Pro1_D_projekt 23.01.2024
Przewidywanie Skali Przestępczości

Autor: Jakub Boczar s23584

Spis treści

1. Cel projektu

Głównym celem badania tego zbioru danych jest zrozumienie zależności między różnymi czynnikami społecznymi a poziomem przestępczości w społecznościach w Stanach Zjednoczonych. Analiza tego zestawu danych może pomóc w identyfikowaniu czynników wpływających na przestępczość oraz w opracowywaniu strategii prewencyjnych.

2. Opis zbioru danych UCI

Zródło danych: https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

Zestaw danych zawiera wszystkie niezbędne informacje do trenowania i testowania modelu. Niemniej jednak, twórcy zbioru danych przefiltrowali dane, które nie były związane z przestępczością. Początkowo, zbiór danych obejmował 122 atrybuty przeznaczone do prognozowania, 5 dodatkowych atrybutów, które nie miały być wykorzystywane do prognoz, oraz kluczowy atrybut decyzyjny - Per Capita Violent Crimes (ViolentCrimesPerPop). Całkowita liczba rekordów w zbiorze wynosi 1994. Dane zawarte w zestawie obejmują informacje takie jak mediana zarobków rodzinnych, liczba interwencji policji oraz stosunek liczby policjantów do ogólnej populacji.

Wszystkie dane przygotowane przez twórców zbioru zostały już znormalizowane w dyskretnym zakresie od 0.00 do 1.00, przy zachowaniu odpowiedniej dystrybucji danych. Jednak dla wartości skrajnych (wszystkie wartości trzykrotności standardowej dewiacji powyżej średniej są normalizowane do 1.00; wszystkie wartości trzykrotności standardowej dewiacji poniżej średniej są normalizowane do 0.00).

Atrybuty:

@attribute state numeric @attribute county numeric @attribute community numeric @attribute communityname string @attribute fold numeric @attribute population numeric @attribute householdsize numeric @attribute racepctblack numeric @attribute racePctWhite numeric @attribute racePctAsian numeric @attribute racePctHisp numeric @attribute agePct12t21 numeric @attribute agePct12t29 numeric @attribute agePct16t24 numeric @attribute agePct65up numeric @attribute numbUrban numeric @attribute pctUrban numeric @attribute medIncome numeric @attribute pctWWage numeric @attribute pctWFarmSelf numeric @attribute pctWInvInc numeric @attribute pctWSocSec numeric @attribute pctWPubAsst numeric @attribute pctWRetire numeric @attribute medFamInc numeric @attribute perCapInc numeric @attribute whitePerCap numeric @attribute blackPerCap numeric @attribute indianPerCap numeric @attribute AsianPerCap numeric @attribute OtherPerCap numeric @attribute HispPerCap numeric @attribute NumUnderPov numeric @attribute PctPopUnderPov numeric @attribute PctLess9thGrade numeric @attribute PctNotHSGrad numeric @attribute PctBSorMore numeric @attribute PctUnemployed numeric @attribute PctEmploy numeric @attribute PctEmplManu numeric @attribute PctEmplProfServ numeric @attribute PctOccupManu numeric @attribute PctOccupMgmtProf numeric @attribute MalePctDivorce numeric @attribute MalePctNevMarr numeric @attribute FemalePctDiv numeric @attribute TotalPctDiv numeric @attribute PersPerFam numeric @attribute PctFam2Par numeric @attribute PctKids2Par numeric @attribute PctYoungKids2Par numeric @attribute PctTeen2Par numeric @attribute PctWorkMomYoungKids numeric @attribute PctWorkMom numeric @attribute NumIlleg numeric @attribute PctIlleg numeric @attribute NumImmig numeric @attribute PctImmigRecent numeric @attribute PctImmigRec5 numeric @attribute PctImmigRec8 numeric @attribute PctImmigRec10 numeric @attribute PctRecentImmig numeric @attribute PctRecImmig5 numeric @attribute PctRecImmig8 numeric @attribute PctRecImmig10 numeric @attribute PctSpeakEnglOnly numeric @attribute PctNotSpeakEnglWell numeric @attribute PctLargHouseFam numeric @attribute PctLargHouseOccup numeric @attribute PersPerOccupHous numeric @attribute PersPerOwnOccHous numeric @attribute PersPerRentOccHous numeric @attribute PctPersOwnOccup numeric @attribute PctPersDenseHous numeric @attribute PctHousLess3BR numeric @attribute MedNumBR numeric @attribute HousVacant numeric @attribute PctHousOccup numeric @attribute PctHousOwnOcc numeric @attribute PctVacantBoarded numeric @attribute PctVacMore6Mos numeric @attribute MedYrHousBuilt numeric @attribute PctHousNoPhone numeric @attribute PctWOFullPlumb numeric @attribute OwnOccLowQuart numeric @attribute OwnOccMedVal numeric @attribute OwnOccHiQuart numeric @attribute RentLowQ numeric @attribute RentMedian numeric @attribute RentHighQ numeric @attribute MedRent numeric @attribute MedRentPctHousInc numeric @attribute MedOwnCostPctInc numeric @attribute MedOwnCostPctIncNoMtg numeric @attribute NumInShelters numeric @attribute NumStreet numeric @attribute PctForeignBorn numeric @attribute PctBornSameState numeric @attribute PctSameHouse85 numeric @attribute PctSameCity85 numeric @attribute PctSameState85 numeric @attribute LemasSwornFT numeric @attribute LemasSwFTPerPop numeric @attribute LemasSwFTFieldOps numeric @attribute LemasSwFTFieldPerPop numeric @attribute LemasTotalReq numeric @attribute LemasTotReqPerPop numeric @attribute PolicReqPerOffic numeric @attribute PolicPerPop numeric @attribute RacialMatchCommPol numeric @attribute PctPolicWhite numeric @attribute PctPolicBlack numeric @attribute PctPolicHisp numeric @attribute PctPolicAsian numeric @attribute PctPolicMinor numeric @attribute OfficAssgnDrugUnits numeric @attribute NumKindsDrugsSeiz numeric @attribute PolicAveOTWorked numeric @attribute LandArea numeric @attribute PopDens numeric @attribute PctUsePubTrans numeric @attribute PolicCars numeric @attribute PolicOperBudg numeric @attribute LemasPctPolicOnPatr numeric @attribute LemasGangUnitDeploy numeric @attribute LemasPctOfficDrugUn numeric @attribute PolicBudgPerPop numeric @attribute ViolentCrimesPerPop numeric

Atrybutem decyzyjnym jest w tym przypadku ViolentCrimesPerPop

3. Użyte Biblioteki

1. Numpy

Numpy to biblioteka do obliczeń numerycznych w języku Python. Zapewnia efektywną obsługę dużych macierzy i tablic oraz zawiera funkcje matematyczne do operacji na nich. Numpy jest kluczowy dla pracy z danymi numerycznymi, obliczeń naukowych i analizy danych.

2. Pandas

Pandas to biblioteka umożliwiająca łatwą i efektywną manipulację danymi w języku Python. Dostarcza struktury danych, takie jak DataFrame, które ułatwiają import, eksplorację, przetwarzanie i analizę danych. Idealna do pracy z danymi tabelarycznymi.

3. Matplotlib

Matplotlib to potężna biblioteka wizualizacji danych w języku Python. Dostarcza narzędzi do tworzenia różnorodnych wykresów, diagramów i map. Biblioteka ta jest wszechstronna i elastyczna, pozwalając użytkownikom dokładnie dostosować wygląd i styl generowanych grafik.

4. Seaborn

Seaborn to narzędzie do wizualizacji danych, które działa na bazie Matplotlib. Posiada wbudowane motywy estetyczne i funkcje do tworzenia atrakcyjnych wykresów statystycznych. Seaborn ułatwia generowanie skomplikowanych wizualizacji z minimalnym wysiłkiem.

5. Sklearn

Scikit-learn to biblioteka do uczenia maszynowego w Pythonie. Oferuje narzędzia do klasyfikacji, regresji, grupowania, redukcji wymiarowości, wydajnego podziału danych i oceny modeli. Sklearn jest przyjazny dla użytkownika, ma dobrą dokumentację i szeroki zakres algorytmów.

4. Metodologia

1. Przetworzenie danych

2. Normalizacja

Dataset był już znormalizowany od samego początku

3. Wybieram atrybuty

Z racji ogromniej ilości atrybutów należało je przefiltrować i ograniczyć do mniejszej ilości. Z tego też powodu użyta zostałą

4. Podział na zbiory treningowe i testowe

Zbiór treningowy: 75% zbioru danych. • Zbiór testowy: 25% (reszta)

5. Modele użyte w projekcie

W projekcie skorzystałem z 4 klasyfikatorów regresyjnych. Przetestowana została ich skuteczność na wcześniej przygotowanych zbiorach danych.

1. Sieć neuronowa

Sieć neuronowa to model matematyczny inspirowany strukturą mózgu, złożony z połączonych ze sobą jednostek zwanych neuronami. Neurony są ułożone w warstwach, a każda z nich przekazuje sygnał do kolejnej warstwy, aż do osiągnięcia warstwy wyjściowej, gdzie uzyskiwany jest wynik.

2. Drzewo regresyjne

Drzewo regresyjne to model predykcyjny, który dzieli zestaw danych na podzbiory, a następnie stosuje regresję do każdego z podzbiorów. Każdy węzeł drzewa reprezentuje decyzję, a liście zawierają prognozy.

3. Regresja liniowa

Regresja liniowa to technika statystyczna służąca do modelowania związku pomiędzy jedną zmienną zależną a jedną lub więcej zmiennymi niezależnymi. Zakłada się, że związek ten można opisać liniową funkcją.

4. Regresja nieliniowa

Regresja nieliniowa to technika modelowania związku pomiędzy zmienną zależną a jedną lub więcej zmiennymi niezależnymi, przy czym zakłada się, że związek ten jest nieliniowy. Może przyjmować różne formy funkcji, nie tylko liniowe.

6. Metody oceniania jakości modelu w projekcie

1. Mean Squared Error (MSE):

Użycie: Ocenia średnią kwadratową różnicę między rzeczywistymi a przewidywanymi wartościami.

Interpretacja: Im niższa wartość MSE, tym lepiej. Oznacza to mniejsze błędy kwadratowe między rzeczywistymi a przewidywanymi wartościami.

2. R^2 Score (Coefficient of Determination):

Użycie: Mierzy, jak dobrze model dostosowuje się do danych, porównując go do średniej wartości celu.

Interpretacja: R^2 wynoszące 1 oznacza idealne dopasowanie, a 0 oznacza, że model nie jest lepszy niż przewidywanie średniej. Im bliżej 1, tym lepiej.

3. Cross-Validation Mean Score MSE:

Użycie: Średni błąd kwadratowy uzyskany w wyniku walidacji krzyżowej.

Interpretacja: Daje informacje o ogólnej skuteczności modelu na różnych podzbiorach danych.

4. Cross-Validation Mean Score R^2:

Użycie: Średni współczynnik determinacji uzyskany w wyniku walidacji krzyżowej.

Interpretacja: Podobnie jak R^2, ocenia, jak dobrze model generalizuje się do nowych danych.

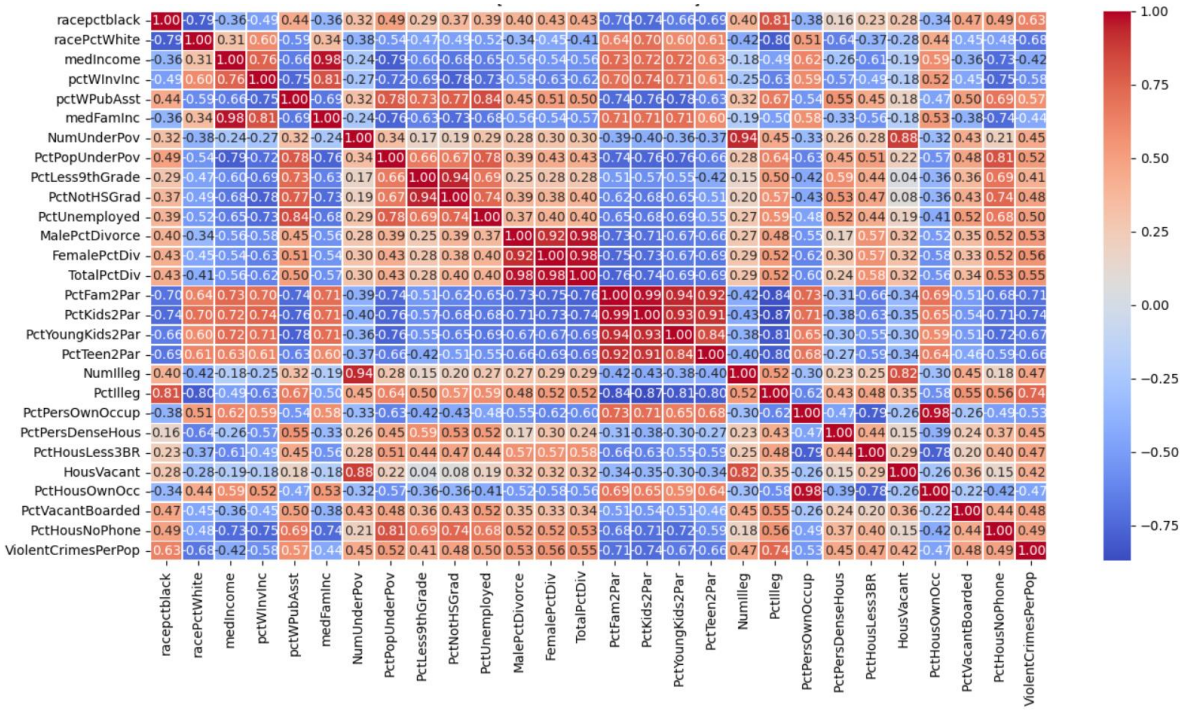
7. Rezultaty

Po przefiltrowaniu wszystkich atrybutów za pomocą mojej funkcji feature_selector_corelation wybrałem 27 atrybutów, które mają odpowiedni w moim przekonaniu poziom korelacji i przekraczczają poziom korelacji = 0,4.
Lista wybranych atrybutów:

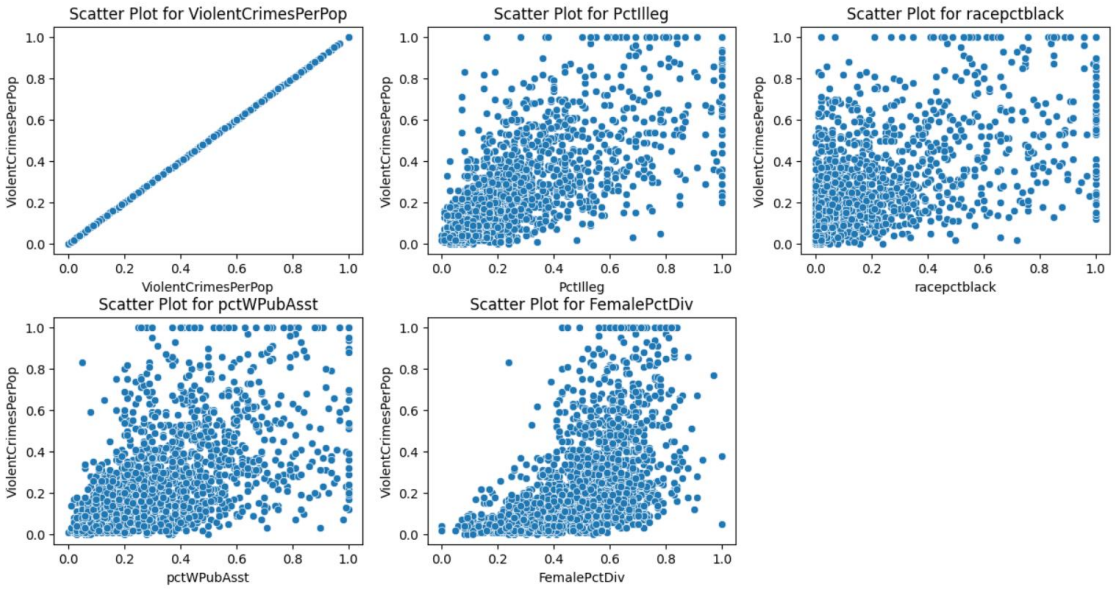
- racepctblack
- racePctWhite
- medIncome
- pctWInvInc
- pctWPubAsst
- medFamInc
- NumUnderPov

- PctPopUnderPov
- PctLess9thGrade
- PctNotHSGrad
- PctUnemployed
- MalePctDivorce
- FemalePctDiv
- TotalPctDiv
- PctFam2Par
- PctKids2Par
- PctYoungKids2Par
- PctTeen2Par
- NumIlleg
- PctIlleg
- PctPersOwnOccup
- PctPersDenseHous
- PctHousLess3BR
- HousVacant
- PctHousOwnOcc
- PctVacantBoarded
- PctHousNoPhone

Dodatkowo stworzyłem heatmapę która obrazuje poziom korelacji miedzy poszczególnymi atrybutami:



Rys.1 Matrix korelacji miedzy atrybutami



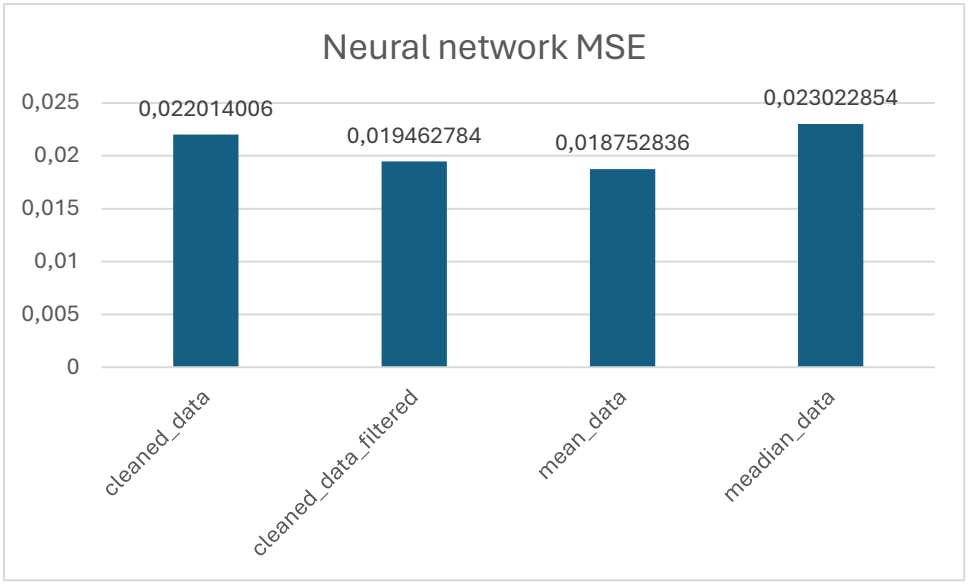
Rys 2. Scatter 5 atrybutów z największa korelacja

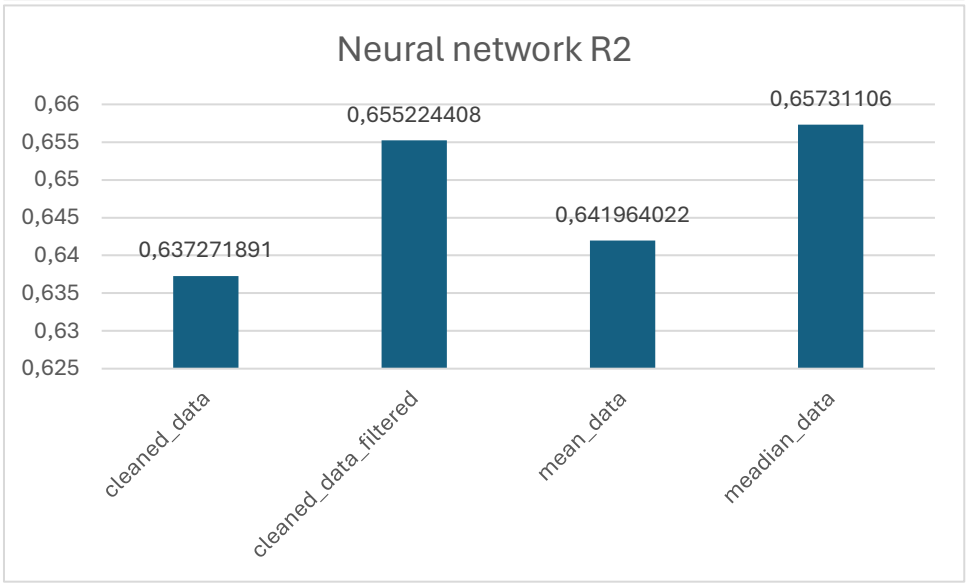
Na podstawie tych danych można określić jakie atrybuty mają największy wpływ na wysokość przestępczości w danych miejscach.

8. Wyniki Programu:

We wszystkich przypadkach zastosowano miedzy innymi 10 krotną walidacje krzyżową

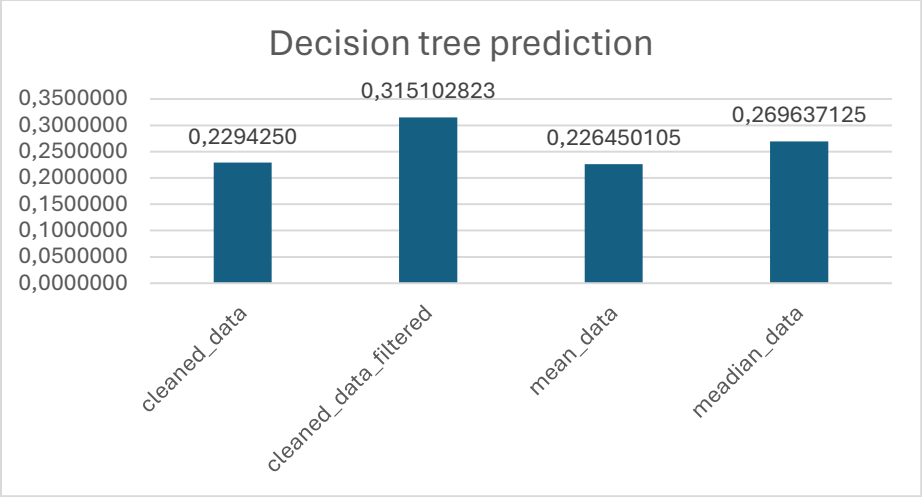
1. Dla sieci neuronowej:

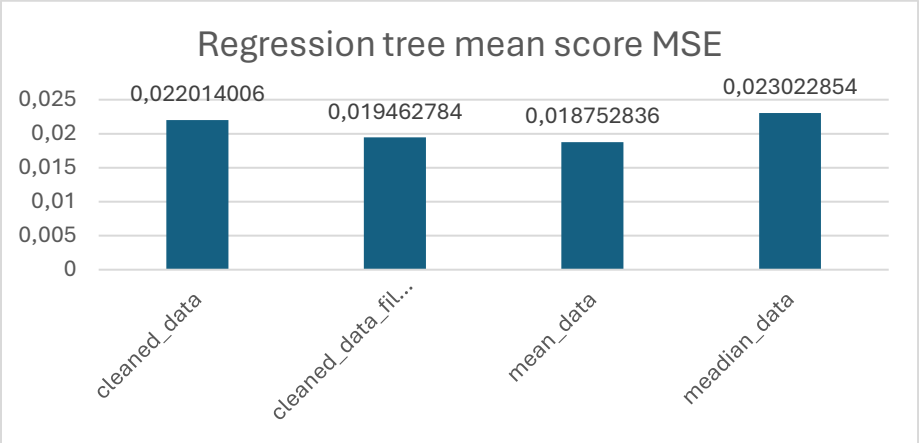




W sieci neuronowej najlepsze są zbiory danych z mediana median_data i z przefiltrowaniem cleared data filtered.
w wypadku sieci najlepiej sprawdził się zbiór median biorac pod uwagę R2 i mse.

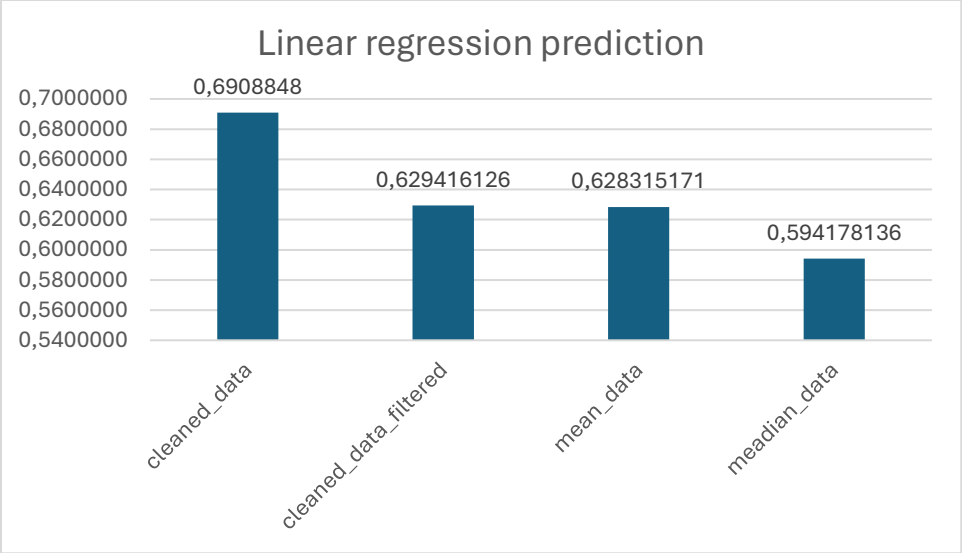
2. Dla drzewa regresyjnego

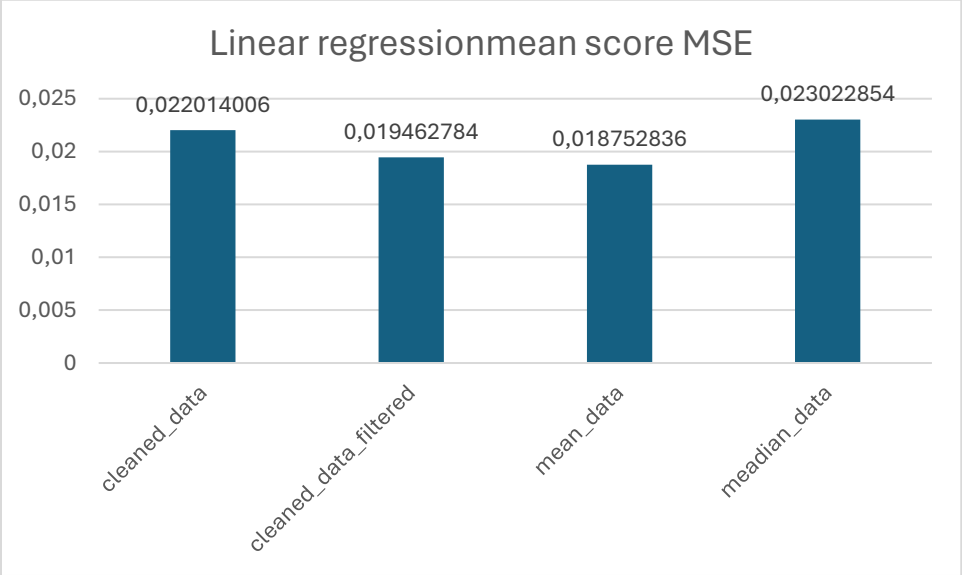


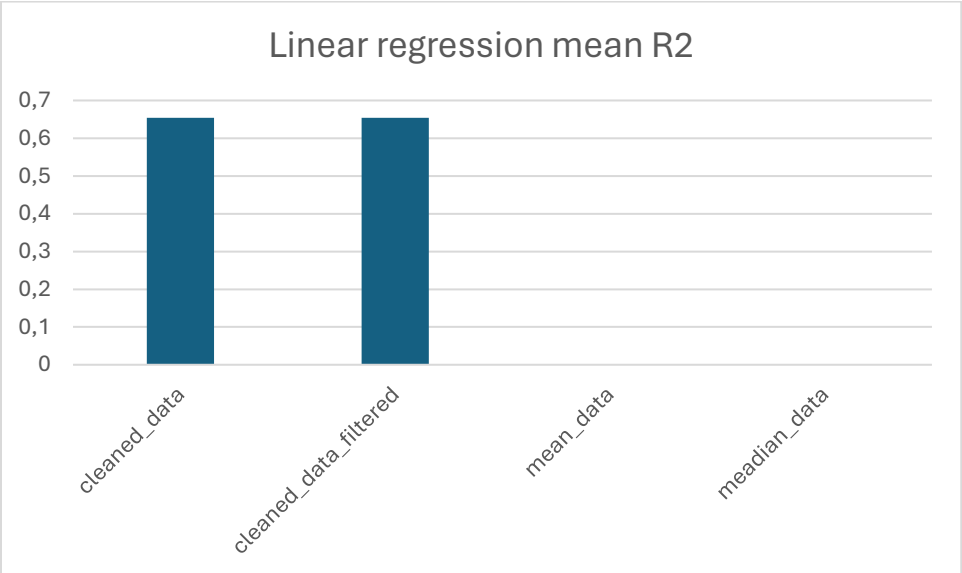


W drzewie zaobserwować można zdecydowana wygraną Cleared_data_filtered i jest to najlepsze rozwiązanie.

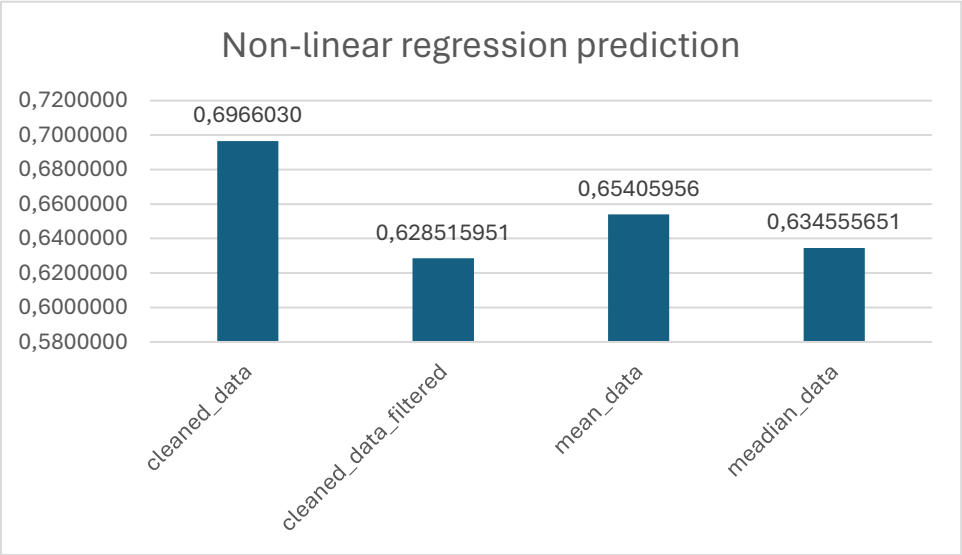
3. Dla regresji liniowej

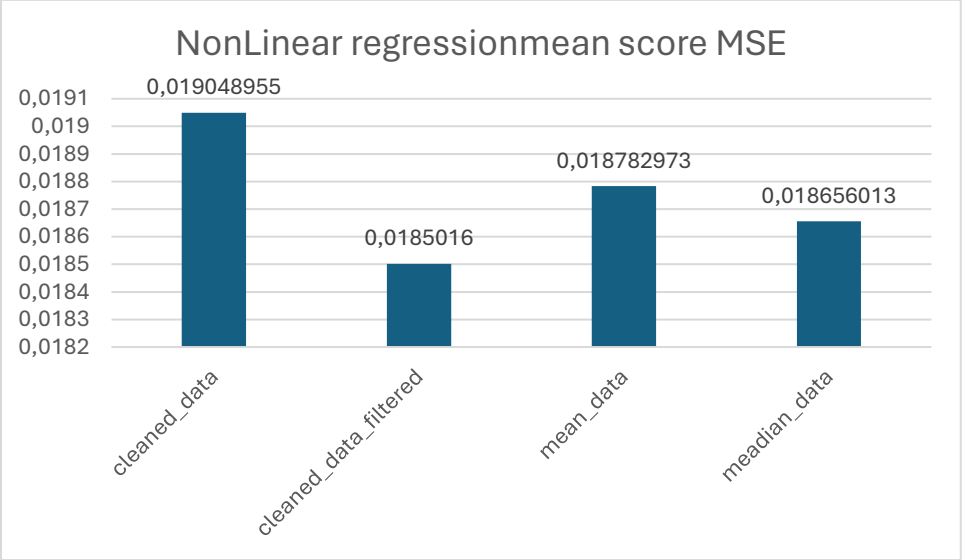


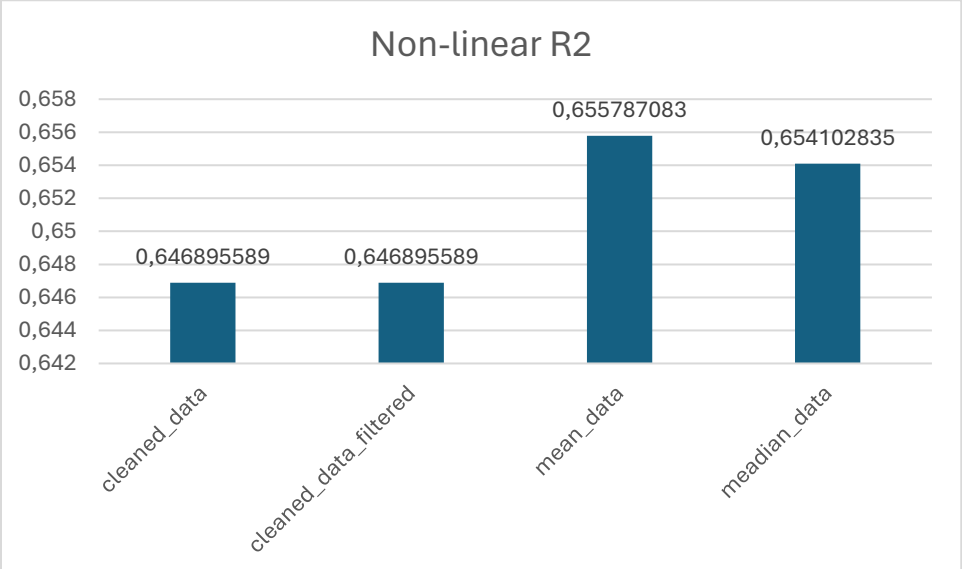




4. Dla regresji nieliniowej

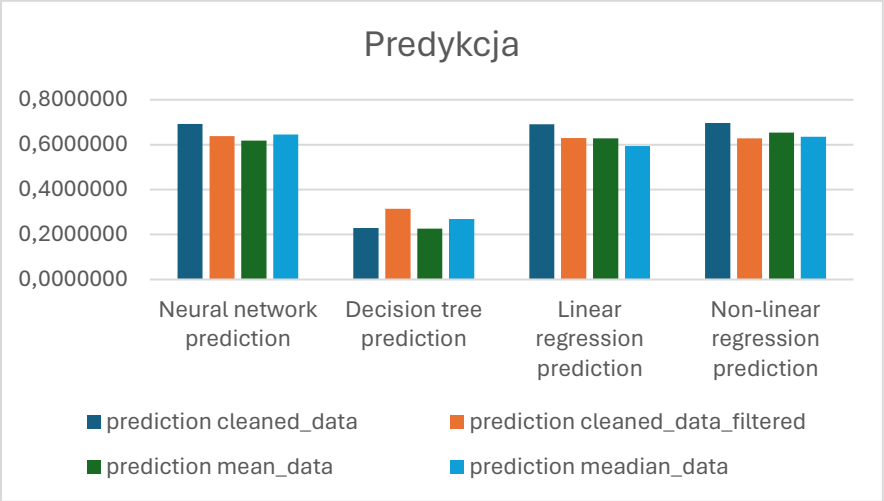


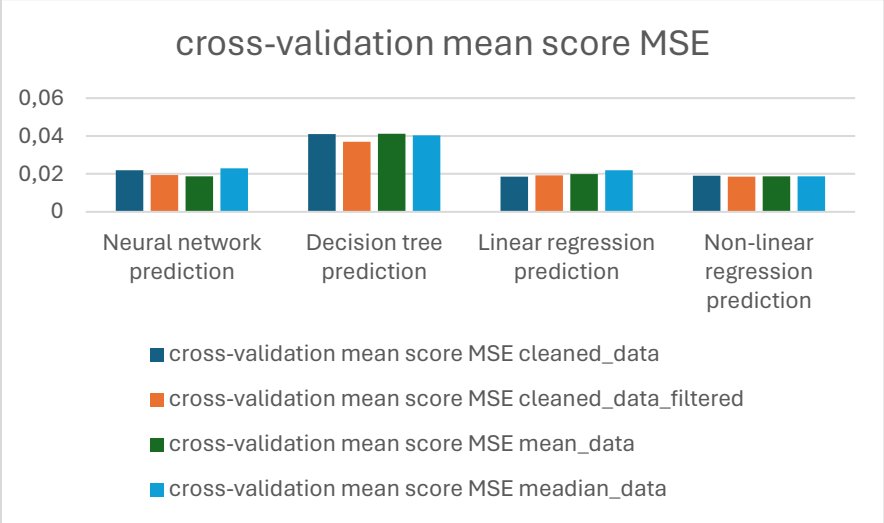




Dla regresji nieliniowej otrzymaliśmy najlepsze wyniki dla cleared_data, natomiast na drugim miejscu znalazło się mean data

9. Porównanie Wyników





Na podstawie powyższych wyników możemy zauważyć ze drzewa regresyjne poradziły sobie w tym wypadku najgorzej, natomiast najbardziej czułe były sieci neuronowe.

W ten sposób możemy stwierdzić że drzewa regrezyjne nie są najlepszym roziązaniem w tym wypadku. Wynik mogłby być inny w wypadku lasu random forest jednak jest to temat na dalsze badania.