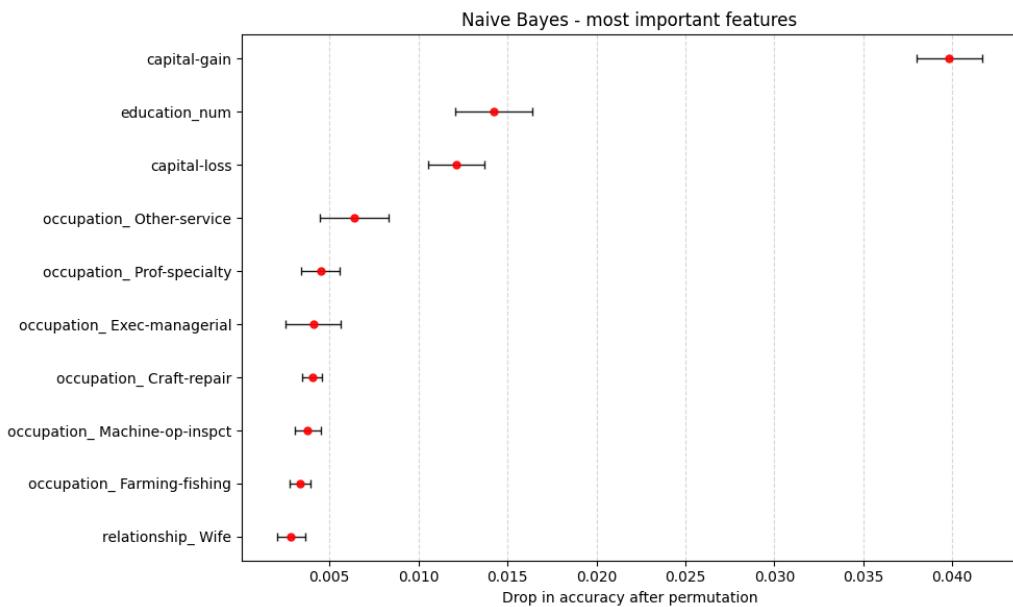
Rysunek 21: Krzywe ALE dla zmiennej *capital-gain*.



Rysunek 22: Krzywe ALE dla zmiennej *capital-loss*.



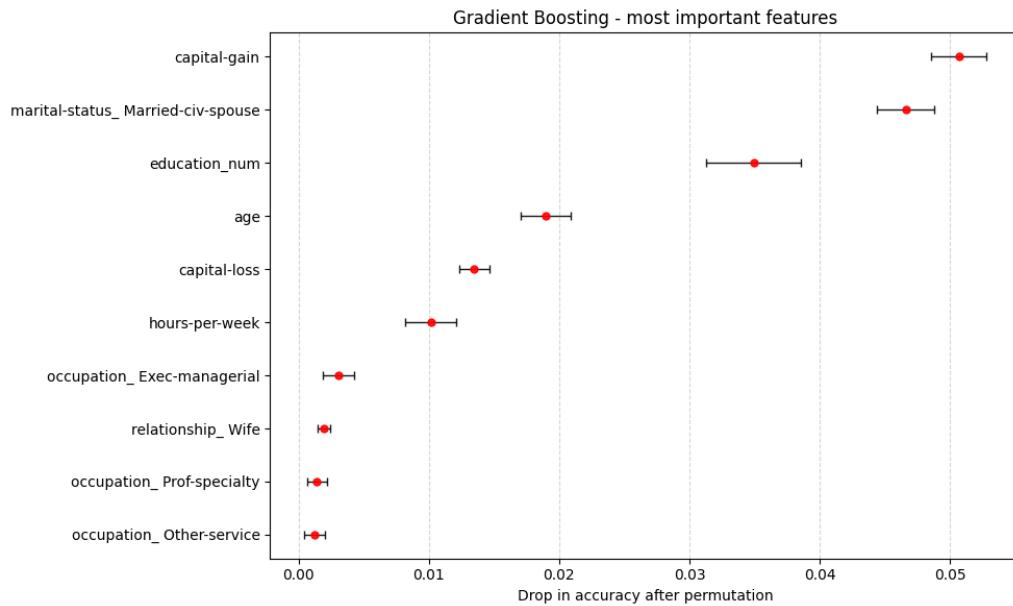
Rysunek 23: Krzywe ALE dla zmiennej *education-num*.



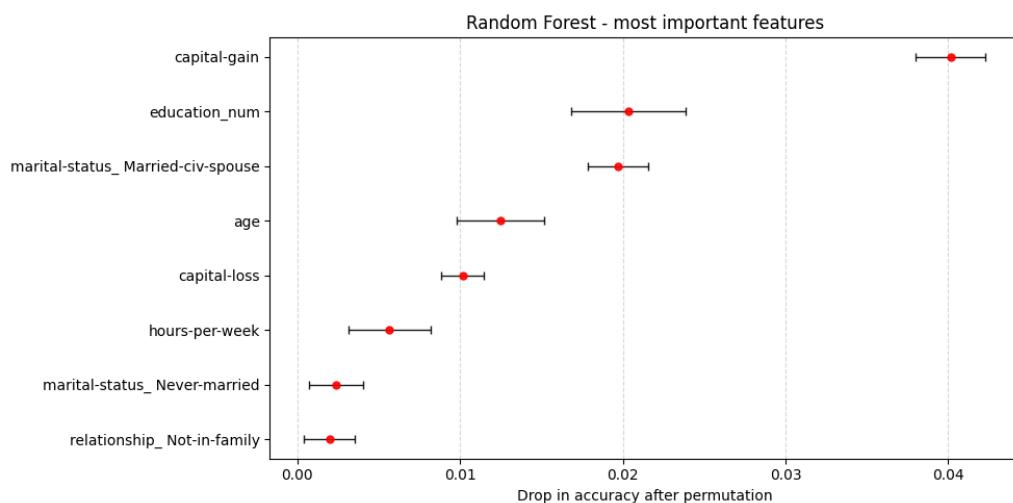
Rysunek 25: Permutacyjna istotność dla ‘GaussianNB’

4.5 Permutacyjna istotność zmiennych

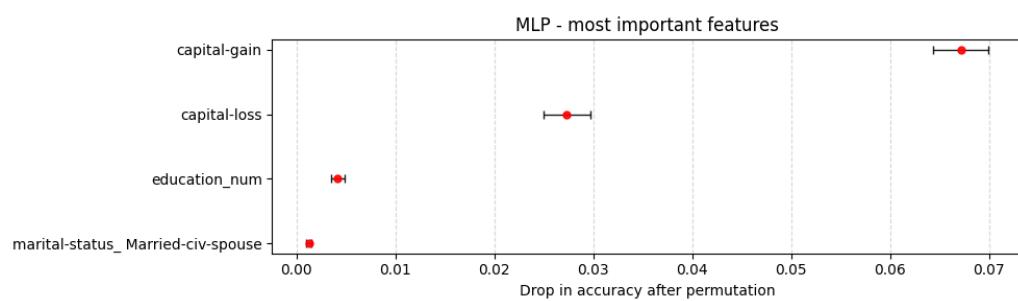
Poniżej zamieszczone zostały wykresy najbardziej wpływowych zmiennych. Wykresy zostały skonstruowane dla wszystkich modeli poza SVC, dla którego program nie potrafił policzyć permutacyjnej istotności. Na wykresach zostały zaznaczone średnie odchyły *accuracy* (czerwone kropki) oraz odchylenia standardowe (czarne kreski). To co rzuca się w oczy, to największy wpływ zmiennej ‘capital-gain’ w każdym modelu, co jest dosyć oczywiste zważywszy na to, co modelujemy. Dodatkowo, widać, że w modelu ‘GaussianNB’ zmienne mają, co do wartości, znacznie mniejszy wpływ niż przy innych modelach, co odpowiada założeniu warunkowej niezależności w założeniach modelu.



Rysunek 26: Permutacyjna istotność dla ‘GradientBoosting’



Rysunek 27: Permutacyjna istotność dla ‘RandomForest’

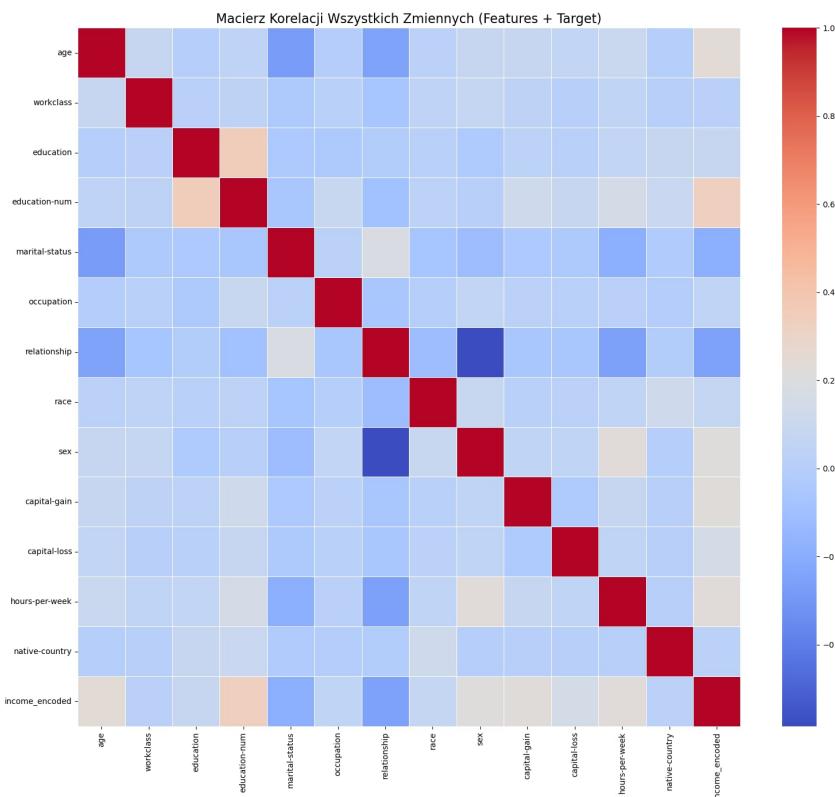


Rysunek 28: Permutacyjna istotność dla ‘MLP’

5 Wnioski

Mimo wykorzystania tego samego zbioru treningowego, każdy z modeli dostarczał nieco inne wyniki w ramach metod wyjaśniającego uczenia maszynowego, co wynika bezpośrednio z różnic w ich architekturze i sposobie działania. Niemniej jednak, w wielu przypadkach obserwowane trendy były zbliżone. Porównanie tych trendów pomiędzy modelami może stanowić dodatkowy argument wspierający ocenę istotności poszczególnych zmiennych.

Analiza macierzy korelacji (Rysunek 29.) wykazała, że choć większość zmiennych w naszym zbiorze danych ma niskie korelacje, istnieją istotne wyjątki. Wobec tego, preferowane jest stosowanie **wykresów ALE** do interpretacji wpływu cech na predykcje modelu. Ich konstrukcja zapewnia większą wiarygodność i odporność na zniekształcenia wynikające z ukrytych zależności.



Rysunek 29: Korelacja wszystkich zmiennych

Przykład: Dla zmiennej *hours-per-week* w modelu GradientBoostingClassifier, zarówno wykres PDP (Rysunek 19.), jak i ALE (Rysunek 21.) ukazują podobny trend:



prawdopodobieństwo wysokiego dochodu gwałtownie wzrasta dla osób pracujących około 40-50 godzin tygodniowo. Ta zbieżność jest zrozumiała ze względu na niską korelację tej cechy z większością innych zmiennych wejściowych. Jednakże, w przypadku cech silnie skorelowanych, różnice między PDP a ALE byłyby znacznie bardziej widoczne, a ALE dostarczyłoby bardziej rzetelnej interpretacji.

Analiza za pomocą metod XAI wykazała, że dla większości zastosowanych klasyfikatorów (Gradient Boosting, Random Forest, MLP, k-NN), kluczowymi zmiennymi wpływającymi na przewidywanie wysokiego dochodu ($>50K$) były: *marital-status_Married-civ-spouse* (zazwyczaj pozytywny wpływ), *capital-gain* (pozytywny wpływ), *education_num* (pozytywny wpływ, co oznacza, że wyższe wykształcenie koreluje z wyższym dochodem) oraz *age* (zwykle pozytywny wpływ w średnim wieku). Wykresy SHAP konsekwentnie pokazywały te cechy jako najbardziej wpływowe w ogólnym rozkładzie. Indywiduálne krzywe ICE oraz skumulowane efekty lokalne (ALE) potwierdziły te zależności, precyzując nielinowy wpływ zmiennych numerycznych, takich jak *age* i *hours-per-week*. Podsumowując, niezależnie od wybranego klasyfikatora, metody XAI spójnie wskazały na status cywilny, zyski kapitałowe, poziom wykształcenia oraz wiek jako najważniejsze czynniki w predykcji poziomu dochodów.



Literatura

- [1] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” 2014.
- [2] B. Becker and R. Kohavi, “Adult.” UCI Machine Learning Repository, 1996. DOI: <https://doi.org/10.24432/C5XW20>.