Spis treści

I. Eksploracyjna analiza danych

1. Wstępne informacje o zbiorze danych

Zbiór danych pochodzi z instytucji szkolnictwa wyższego i zawiera informacje o studentach. Zawiera dane zebrane podczas zapisu studenta (ścieżka akademicka, dane demograficzne i czynniki społeczno-ekonomiczne) oraz wyniki akademickie na koniec pierwszego i drugiego semestru. Zbiór również informuje, o statusie studenta (tzn. czy dalej studiuje, czy porzucił studia czy już jest absolwentem).

2. Informacje o zbiorze danych:

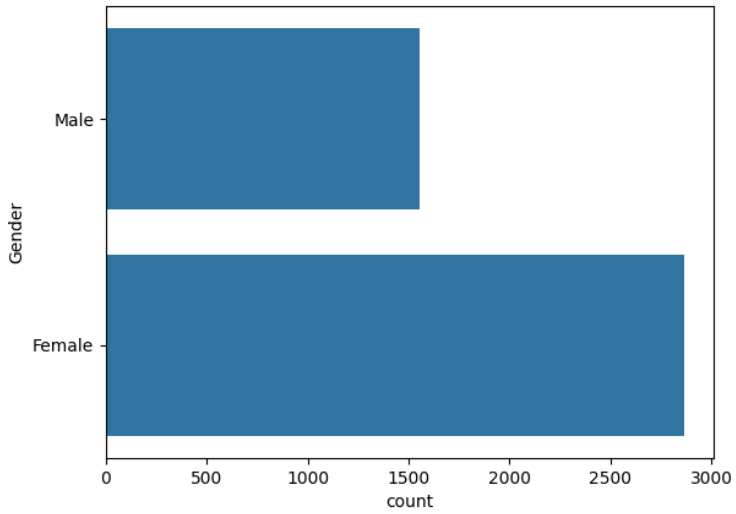
Zbiór składa się z 4424 wierszy oraz z 35 kolumn:

Kolumny reprezentują poszczególne kategorie:

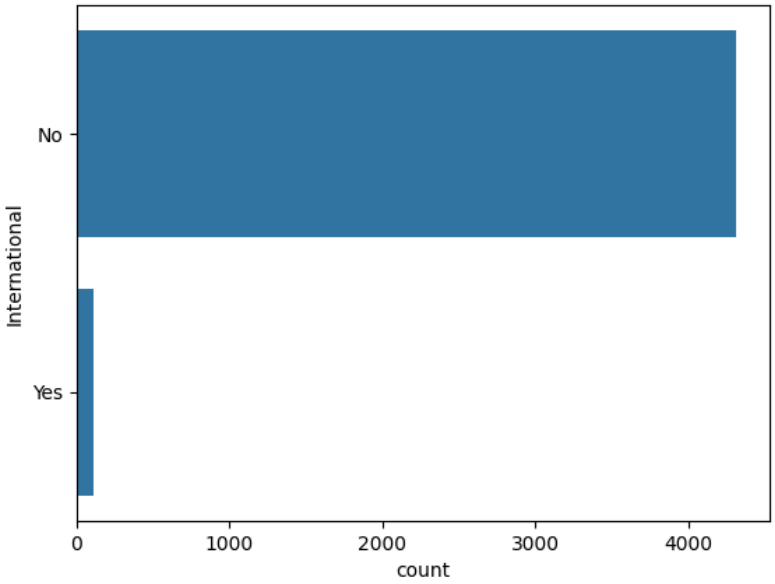
- Marital status
- Application mode
- Application order
- Course
- Daytime/evening attendance
- Previous qualification:
- Nationality
- Mother's / Father’s qualification
- Mother's / Father’s occupation
- Displaced
- Educational special needs
- Debtor
- Tuition fees up to date
- Gender
- Scholarship holder
- Age at enrollment
- International
- Curricular units 1st / 2nd sem (credited)
- Curricular units 1st / 2nd sem (enrolled)
- Curricular units 1st / 2nd sem (evaluations)
- Curricular units 1st / 2nd sem (approved)
- Curricular units 1st / 2nd sem (grade)
- Curricular units 1st / 2 nd sem (without evaluations)
- Unemployment rate
- Inflation rate
- GDP
- Target

3. Podstawowe zależności zbioru danych przedstawione na wykresach

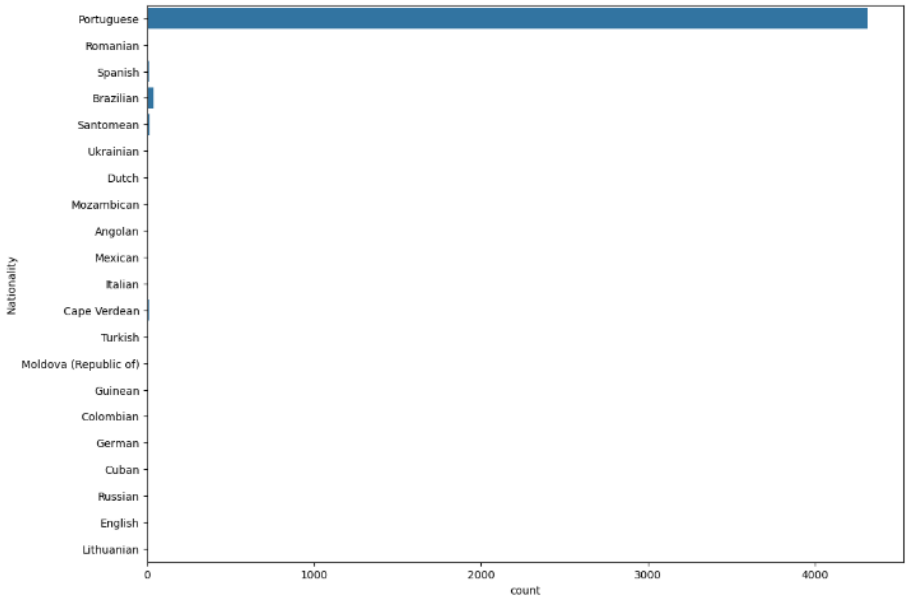
a) Wykres liczności studentów z podziałem na płeć. Z niego wprost wynika, że większość studentów ujętych w tym zbiorze danych to kobiety



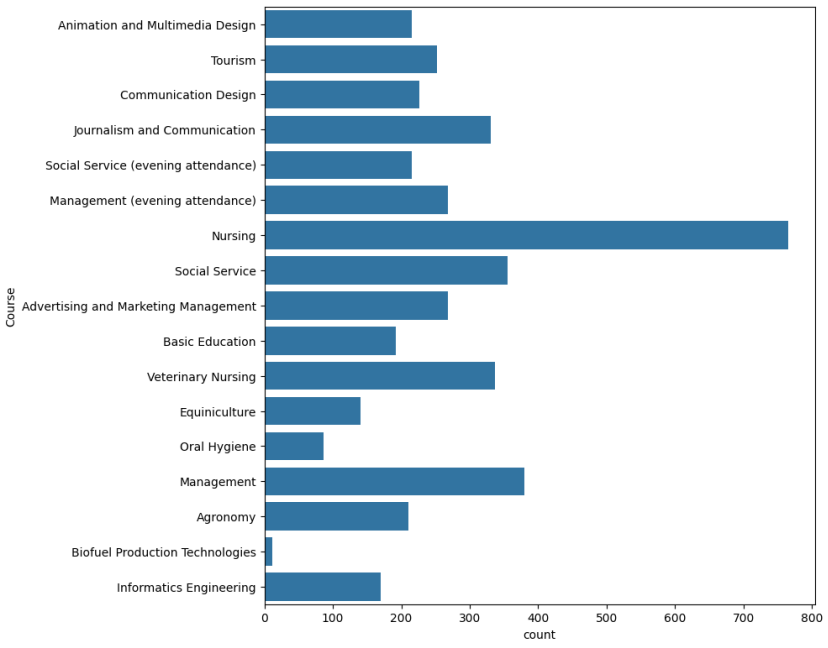
b) Wykres liczności studentów z podziałem na to czy są z zagranicy. Z niego wynika, że znaczna większość studentów jest jednak z kraju, w którym studiują.



c) Wykres liczności studentów z podziałem na kraj pochodzenia. I tutaj mamy dowód na to, że studenci pochodzący z jednego kraju – z Portugalii dominują. Więc możemy założyć, że zbiór danych był oparty na portugalskich uczelniach



d) Wykres przedstawiający to, na jaki kurs najczęściej zapisywali się studenci. Najchętniej wybieralnym kursem było pielęgniarstwo.

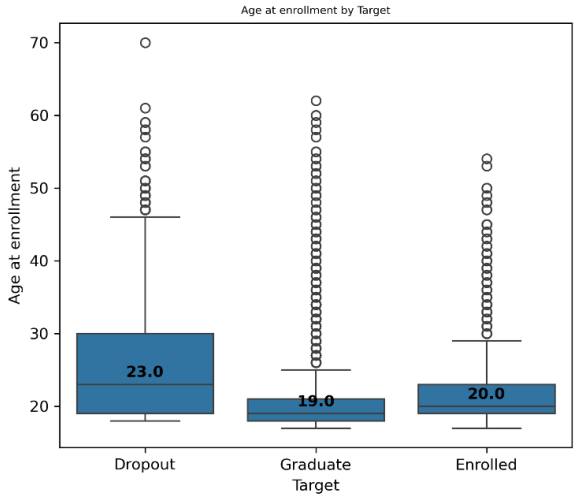


4. Wizualizacja danych z pomocą wykresów boxplot

a) Wykres przedstawia rozkład wieku studenta przy zapisie na studia w trzech grupach. Dane z wykresu możemy zinterpretować następująco:

I. Studenci, którzy skończyli studia, są najmłodsi w momencie zapisu (mediana 19 lat)

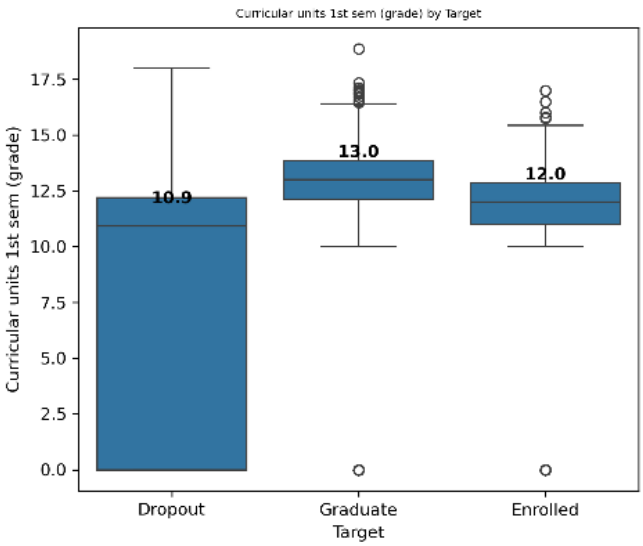
II. Osoby, które rezygnują, są najstarsze w momencie zapisu (mediana 23 lata). Może to wskazywać na osoby, które podjęły studia później niż rówieśnicy lub wróciły do nauki po przerwie.



b) Wykres przedstawia rozkład średniej ocen za 1 semestr studenta w trzech grupach. Dane z wykresu możemy zinterpretować następująco:

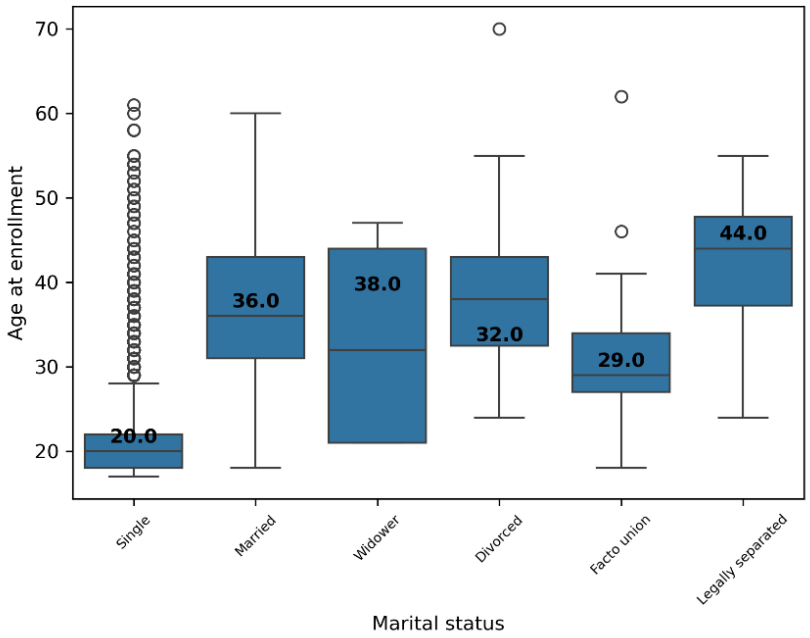
I. Osoby, które zrezygnowały, miały najniższą medianę (10.9) i największą zmienność wyników, co może oznaczać, że część z nich miała trudności akademickie.

II. Studenci, którzy ukończyli studia, mieli najwyższą medianę (13.0), co sugeruje, że dobre wyniki w pierwszym semestrze zwiększają szansę na ukończenie studiów



c) Wykres przedstawia rozkład wieku studenta przy zapisie na studia w grupach określających stan cywilny. Dane z wykresu możemy zinterpretować następująco

I. Singiel ma najniższą medianę (20 lat) – większość młodych studentów to osoby niepozostające w związkach małżeńskich.
II. Osoby w małżeństwie mają wyższą medianę (36 lat), co sugeruje, że często zaczynają studia w późniejszym wieku.

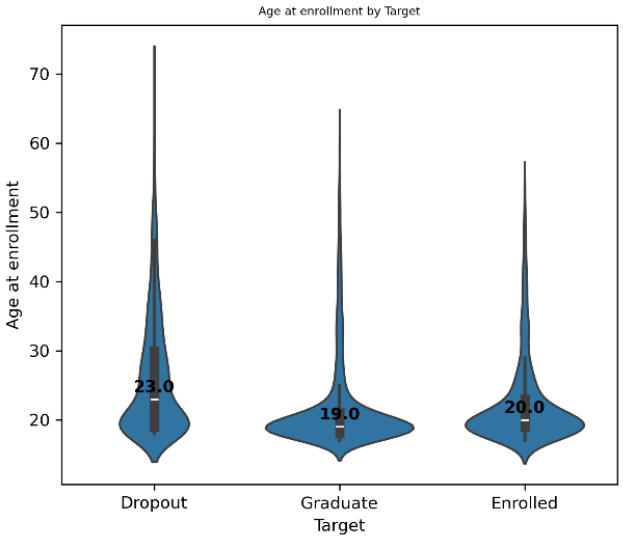


Na powyższych wykresach mogliśmy, zauważyć wiele odstających wartości (zaznaczone jako kółeczka). Wartości odstające znacząco zwiększają średni wiek w każdej kategorii i mogą zaburzać analizę statystyczną. Szczególnie (na wykresie a.) kategorie "Dropout" i "Graduate" mają dużą liczbę wartości odstających, sięgających nawet 70 lat w przypadku "Dropout" i ponad 60 lat w przypadku "Graduate". Lecz przedstawiona mediana jest odporna na ekstremalne wartości odstające

5. Wizualizacja danych za pomocą wykresów violinplot

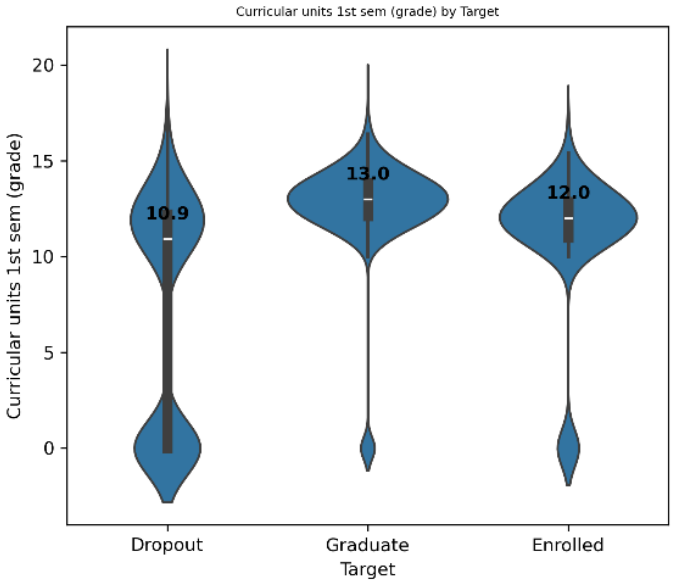
a) Wykres przedstawia rozkład wieku studentów w momencie rozpoczęcia nauki w trzech grupach

I. Wykres ten dobitniej pokazuje, że większość studentów we wszystkich kategoriach to osoby młode (im szersza część tym więcej danych w tym zakresie), ale występują też wartości odstające, sięgające nawet 70 lat, szczególnie w grupie Dropout



b) Wykres przedstawia rozkład średniej ocen za 1 semestr studentów w trzech grupach

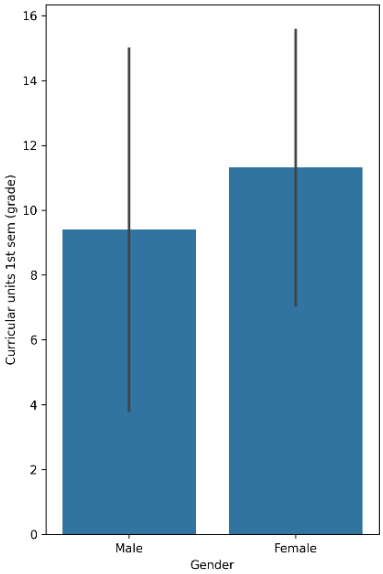
I. Absolwenci mają najwyższe średnie ocen w pierwszym semestrze, osoby przerywające naukę - najniższe. Interesujący jest też drugi szczyt w dolnej części wykresu dla grupy Dropout - może to wskazywać na grupę osób, które uzyskały bardzo niskie wyniki lub zerowe (np. w ogóle nie uczestniczyły w zajęciach).



6. Wizualizacja cech numerycznych za pomocą errorbars

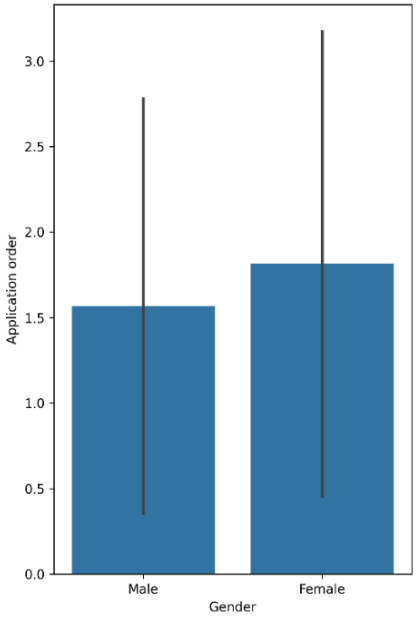
a) Wykres przedstawia średnie oceny z pierwszego semestru dla dwóch grup
Słupki błędu reprezentują odchylenie standardowe:

Wykres pokazuje, że to kobiety mają większą średnią ocen za 1 semestr.
Duże słupki błędu na wykresie sugerują, że istnieje duża zmienność wyników w każdej grupie (tzn. średnie ocen poszczególnych studentów znacznie różnią się od siebie



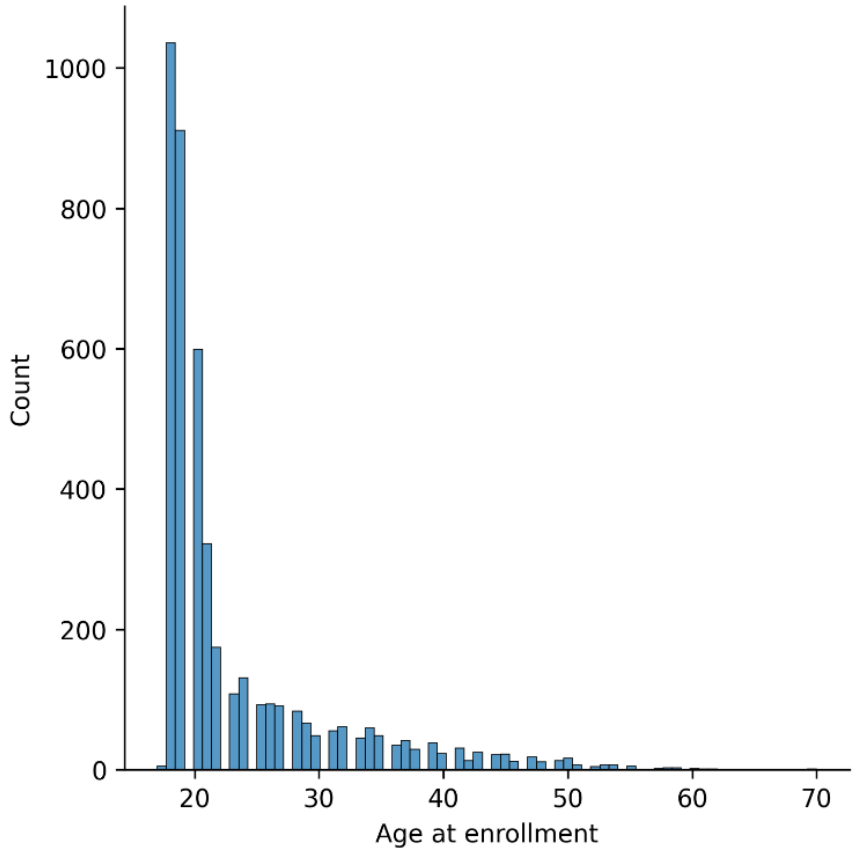
b) Oba wykresy przedstawiają kolejność wyboru danego kierunku studiów dla dwóch grup

Wykres pokazują, że to średnio mężczyźni częściej wybierali dane studia jako priorytetowe (tzn. umieszczali je na pierwszych miejscach podczas swoich wyborów)
W grupie istnieje duża zmienność indywidualnych wyników (większa zmienność występuję w grupie kobiet, czyli ich kolejność wyboru nie była taka jednoznaczna)



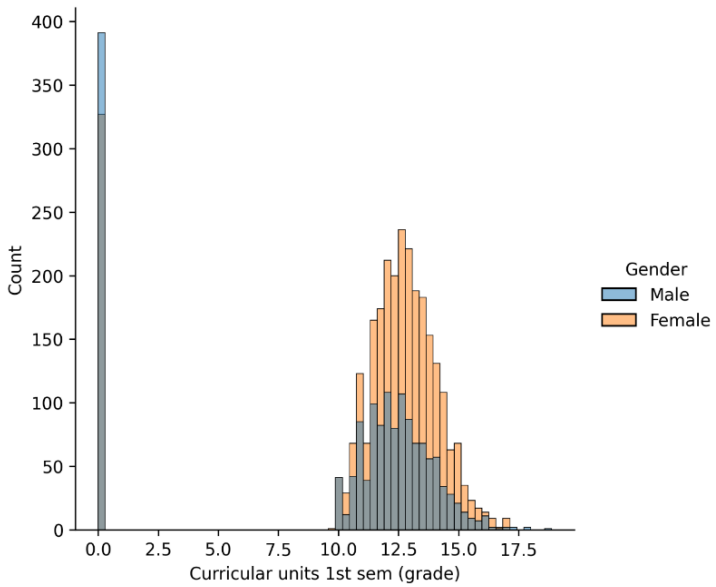
7. Prezentacja histogramów dla cech numerycznych

a)



Na powyższym wykresie widoczne jest to że większość studentów zapisuje się na studia w młodym wieku, ale istnieją również jednostki podejmujące edukację w późniejszych etapach życia. Obserwowana struktura jest zgodna z typowym modelem zapisów na studia, gdzie młodsze osoby stanowią większość, a starsze zapisują się znacznie rzadziej.

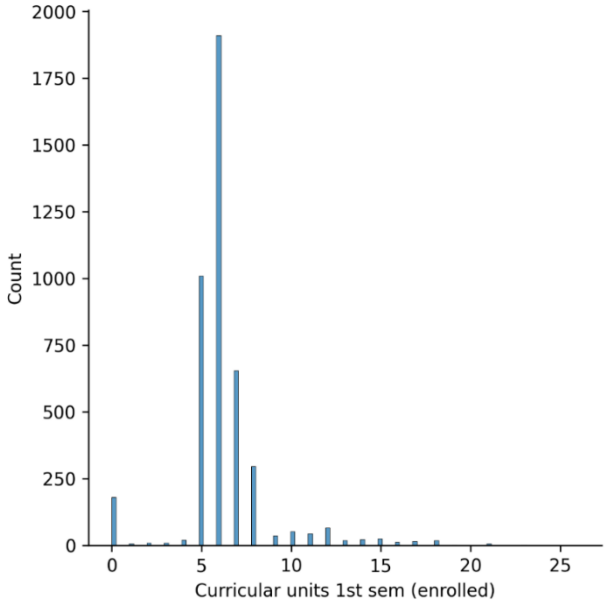
b)



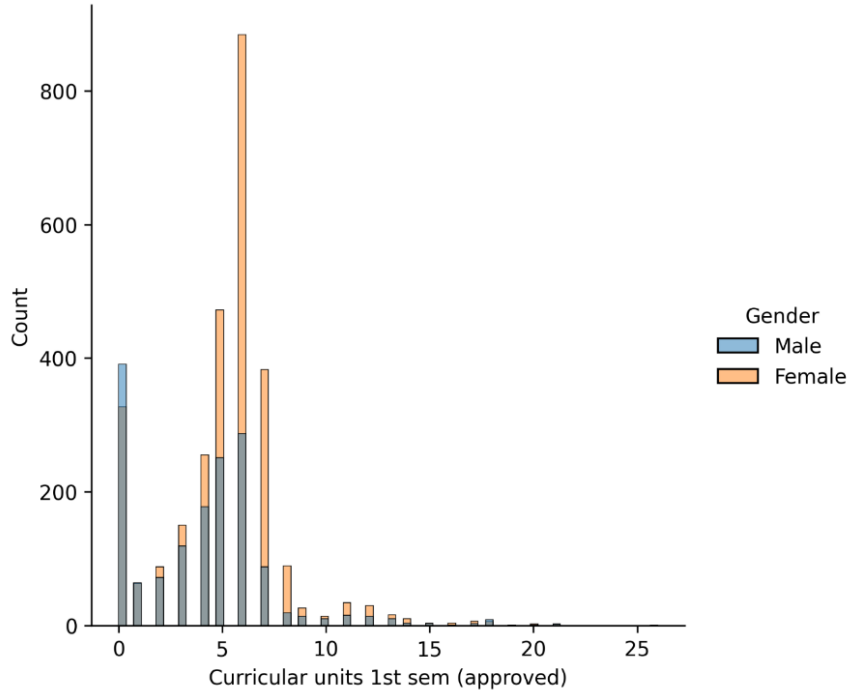
Na histogramie przedstawiono rozkład średniej za pierwszy semestr, ale z rozróżnieniem na płeć. Wykorzystanie warstwowego grupowania umożliwia identyfikację potencjalnych różnic między mężczyznami a kobietami w wartości średniej za 1 semestr:

Na wykresie możemy zobaczyć, że to kobiety osiągają większą średnią w porównaniu do mężczyzn. Oraz, że to mężczyźni częściej osiągają średnią ocen na poziomie ~ 0

c)



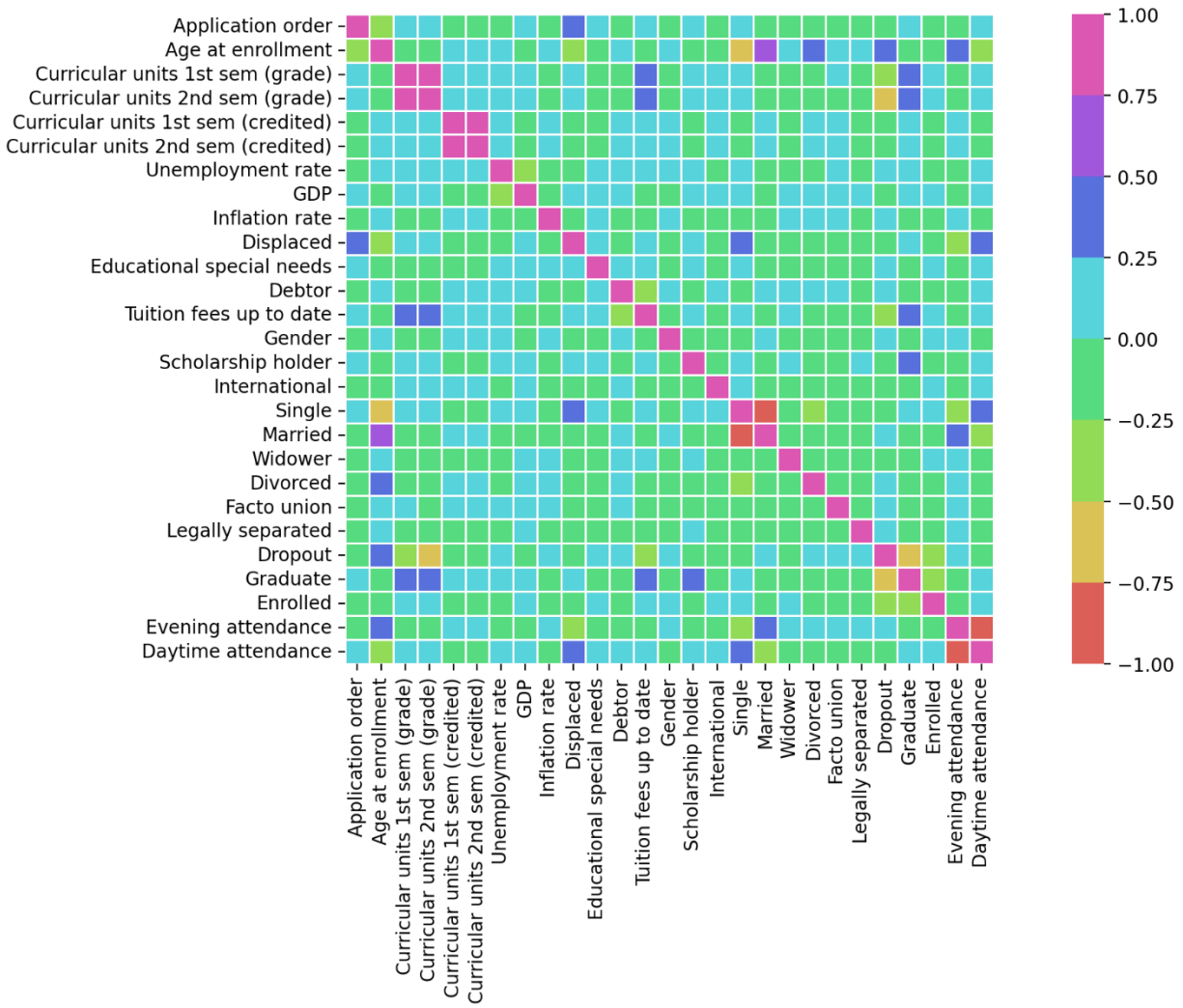
Na powyższym wykresie widoczne jest to że studenci na pierwszym semestrze najchętniej zapisują się na od 5 do 8 przedmiotów. Ale istnieją również niemarginalne przypadki, kiedy studenci zapisują się na więcej niż 12 przedmiotów.



Powyższy wykres nawiązuje do poprzedniego, ponieważ przedstawia liczbę jednostek zaliczonych w 1 semestrze. Widać tutaj, że istnieje sporo osób które prawdopodobnie zaliczyły wszystkie przedmioty na które się zapisały, ale pojawią się też znaczny odsetek osób które nie zaliczyły kilku przedmiotów, oraz widoczna duża liczba studentów którzy nie zaliczyli żadnego (prawdopodobnie porzucili studia wcześniej

Widoczny jest również wyraźny wzorzec różnic między płciami, sugerujący, że kobiety osiągają lepsze wyniki akademickie w pierwszym semestrze, zaliczając więcej jednostek dydaktycznych niż mężczyźni.

8. Heatmap korelacji danych



Heatmap przedstawia macierz korelacji pomiędzy różnymi zmiennymi w zbiorze danych, z której to możemy odczytać następujące zależności:

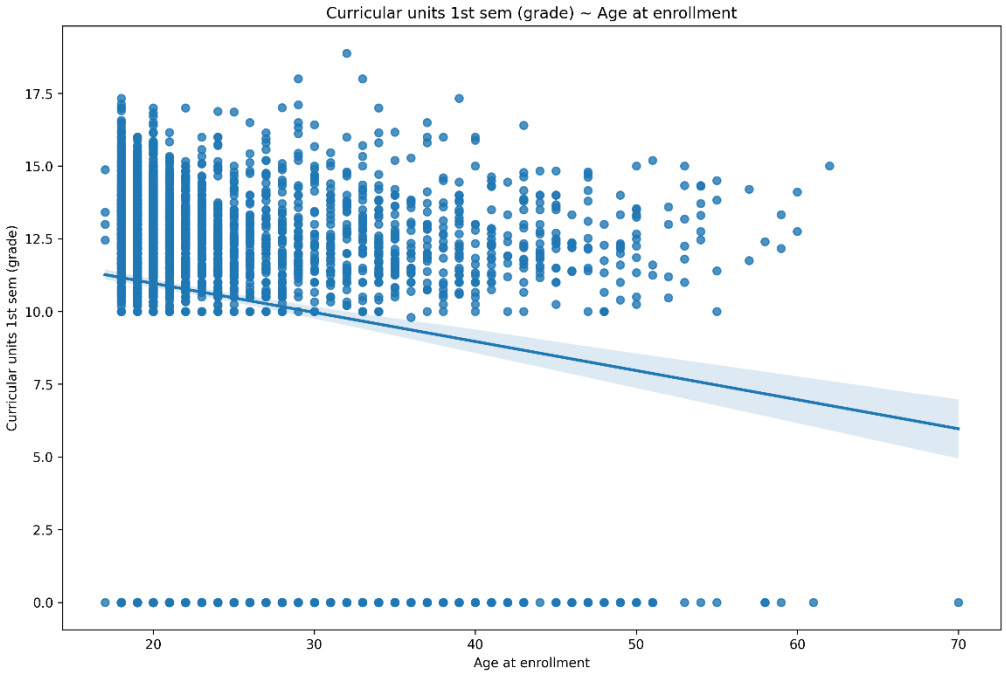
a) Scholarship holder a Curricular units (grade): Istnieje pozytywna korelacja między posiadaniem stypendium a wynikami średniej za semestr. Studenci, którzy otrzymują stypendia, zwykle uzyskują lepsze wyniki w nauce.
b) Age at enrollment a Married: Korelacja pomiędzy wiekiem zapisu na studia a stanem cywilnym, oznacza to, że starsi studenci, którzy zaczynają studia w późniejszym wieku, częściej są już w związku małżeńskim.
c) Związek między rodzajem studiowania a stanem cywilnym: Analiza pokazuje silną zależność między typem studiowania (np. studia dzienne vs wieczorowe) a stanem cywilnym studenta. Osoby będące w związku małżeńskim częściej wybierają studia wieczorowe, podczas gdy single skłaniają się ku studiowaniu dziennemu.

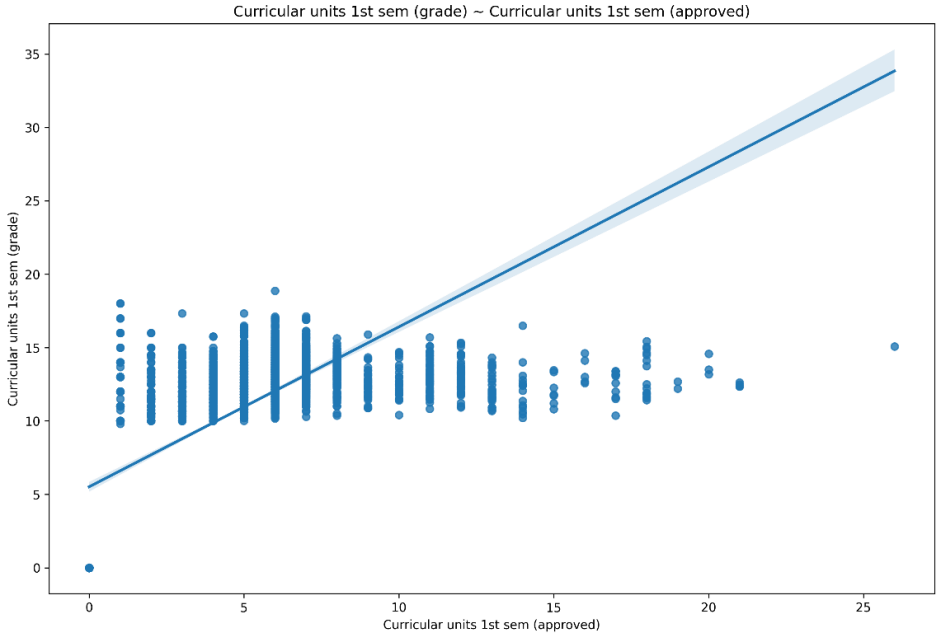
d) Curricular units 1st sem (grade) oraz Curricular units 2nd sem (grade): Zauważamy pozytywną korelację pomiędzy wynikami średniej w pierwszym a drugim semestrze. Oznacza to, że wysokie średnie w drugim semestrze często idą w parze z wysokimi średnimi w pierwszym semestrze.

9. Analiza i wizualizacja korelacji liniowej

a)

Przedstawiony wykres pokazuje regresję liniową badającą zależność między wiekiem studentów w momencie zapisania się na studia a ich średnia ocen za pierwszy semestr. Linia trendu ma wyraźne nachylenie ujemne, oznacza to, że statystycznie rzecz biorąc, im starszy student, tym niższa średnia w pierwszym semestrze.



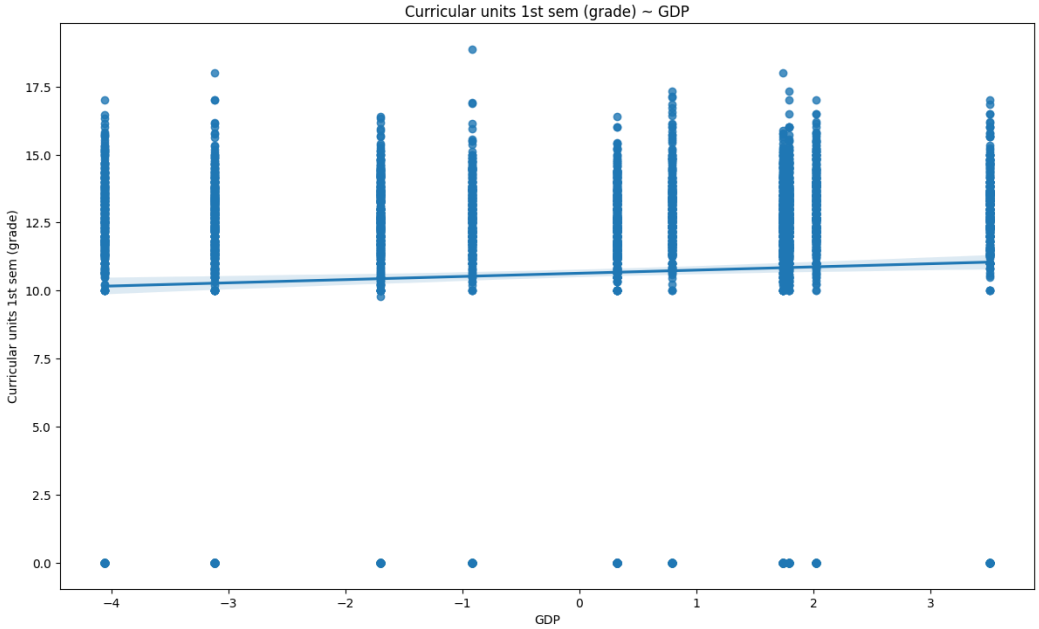


b)

Przedstawiony wykres pokazuje przedstawia regresje liniową badającą zależność między liczbą zaliczonych przedmiotów w 1 semestrze, a średnią za pierwszy semestr. Liniowa regresja wskazuje na dodatnią korelację – im więcej przedmiotów student zdał, tym zazwyczaj wyższa była jego średnia ocen (ale nie gwarantuje bardzo wysokiej średniej)

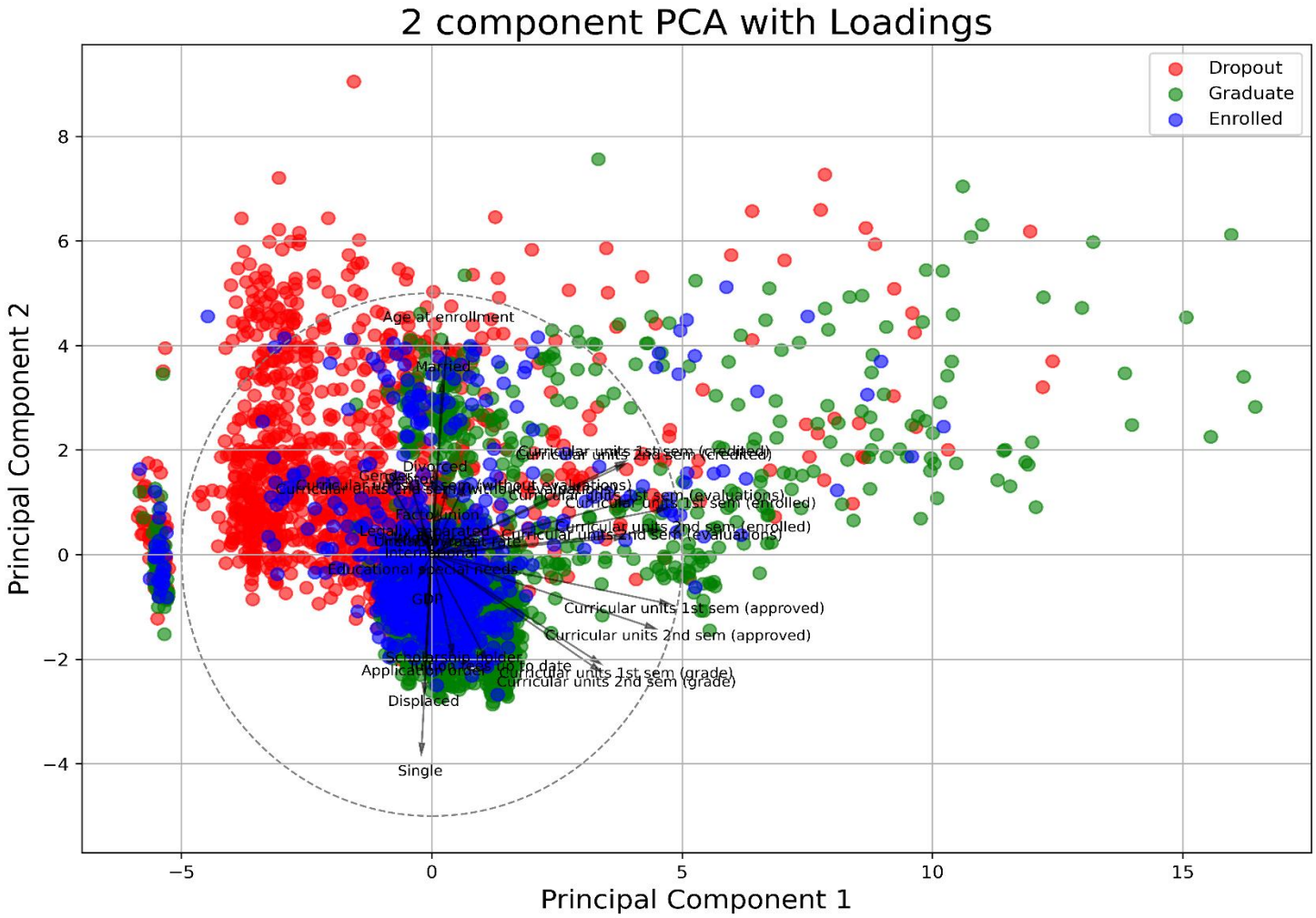
c)

Przedstawiony wykres pokazuje analizę regresji liniowej badającą zależność między PKB (GDP) a średnią za pierwszy semestr. Punkty danych są bardzo rozproszone wokół linii trendu, co wskazuje na słabą moc predykcyjną tego modelu. Na wykresie widoczna jest linia trendu o niewielkim nachyleniu dodatnim, co sugeruje bardzo słabą pozytywną korelację między PKB a ocenami studentów. Wraz ze wzrostem PKB występuje nieznaczny wzrost średnich ocen.



10. PCA

Po dokonaniu PCA, otrzymujemy wykres, który składa się tylko z dwóch głównych składowych. Te dwie osie są uproszczeniem wielowymiarowości zbioru danych, który składał się z wielu cech. Na wykresie, również naniesione są loadings (czyli strzałki, które pokazują jak każda oryginalna zmienna wpływa na daną główną składową)



I. Najdłuższe wektory wskazujące w prawo to "Curricular units 1st/2nd sem (approved)" i "Curricular units 1st/2nd sem (grade)" - oznacza to, że studenci z wysokimi ocenami i zaliczonymi przedmiotami częściej kończą studia. (widzimy te zielone punkty). Widać wyraźną korelację między osiągnięciami akademickimi a ukończeniem studiów.

II. Przesuwając się na lewo od środka PC1 (czyli naszej ścieżki akademickiej) napotykamy na duża grupę studentów, którzy porzucili studiowanie

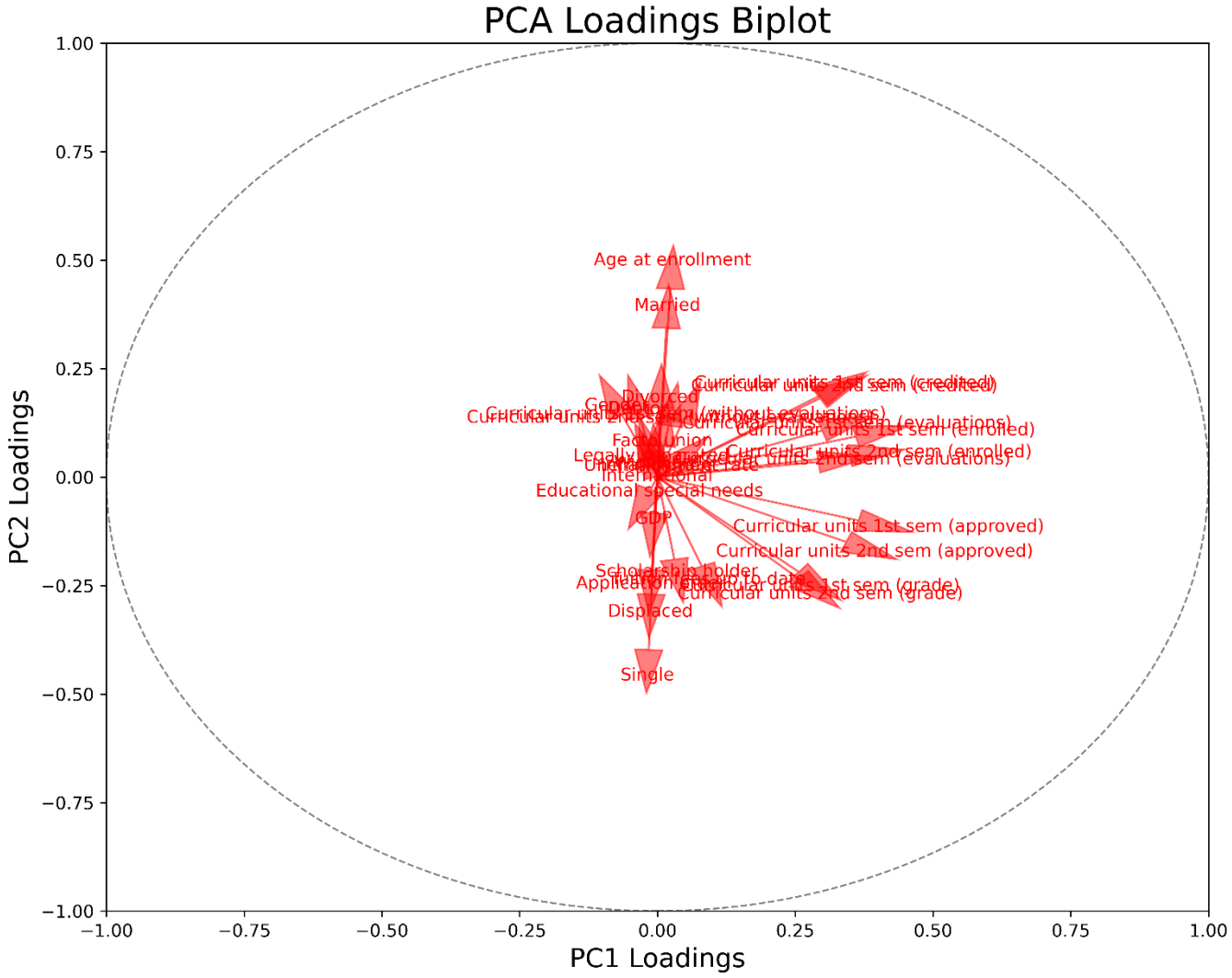
III. Przesuwając się w górę kierunku PC2 widzimy długie wektory odpowiadające za czynniki socjalne (wiek, stan cywilny, płeć)

W tym przypadku dwa główne komponenty z variance ratio odpowiednio:
a) PC1 - **0.20798626**
b) PC2 - **0.1097144**

Oraz tabela jak poszczególne cechy wpływają na nowe składowe:

Variable	***PC1***	***PC2***
Curricular units 1st sem (approved)	0.911833	-0.181255
Curricular units 1st sem (enrolled)	0.899333	0.169720
Curricular units 2nd sem (enrolled)	0.873261	0.089968
Curricular units 2nd sem (approved)	0.855741	-0.271153
Curricular units 1st sem (evaluations)	0.746369	0.194764
Curricular units 1st sem (credited)	0.737495	0.341232
Curricular units 2nd sem (credited)	0.736830	0.329778
Curricular units 2nd sem (evaluations)	0.729868	0.063814
Curricular units 2nd sem (grade)	0.641248	-0.425444
Curricular units 1st sem (grade)	0.641036	-0.396588
Tuition fees up to date	0.201292	-0.372869
Curricular units 1st sem (without evaluations)	0.111605	0.228289
Scholarship holder	0.076540	-0.344335
Age at enrollment	0.058891	0.787238
Unemployment rate	0.055151	0.040712
Curricular units 2nd sem (without evaluations)	0.050712	0.215575
Married	0.039926	0.622864
Facto union	0.019334	0.130018
Divorced	0.011282	0.289889
International	-0.001525	0.003296
Inflation rate	-0.003487	0.041837
GDP	-0.015581	-0.151621
Widower	-0.023696	0.052098
Legally separated	-0.024738	0.075845

Application order	-0.025798	-0.388232
Displaced	-0.028260	-0.490035
Educational special needs	-0.031230	-0.051727
Single	-0.039631	-0.721944
Debtor	-0.077561	0.247839
Gender	-0.160301	0.259906



II. Implementacja modeli uczenia maszynowego

1. Implementacja regresji liniowej za pomocą formuły zamkniętej

Regresja liniowa za pomocą zamkniętej formuły jest sposobem na obliczanie wag dla wejściowych cech, który opiera się na równaniach funkcji liniowej, które idealnie miałby przechodzić przez wszystkie punkty z *X*. Formuła zamknięta wyklucza udział iteracji w procesie uczenia, ale jest ograniczona, (np. nie zawsze istnieje możliwość odwrócenia macierzy)
Wzór na zamkniętą formułę:

$$\theta = (X^T X)^{-1} X^T y$$

(linear_regression.ipynb)

Typ zbioru	Wynik własnej implementacji	Wynik oryginalnej implementacji	Różnica procentowa (%)
Treningowy	0.836407	0.836410	0.00036
Testowy	0.831219	0.831318	0.0119
Walidacyjny	0.809923	0.809940	0.0021

Table 1: Porównanie wyników R^2 dla własnej implementacji regresji liniowej oraz dla implementacji oryginalnej

2. Implementacja wieloklasowej regresji logistycznej

Wieloklasowa regresja logistyczna jest rozszerzeniem regresji logistycznej, gdzie poprzez funkcje sigmoidalną wybieramy jedną z dwóch klas (0 lub 1). Z racji tego, że nasz problem ma więcej niż 2 klasy, rozwiązanie to używa funkcji aktywacji sotfmax, która będzie generowała odpowiednie prawdopodobieństwa dla danych wejściowych, dając możliwość przypisania etykiety dla rekordów. Jest to proces czysto iteracyjny z użyciem matematycznego gradientu funkcji, a za wyznacznik czy kierujemy się w dobrą stronę przyjmujemy CrossEntropyError function.

Typ zbioru	Wynik własnej implementacji	Wynik oryginalnej implementacji	Różnica procentowa (%)
Treningowy	0.804974	0.822498	2.13
Testowy	0.765060	0.769578	0.59
Walidacyjny	0.755086	0.759608	0.60

Table 2: Porównanie wyników skuteczności dla własnej implementacji regresji logistycznej oraz dla implementacji oryginalnej. Dla parametrów iterations=10000

3. Implementacja wieloklasowej regresji logistycznej w PyTorch

Koncept zostaje ten sam, tylko przewaga pytorcha jest w tym, że posiada m.in. automatyczne obliczanie gradientu, oraz funkcji błędu. Jak i daje możliwość trenowania swoich modelów z użyciem GPU, co znacząco przyśpiesza czasy treningu.

Zbiór	**Skuteczność**
Treningowy	0.818001
Testowy	0.744812

Table 3: Porównanie wyników skuteczności modelu regresji logistycznej PyTorch.Dla parametrów iterations=2000

Parametr	**Czas trwania treningu**
CPU	120 s
GPU	72 s

Table 4: Porównanie czasów treningu modelu w PyTorch bez GPU oraz z GPU

III. Optymalizacja modeli uczenia maszynowego

Uwaga: Jeśli nie będzie zaznaczone inaczej to przyjmujemy następujące parametry modelów:
(n_iters=5000, lr=0.001, batch_size=64)

Po implementacji modeli regresji logistycznej, możemy zastosować rożne techniki, aby poprawić wydajność i skuteczność predykcji danych modeli. Do rozważań będzie brany poprzedni model wieloklasowej regresji logistycznej.

1. Cross-validation i ewaluacja modelu

Wyniki pojedynczych wywołań danego modelu zależą głównie od elementu losowego, czyli jak z całego zbioru wylosuje się część treningowa i testowa. Dlatego aby wyniki dla danego modelu były bardziej przewidywalne wykonuje się cross-validacje naszego modelu.

accuracy(random_state=***42***)	0.7665
accuracy(random_state=***11***)	0.7890

Tabela 1 Pojedyńcze wywołania modelu BaseLogisticRegression

accuracy(**first_fold**)	0.7761
accuracy(**second_fold)**	0.7722
accuracy(**third_fold**)	0.7664
CV accuracy	0.7716

Tabela 2 Wyniki k3-cross-validacji modelu BaseLogisticRegression

Warto korzystać z CV, aby nie opierać się na możliwie złudnym przekonaniu, że nasz model osiągnął bardzo wysoką dokładność.

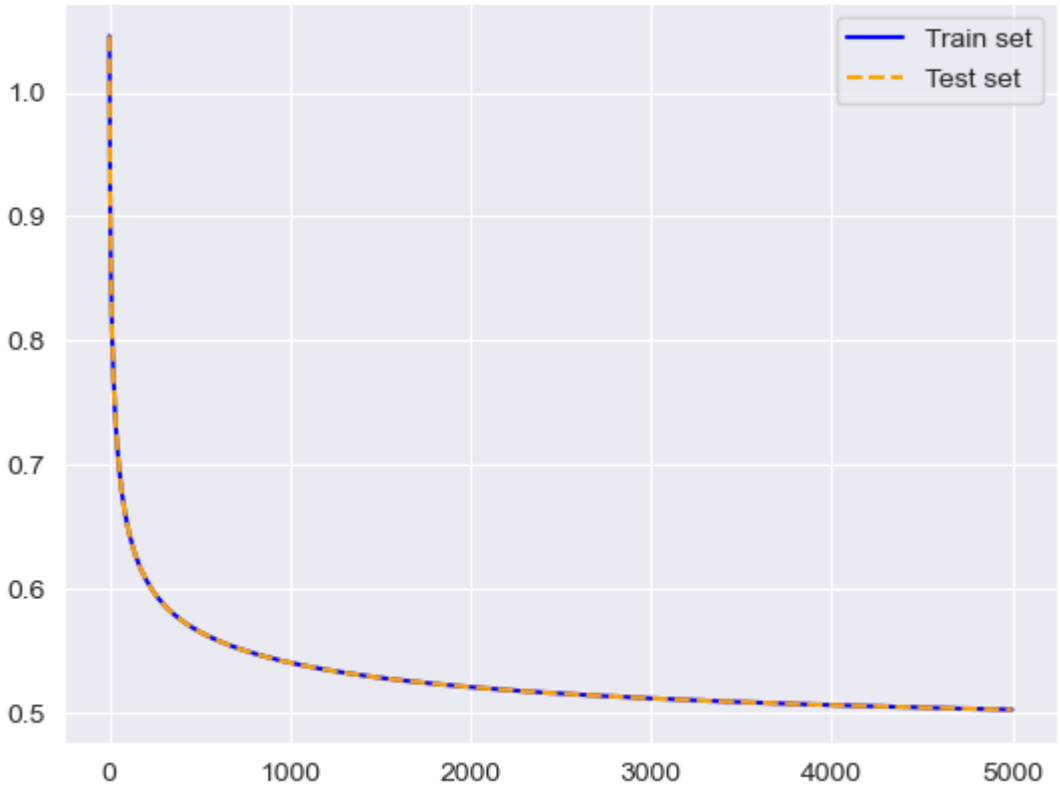
2. Wykresy zbieżności i analiza błędów

Aby stwierdzić czy nasz BaseLogisticRegression obarczony jest problemem overfittingu lub underfittingu, zostały przeprowadzone następujące treningi tego modelu zakończone następującymi wynikami:

accuracy(**train_set**)	0.8019
accuracy(**test_set**)	0.7695
accuracy(**CV**)	0.7719

Tabela 3 Wyniki predykcji dla modelu BaseLogisticRegression

Widać, że wyniki te są bardzo zbliżone do siebie, więc już na podstawie tych danych możemy stwierdzić, że żadne z wymienionych dwóch problemów nie występują w tym testowanym modelu. (W przypadku overfittingu wynik dla zbioru treningowego byłby znacznie większy niż dla zbioru testowego, w przypadku underfittingu obydwa wyniki byłby równo słabe i gorsze)



Rysunek 1 Koszt w zależności od zbioru i epoki

Również możemy zaobserwować ten brak problemów na wykresie kosztu w zależności od epoki:

Spadek błędu w czasie jest znaczący oraz płynny, nie ma, żadnego odchyłu w stronę dodatnią wartości błędu, oraz dwie krzywe praktycznie pokrywają się więc problem overfittingu oraz underfittingu nie występuje.

Wciąż aby usprawnić nasz model, możemy zwiększyć jego „elastyczność" poprzez rozszerzenie go o PolynomialFeatures (stopnia *n=2*):

accuracy(**train_set**)	0.8112
accuracy(**test_set**)	0.7718
accuracy(**CV**)	0.7723

Tabela 4 Wyniki predykcji dla modelu BaseLogisticRegression z PolynomialFeatures n=2

Porównując te wyniki można stwierdzić, że dokładność w małym stopniu wzrosła, więc metoda przyniosła pośrednie korzyści i będzie wykorzystywana we wszystkich dalszych usprawnieniach.

Od początku pracujemy na oryginalnym zbiorze danych opisanym w *części I,* czyli pracujemy na wszystkich kolumnach (cechach). Może to wprowadzić zbyteczną złożoność i wydłużyć czas uczenia modelu.
Po testach wyszły następujące wnioski, że możemy ograniczyć zbiór kolumn, które bierzemy pod uwagę, a wyniki predykcji nie będą gorsze, a nawet staną się minimalnie lepsze.

```
numerical_features_modified = [
    "Application order","Age at enrollment", "Curricular units 1st sem (credited)",
    "Curricular units 1st sem (enrolled)",
    "Curricular units 1st sem (evaluations)","Curricular units 1st sem (approved)","Curricular units 1st sem (grade)",
    "Curricular units 1st sem (without evaluations)","Curricular units 2nd sem (credited)",
    "Curricular units 2nd sem (enrolled)",
    "Curricular units 2nd sem (evaluations)","Curricular units 2nd sem (approved)",
    "Curricular units 2nd sem (without evaluations)","Curricular units 2nd sem (grade)"
]
categorical_features_modified = [
    "Marital status","Application mode","Course","Daytime/evening attendance","Previous qualification",
    "Mother's qualification","Father's qualification","Mother's occupation","Father's occupation","Displaced",
    "Educational special needs","Debtor","Tuition fees up to date","Gender","Scholarship holder"
]
```

Rysunek 2 Nowy zbiór cech

Wyniki predykcji:

accuracy(**first_fold**)	0.7839
accuracy(**second_fold)**	0.7672
accuracy(**third_fold**)	0.7858
CV accuracy	0.7700

Tabela 5 Wyniki predykcji dla modelu BaseLogisticRegression dla mniejszego zbioru cech wejściowych

3. Regularyzacja L1 i L2

Uwaga: Jeśli nie będzie zaznaczone inaczej to przyjmujemy następujące parametry modelów: (n_iters=5000, lr=0.001,batch_size=64, l =0.001)

Metody regularyzacji stosujemy, żeby mieć większą kontrolę nad wagami, które samodzielnie wylicza i ciągle zmienia uczący się model. Z racji tego, że wagi odpowiadają za określenie „ważności" danej cechy wejściowej, to poprzez manipulację możemy albo coś zrobić bardziej ważne, albo możemy coś całkowicie wyeliminować.

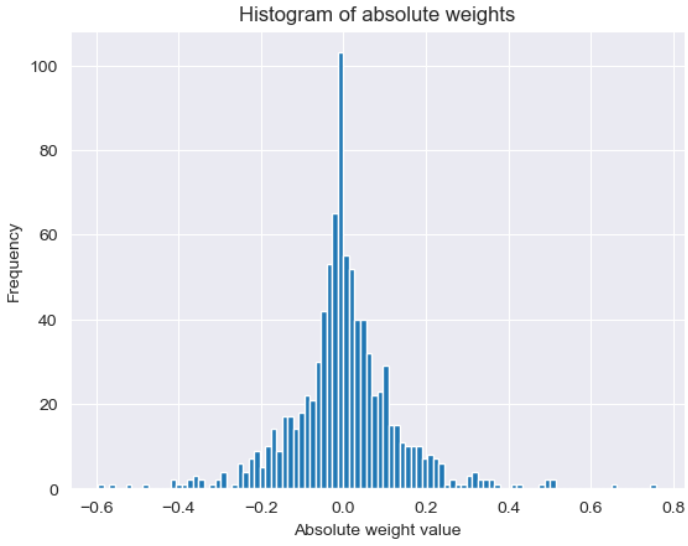
Pod uwagę będą brane modele podstawowe rozszerzone następująco:

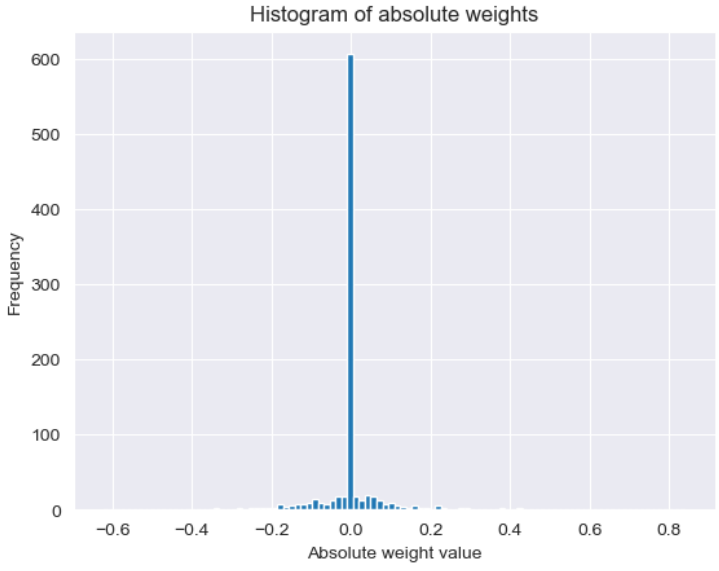
a) BaseLogisticRegressionL1 (lasso)
b) BaseLogisticRegressionL2 (ridge)
c) BaseLogisticRegressionCombined (L1+L2)

accuracy(**first_fold**)	0.7800
accuracy(**second_fold)**	0.7751
accuracy(**third_fold**)	0.7677
CV accuracy	0.7731

Tabela 6 Wyniki skuteczności dla modelu z regularyzacją L1

Wynik po regularyzacji L1 może sugerować, że już nie ma cech, które są warte usunięcia ze zbioru, po tym jak już raz ręcznie ten zbiór cech zmniejszaliśmy, dlatego skuteczność jest niezmieniona praktycznie.





Widzimy tutaj rozkład wag przed regularyzacją L1 oraz po regularyzacji, widać wyraźnie jak spadły wartości wag, oraz, że wag w okolicach 0 przybyło ogromnie dużo

accuracy(**first_fold**)	0.7722
accuracy(**second_fold)**	0.7891
accuracy(**third_fold**)	0.7834
CV accuracy	0.7811

Tabela 7 Wyniki skuteczności dla modelu z regularyzacją L2

Na obecną chwile jest to najlepszy wynik modelu, L2 mogło proporcjonalnie zbić wysokie wagi, przez co ustabilizowało niektóre, pozwalając na „uspokojenie i ustabilizowanie" predykcji, przez co ilość poprawnych predykcji wzrosła.

Możemy sprawdzić, jak wyglądają tablice wag przed i po regularyzacji:

Przed i po:

[[-8.97527571e-02 -4.49842253e-02 1.34736982e-01] [2.43500090e-01 -1.47497424e-01 -9.60026660e-02] [-2.55750105e-02 5.82487069e-02 **3.26736963e-02**] [**1.96075392e-01** 6.10108196e-02 -2.57086211e-01] [-9.70726496e-02 3.17800591e-01 -2.20727941e-01]	[[6.80675346e-02 -1.69309896e-01 1.01242361e-01] [1.89768116e-01 -9.86311071e-02 -9.11370091e-02] [-3.61905115e-02 6.02187619e-02 **-2.40282505e-02**] [**1.56276966e-03** 7.83625307e-02 -2.34639497e-01] [-1.26349846e-01 2.59835740e-01 -1.33485894e-01]

Widzimy, że w miejscach są widoczne mniejsze wagi po zastosowaniu regularyzacji L2

accuracy(**first_fold**)	0.7936
accuracy(**second_fold)**	0.7839
accuracy(**third_fold**)	0.7558
CV accuracy	0.7777

Tabela 8 Wyniki skuteczności dla modelu z regularyzacją L1+L2 (alpha=0.5)

Połączenie L1 L2 (elastic net) jest kompromisem między stabilizacją a utratą informacji poprzez zerowanie wag. Jednak w naszym przypadku nie przynosi znaczących lepszych efektów.

4. Usprawnienie danych - balansowanie zbiorów

Balansowanie zbioru danych jest ważne, aby przez element losowy model nie uczył się na danych gdzie jedna klasa występuję więcej razy niż inne i przez to żeby nie była błędnie promowana.

Do balansowania zbioru danych wykorzystane zostaną dwie techniki:

a) SMOTENC (wariancja SMOTE ale też dla cech kategorialnych)
b) RandomUnderSampler

A test będzie przeprowadzany na modelu BaseLogisticRegressionRegL2

	Precision	Recall	F1-score	Support
0	0.8113	0.7800	0.7954	441
1	0.5500	0.3592	0.4346	245
2	0.7917	0.9174	0.8499	642
Accuracy			**0.7688**	1328
Macro avg	0.7177	0.6856	0.6933	1328
Weighted avg	0.7536	0.7688	0.7552	1328

Tabela 9 Raport bez sampling

	Precision	Recall	F1-score	Support
0	0.8406	0.7415	0.7880	441
1	0.4533	0.5551	0.4991	245
2	0.8372	0.8333	0.8353	642
Accuracy			**0.7515**	1328
Macro avg	0.7104	0.7100	0.7074	1328
Weighted avg	0.7675	0.7515	0.7575	1328

Tabela 10 Raport dla sampling-SMOTE

	Precision	Recall	F1-score	Support
0	0.8464	0.7120	0.7734	441
1	0.4448	0.6245	0.5195	245
2	0.8450	0.8069	0.8255	642
Accuracy			**0.7417**	1328
Macro avg	0.7121	0.7145	0.7061	1328
Weighted avg	0.7716	0.7417	0.7517	1328

Tabela 11 Raport dla sampling-RandomUnderSampler

Wnioski:

Wyniki CV dla każdej metody sampling'u są znacząco niższe niż przed samplingiem.

Widać, że kiedy używamy SMOTE albo RandomUnderSampler to nasze wyniki skuteczności jednoznacznie spadają. Jest to spowodowane wyrównaniem „szans" na predykcję wszystkich klas, a dokładnie:

Model, przed zastosowaniem samplingu widocznie „ignoruje" klasę 1, (widać to na współczynniku recall), skupiając się między klasą 0 i 2,ale jak widać przynosi mu to najlepsze efekty w postaci najwyższego wyniku skuteczności. Jednak w przypadku zbiorów danych, gdzie znaczenie nawet jednego poprawnego przypisania jest bardzo ważne, nie możemy sobie pozwolić na takie ignorowanie jednej klasy.

Model po zastosowaniu SMOTE ma już bardziej wyrównany współczynnik recall i powiązaną z tym spadek skuteczności modelu. Podobnie jest po zastosowaniu RandomUnderSampler, ale w tym przypadki recall jest jeszcze bardziej równy.

5. Optymalizacja hiperparametrów

Modele, które staramy się zoptymalizować posiadają w sobie parametry, które są stałe przez cały cykl uczenia się modelu, i to użytkownik obsługujący model musi je samodzielnie ustalić. Parametry, które będziemy optymalizować:

a) BaseLogisticRegression: *n_iters, lr, batch_size*
b) BaseLogisticRegressionRegL2: *n_iters, lr, batch_size,l*
c) BaseLogisticRegressionRegCombined: *n_iters, lr, batch_size,l, alpha*

Żeby stwierdzić jakie wartości będą najlepsze dla danego modelu najprostszym sposobem jest sprawdzenie wszystkich kombinacji wybranych wartości.

n_iters: [3000,4000,5000,6000,7000,8000,9000,10000] ***lr:*** [0.001,0.0001,0.00001] ***batch_size:*** [32,64,128]		
	n_iters	7000
	lr	0.001
	batch_size	64
	value	0.7701
	TIME	2h36min

Tabela 12 Wyniki poszukiwania hiperparametrów dla BaseLogisticRegression

n_iters:[4000,5000,6000], ***l***r:[0.001,0.0001,0.00001], ***batch_size***: [64,128], ***l*** :[0.001,0.0001,0.00001],		
	n_iters	5000
	lr	0.001
	batch_size	64
	l	0.00001
	value	0.7755
	TIME	46min

Tabela 13 Wyniki poszukiwania hiperparametrów dla BaseLogisticRegressionRegL2

n_iters:[4000,5000,6000], ***lr***:[0.001,0.0001,0.00001], ***batch_size***:[64,128], ***l***:[0.001,0.0001,0.00001], ***alpha***: [0.5,0.6,0.7,0.8,0.9]		
	n_iters	4000
	lr	0.001
	batch_size	128
	l	0.001
	alpha	0.5
	value	0.7784
	TIME	4h15min

Tabela 14 Wyniki poszukiwania hiperparametrów dla BaseLogisticRegressionRegCombined

Dalej trzeba mieć na uwadze element losowy uruchomienia modelu, więc może się zdarzyć, że lekko inne parametry również po CV będą miały lepszy wynik, co widać porównując poprzednie wyniki CV na lekko innych hiperparametrach.

*** Dodatkowo optymalizacja hiperparametrów dla modeli nie wchodzących w skład oceny:

n_estimators:[100,200,300], ***max_depth***:[None,10,20], ***min_samples_split***:[2,5,10], ***min_samples_leaf***:[1,2,4],		
	n_estimators	200
	max_depth	20
	min_samples_split	2
	min_samples_leaf	1
	value	0.7742
	TIME	4min

Tabela 15 Wyniki poszukiwania hiperparametrów dla RandomForest

6. Metody ensemble

Metody ensemble mają za zadanie sprawić, że kilka modelów będzie uzupełniać się w celu uzyskania końcowo lepszego wyniku predykcji.

Wykorzystywane metody ensemble:

a) HardVotingClassifier
b) StackingClassifier

Przed wykonaniem uczenia z użyciem tych technik trzeba było wybrać modele składowe. Idealnym podejściem jest kiedy modele popełniają błędy w innych miejscach:
a) BaseLogisticRegression
b) BaseLogisticRegressionCombined

To one po stworzeniu macierzy korelacji błędów okazały się mieć najniższy wynik (lecz jednak nadal wysoki)

Otrzymane wyniki są porównywalne, a nawet gorsze od podstawowych metod bez optymalizacji, wynikać to może właśnie z tego, że modele składowe są do siebie zbyt implementacyjnie podobne, i popełniają wspólne wspólnie błędu na wspólnych obszarach

Metoda:	**Accuracy:**
HardVotingClassifier	0.7764
StackingClassifier	*0.7445*

Tabela 16 Wyniki CV dla ensemble metod

7. Mixture of Experts

W tej metodzie modele, które są nazywane ekspertami w czasie treningu, zyskują zaufanie co do poszczególnych rekordów, następnie na podstawie tego zaufania trenuje się wynikowy model, który na podstawie różnych wejść wybiera to co wyliczył już lepszy model dla danego wejścia.
W MoE określamy gating model - RandomForestClassifier
Jednak tutaj problem jest również podobny, modele składowe zbyt mało różnią się od siebie w obszarach gdzie popełniają błędy, dlatego końcowa skuteczność waha się w okolicy standardowej **0.7673**

8. Podsumowanie wyników

Spośród wszystkich modeli i ich wariancji, które zostały stworzone podczas wykonywania optymalizacji modeli uczenia maszynowego, można sporządzić następujące zestawienie:

Wszystkie przedstawione wyniki otrzymane zostały po *cross-validation[k=3]*

	BaseLogisticRegression	BaseLogisticRegressionRegL1	BaseLogisticRegressionRegL2	BaseLogisticRegressionRegL1L2	EnsembleVotingClassifier	EnsebleStackingClassifier	MoE
Opis:	iters = 5000 lr = 0.001 batch = 64	iters = 5000 lr=0.001 batch=64 l=0.00001 +smaller dataset +PolynomialFeatures	iters = 5000 lr =0.001 batch =64 l =0.001 +smaller dataset +PolynomialFeatures	Iters = 4000 lr = 0.001 batch = 128 l=0.001 alpha=0.5 +smaller dataset +PolynomialFeatures	BaseLogisticRegressionRegL1L2 BaseLogisticRegression	BaseLogisticRegressionRegL1L2 BaseLogisticRegression	BaseLogisticRegressionRegL1L2 BaseLogisticRegression
Skuteczność	0.7719	0.7755	**0.7811**	0.7784	0.7764	0.74450	0.7673

Tabela 17 Podsumowanie wyników wytrenowanych modeli

Najlepszy model: Najlepszym wytrenowanym i przetestowanym modelem został model **BaseLogisticRegressionRegL2.** Czyli pierwotna wersja wieloklasowej regresji logistycznej wzbogacona o trenowanie na ograniczonym zbiorze danych wraz z zwiększoną złożonością przez PolynomialFeatures oraz z wprowadzoną regularyzacją (L2 ridge) – jej najlepszy wynik osiągnął ***78,11%*** skuteczności predykcji. Zastosowanie tej metody regularazji sprawiło, że wagi zostały prawidłowo ustabilizowane podczas procesu uczenia, oraz zapobiegła przeuczenia modelu poprzez zbyt duże przywiązania do wag, jednocześnie nie usuwając żadnych informacji (L2 nie doprowadza do zerowania wag przy elementach wejściowych)

Metody, które miały gratyfikować wiele pracujących modeli na raz nie sprawdziły się najlepiej ze względu na zbyt duże ich podobieństwo w popełnianiu błędów.