****

**MINISTERUL EDUCAȚIEI ȘI CERCETĂRII**

**AL REPUBLICII MOLDOVA**

**Universitatea Tehnică a Moldovei**

**Facultatea Calculatoare, Informatică și Microelectronică**

**Departamentul Informatică și Ingineria Sistemelor**

**Barîmb Eusebiu**

**Raport**

**pentru lucrarea de laborator**

***la cursul de “Medii Virtuale pentru Știința Datelor”***

**Chișinău – 2023**

**CUPRINS**

[1. O INTRODUCERE ÎN ÎNVĂȚAREA AUTOMATĂ CU SCIKIT-LEARN 3](#_Toc116852105)

[2. NORMLAIZAREA DATELOR 4](#_Toc116852106)

[3. MINMAX NORMALIZAREA 5](#_Toc116852107)

[4. DATA VISUALIZATION 6](#_Toc116852108)

[5. REDUCEREA ȘI VIZUALIZAREA DATELOR 10](#_Toc116852109)

[CONCLUZII 11](#_Toc116852110)

[BIBLOGRAFIE 12](#_Toc116852111)

INTRODUCERE

Una dintre caracteristicile esențiale ale oricărei analize imobiliare este înțelegerea valorii medii a locuințelor. În setul de date privind locuințele din Boston, această valoare este reprezentată de coloana **MEDV,** care denotă valoarea mediană a caselor ocupate de proprietari, exprimată în mii de dolari. Prezicerea acestei valori nu este doar o provocare academică, ci și o aplicare practică esențială în domeniul imobiliar și al finanțelor.

Regresia, fiind una dintre tehnicile fundamentale în analiza de date, ne va servi ca punct de plecare în acest proiect. Modelul de regresie urmărește să stabilească o relație între variabilele independente, cum ar fi rata criminalității sau numărul mediu de camere pe locuință, și variabila dependentă - în cazul nostru, valoarea mediană a locuinței (MEDV). Prin înțelegerea acestei relații, putem face previziuni precise și informate despre valoarea locuințelor în funcție de diferiți factori.

Este important de menționat că, în era digitală de astăzi, există o mulțime de platforme și instrumente care permit analiza și modelarea datelor. În cadrul acestui proiect, vom explora capacitatea mai multor platforme, precum IBM SPSS, KNIME, Python și Excel, în crearea de modele de inteligență artificială. Fiecare dintre aceste platforme are propriile sale avantaje și particularități, oferindu-ne o gamă largă de tehnici și metode pentru a aborda setul de date din Boston într-un mod cuprinzător.

* 1. DESCRIEREA MODELULUI

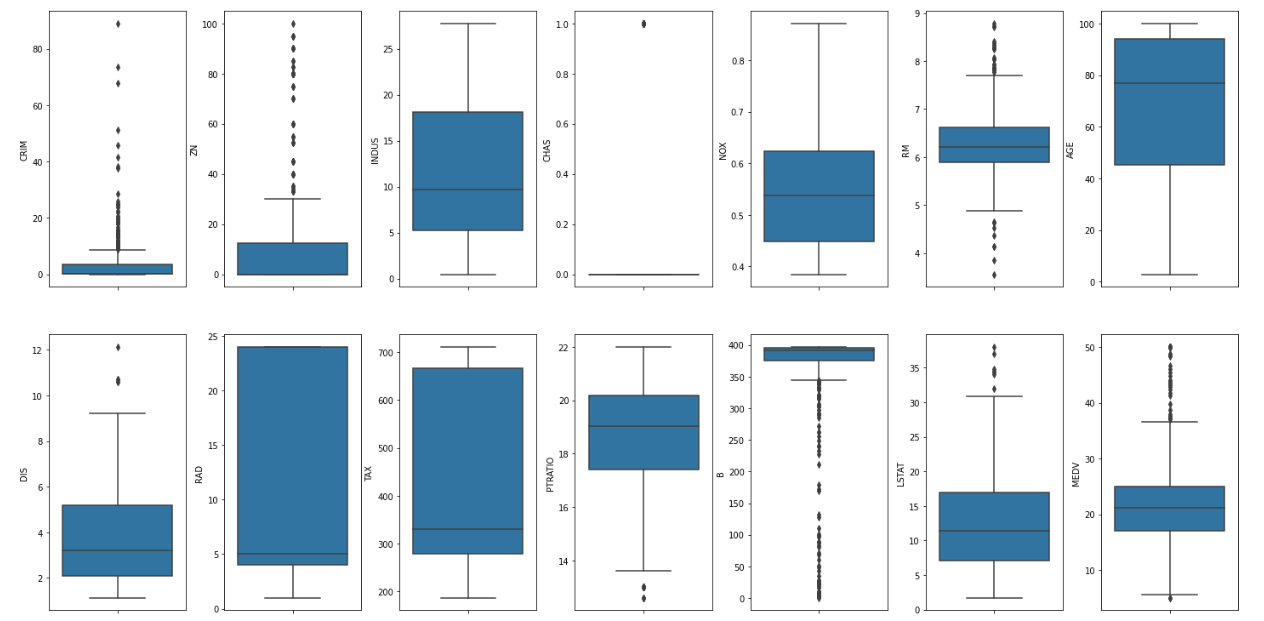
Setul de date pe care îl analizăm reprezintă un eșantion din piața imobiliară a orașului Boston, bazat pe datele colectate de Serviciul de Recensământ al SUA. Acesta constă din mai multe coloane ce caracterizează diferiți factori socio-economici, demografici și geografici care ar putea influența valoarea unei locuințe.

Iată o scurtă privire asupra primelor intrărilor:

1. CRIM variază de la valori foarte mici, cum ar fi 0,00632, până la valori mai mari precum 0,21124, indicând rata per capita a criminalității pe oraș.
2. ZN indică proporția terenului rezidențial zonificat pentru loturi mari, unde vedem valori precum 18.0 și 12.5, dar și multe intrări cu 0.0.
3. INDUS prezintă proporții variate ale afacerilor non-retail pe oraș, cu valori ca 2.31 sau 7.87.
4. CHAS este o variabilă dummy pentru râul Charles, unde o valoare de 1 indică apropiere de râu, dar majoritatea intrărilor sunt 0, cu excepția unor valori lipsă (NaN).
5. NOX prezintă concentrații ale oxidului de azot, cu variații între 0,458 și 0,538.
6. RM oferă o idee despre numărul mediu de camere pe locuință, cu valori cuprinse între 5,631 și 7,185.
7. AGE arată proporția locuințelor ocupate de proprietar construite înainte de 1940, variind de la 45,8 la 100.
8. DIS prezintă distanțe ponderate până la cinci centre de angajare din Boston, cu valori cuprinse între 4,0900 și 6,5921.
9. RAD și TAX variază în funcție de accesibilitatea la autostrăzi și rata impozitului pe proprietate.
10. PTRATIO prezintă variații ale raportului elev-profesor în diferite orașe. B este o transformare a proporției de populație de culoare, cu valori apropiate de 400 în multe intrări.
11. LSTAT și MEDV arată procentul populației cu un statut inferior și, respectiv, valoarea medie a locuințelor. Este important de remarcat că există valori lipsă (NaN) în setul de date, în special în coloanele CHAS și LSTAT.
    1. IMPLEMENTAREA REGRESILOR CU PYTHON

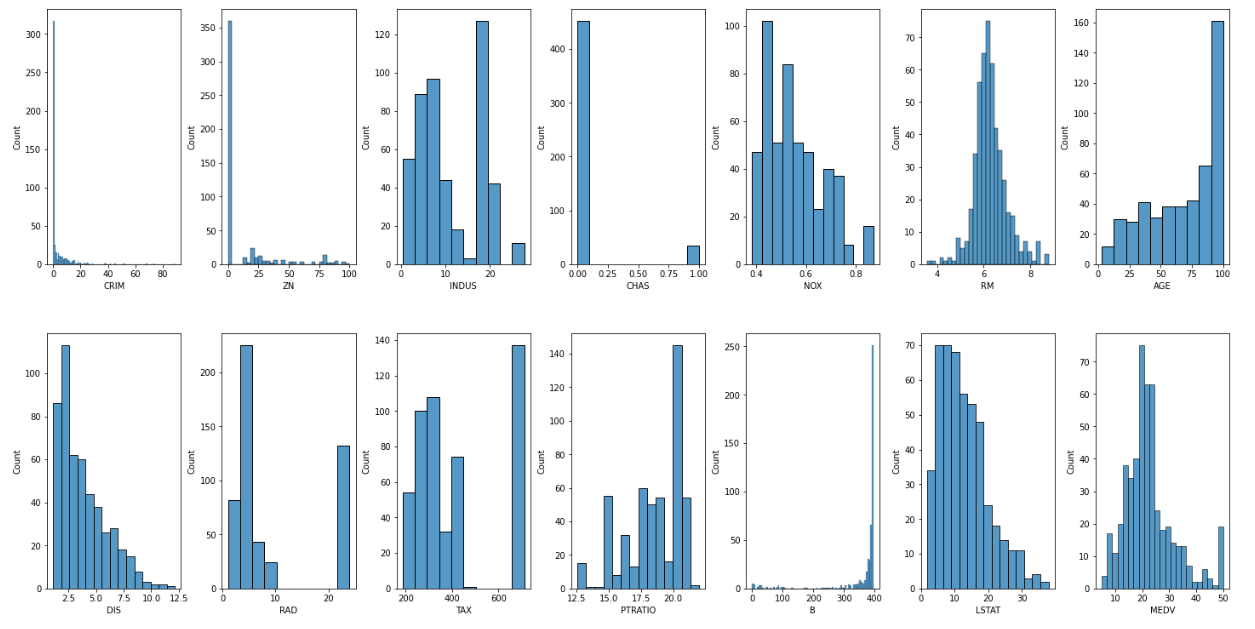
În capitolul ce urmează vor fi prezentate rezultatele obținute în urma crearea modelelor cu ajutor la python.

Primul pas care l-am făcut a fost vizualizarea datelor și analiza lor, datele le putem observa în figura 1. În această analiză, se pot distinge câteva tendințe și caracteristici notabile: La prima vedere, rata criminalității (CRIM) pentru multe zone pare a fi scăzută, având majoritatea valorilor concentrate spre partea de jos a graficului. Similar, proporția de teren rezidențial (ZN) pare a fi redusă pentru multe dintre înregistrări, având valori preponderent mici. Acest lucru sugerează că o mare parte din zonele din setul de date pot fi mai dens populate sau urbane. O caracteristică interesantă este legată de proximitatea față de râul Charles (CHAS). Acest atribut, care indică dacă o casă este în apropiere de râu sau nu, prezintă o majoritate covârșitoare a caselor situate departe de râu. Un alt aspect remarcabil este nivelul de poluare, măsurat ca concentrație de oxid de azot (NOX). Distribuția arată că există zone cu niveluri variate de poluare, având o medie în jurul valorii de 0.55. Când ne uităm la numărul mediu de camere per locuință (RM), observăm o distribuție destul de echilibrată, cu o medie în jurul valorii de 6 camere. Aceasta poate indica faptul că majoritatea caselor au dimensiuni medii. În încheiere, este demn de menționat că, în toate diagramele, există câteva puncte izolate care reprezintă valori extreme. Aceste valori pot indica anumite particularități sau anomalii în date, care ar putea necesita o analiză ulterioară detaliată.



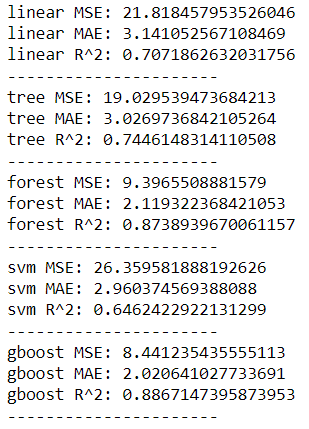
**Figura 1.** Visualizarea datelor cu histograme.

Figura 2, ilustrează distribuțiile a diferitelor atribute ale setului de date cu prețurile caselor din Boston. Observăm o concentrare mare a ratelor scăzute de criminalitate (CRIM), indicând zone mai sigure. Majoritatea proprietăților par să fie departe de râul Charles (CHAS), cu valoarea concentrată la zero. De asemenea, nivelul de poluare (NOX) se înclină spre o concentrație medie de 0,5, iar numărul mediu de camere (RM) indică că cele mai multe locuințe au în jur de 6 camere. În ceea ce privește impozitul pe proprietate (TAX), majoritatea caselor au rate de impozitare mai mari, iar proporția populației cu un statut socio-economic mai scăzut (LSTAT) se înclină spre valori mai mari, ceea ce sugerează zone cu venituri potențial mai scăzute.



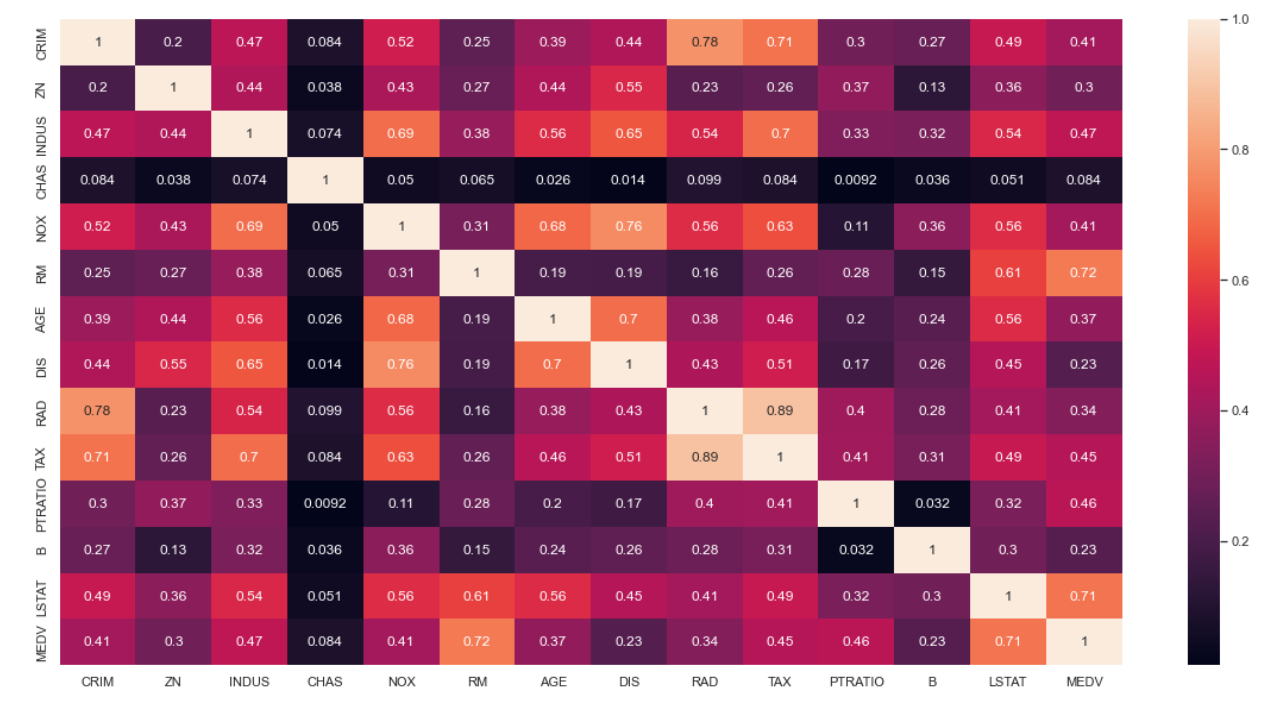
**Figura 2.** Visualizare detelor prin distribuție.

În figura 3, este prezentat rezultatul pentru pipelinul a modelelor pentru metodele regresii liniare, decision tree regresiion, random forest regression, suport vector machine regression și gradient boost regression. Modele au fost create pe baza întregului data set.



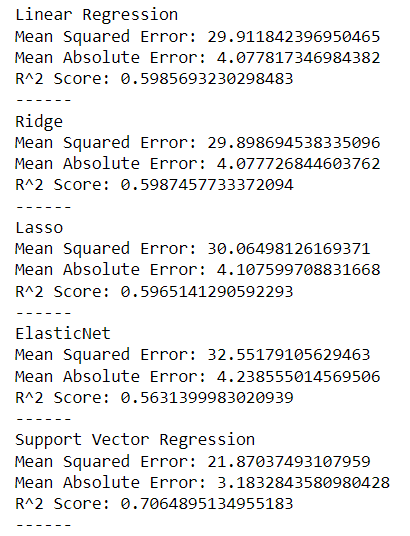
**Figura 3.** Rezultatele pentru primul test

În cele ce urmează am făcut corelați pentru întreg data set, pe a analiza care coloane pot avea un impact mai mare asupra crearea modelelor pentru a prezice coloana MEDV care este de fapt coloană țintă, în figura 4 o putem observa.



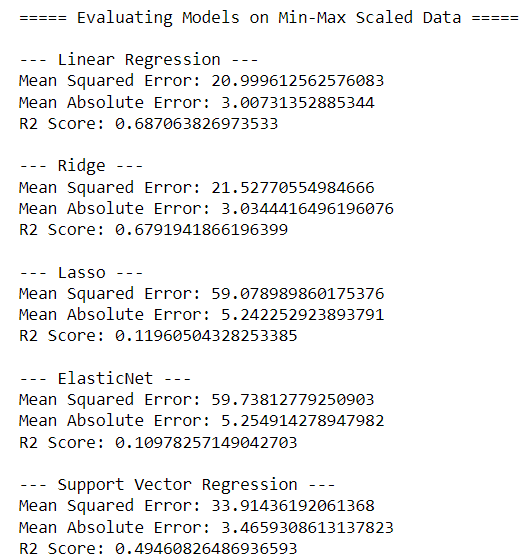
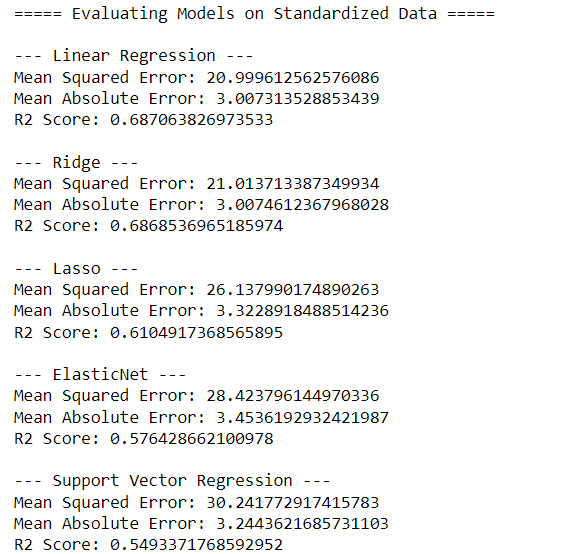
**Figura 4.** Corelația pentru data setul nostru.

Următorul pas care l-am făcut a fost crearea modelor pe baza coloanelor care au fost rezultate din analiza corelației, a fost selectate coloanele care au corelația cea mai înaltă și anume RM și LSTAT, eu pentru acest pas mă așteptam să fie un rezultat mai înalt la acuratețe, ele fiind prezentate în figura 5.



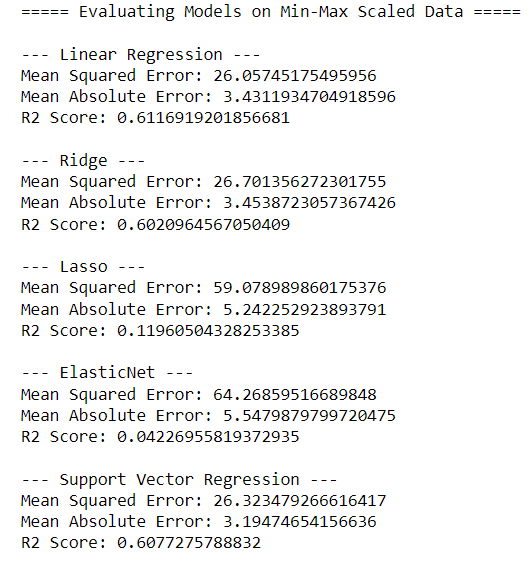
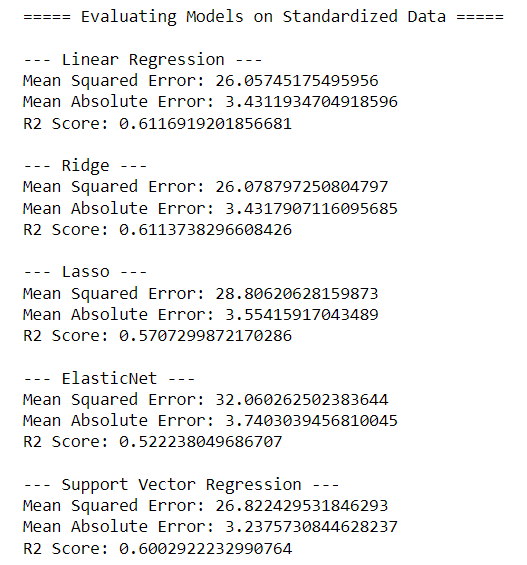
**Figura 5.** Acurateția modelelor pe baza coloanelor RM și LSTAT

În urma crearea modelelor de mai acurateța nu era cea mai bună și am decis să fac anumite teste pentru a îmbunătăți acuratețea. Următorul pas pe care l-am făcut a fost standardizarea și scalarea datelor pe întreg date set. Rezultatele sunt prezentate în figura 6 și figura 7.



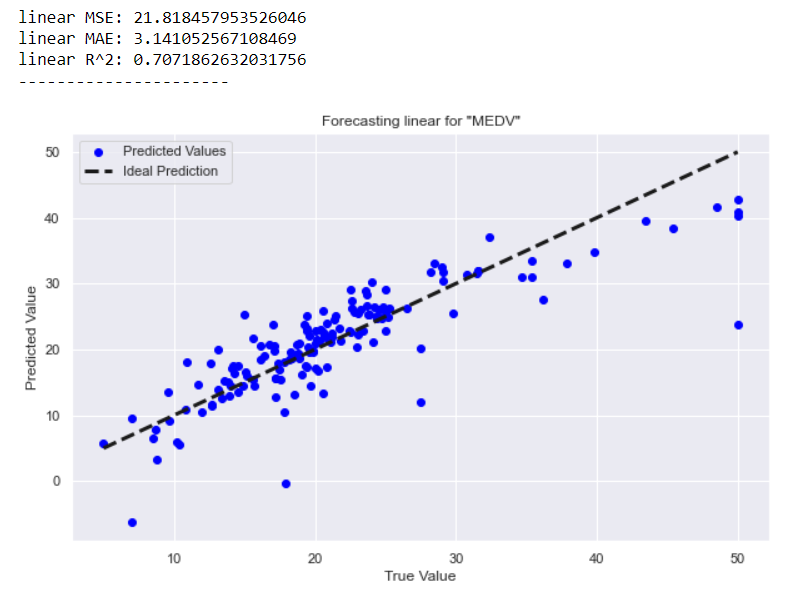
**Figura 6 și figura 7.** Rezultatele pentru standardizarea și scalarea detelor pe întreg date set

Următorul pas care a fost întreprins a fost crearea modelelor cu același metode, de asemenea folosind datele fiind standardizate și scalate, în figura 8 și figura 9 sunt prezentate rezultatele pentru aceste teste.



**Figura 8 și figura 9.** Rezultatele standardizări și scalări pentru coloanele cu o corelație mai mare.

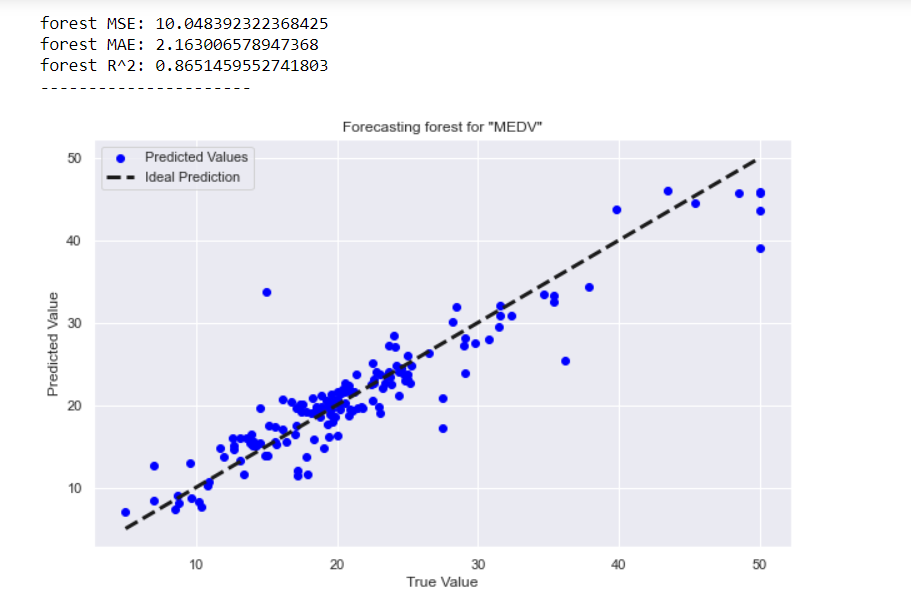
În urma concluziilor care se bazează pe analiza rezultatelor acurateţei obținute de metodele noastre pentru creare de modele am ajuns la concluzia că cel mai corect pas este anume să nu fie făcut standardizare și scalare datelor, de asemenea să nu se excludă outliers și să fie creat modelul pe întreg date set doar recrearea de date, rezultatele predicțiilor sunt prezentate în figurile: figura 10, figura 11, figura 12, figura 13 și figura 14.



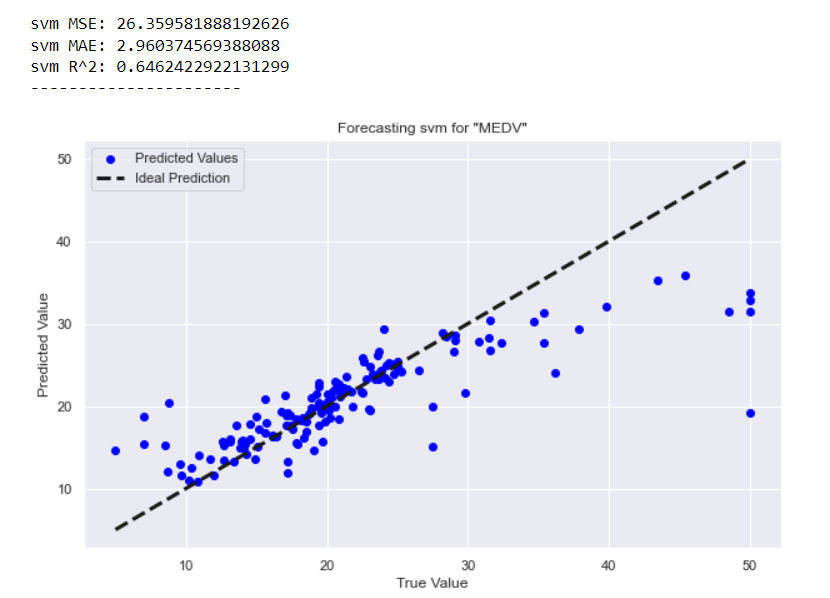
**Figura 10.** Predicțiile pentru regresia liniară

****

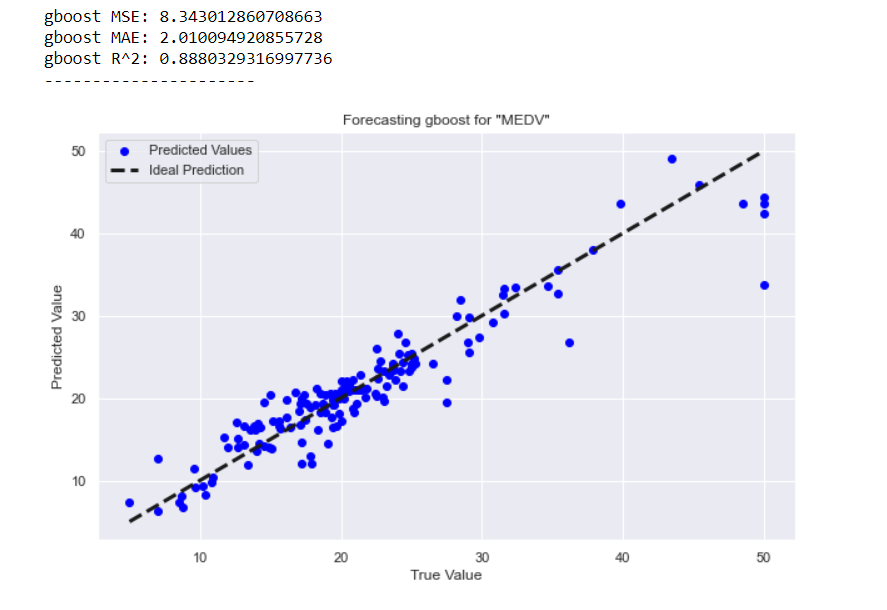
**Figura 11.** Predicțiile pentru Decision Tree regression

****

**Figura 12.** Predicțiile pentru Random Forest Regression

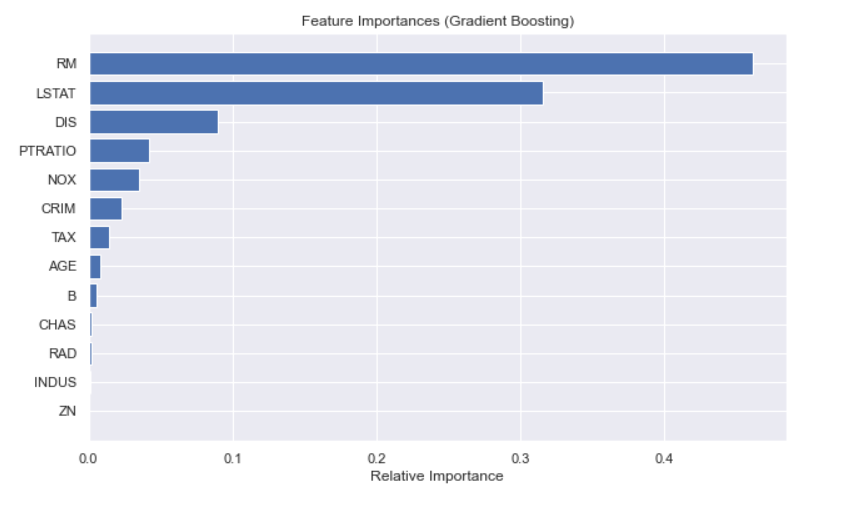
****

**Figura 13.** Predicțiile pentru Support Vector Machine

****

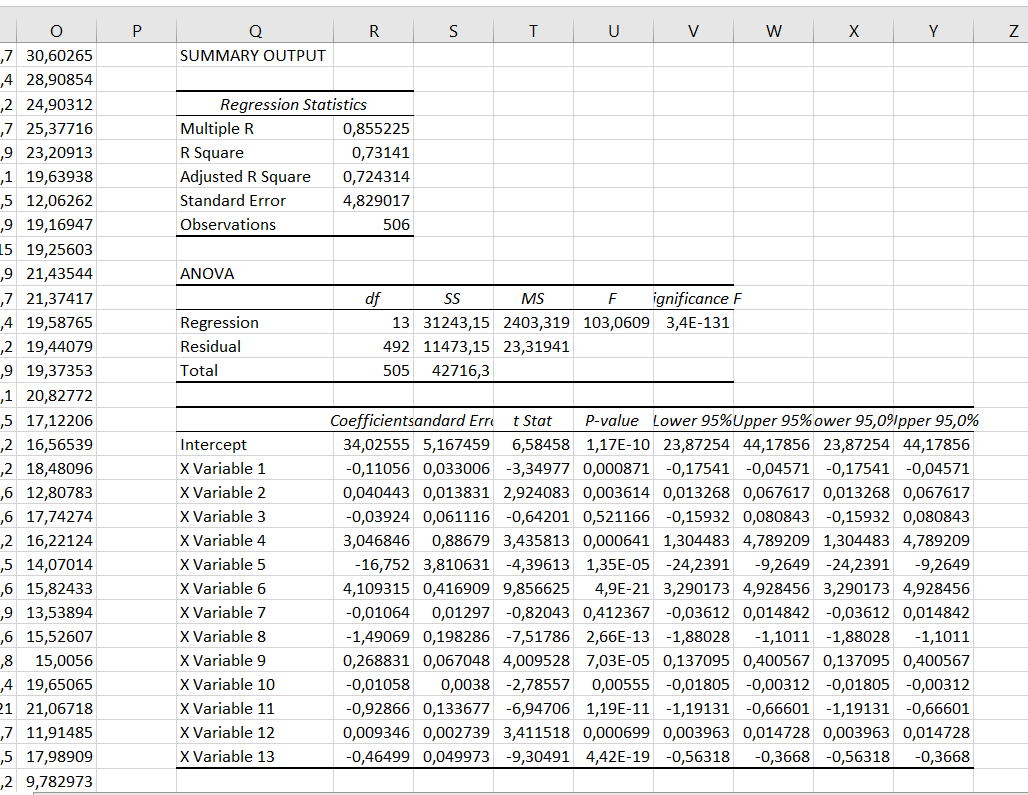
**Figura 14.** Predicțiile pentru Gradien Bost Regression

De pentru a înțelege mai bine modelul care o acuratețe cât mai bună, am luat și am făcut o interpretare de model, unde din figura 15 putem observa că coloanele RM și LSTAT sunt cele mai valorase în modelul nostru.



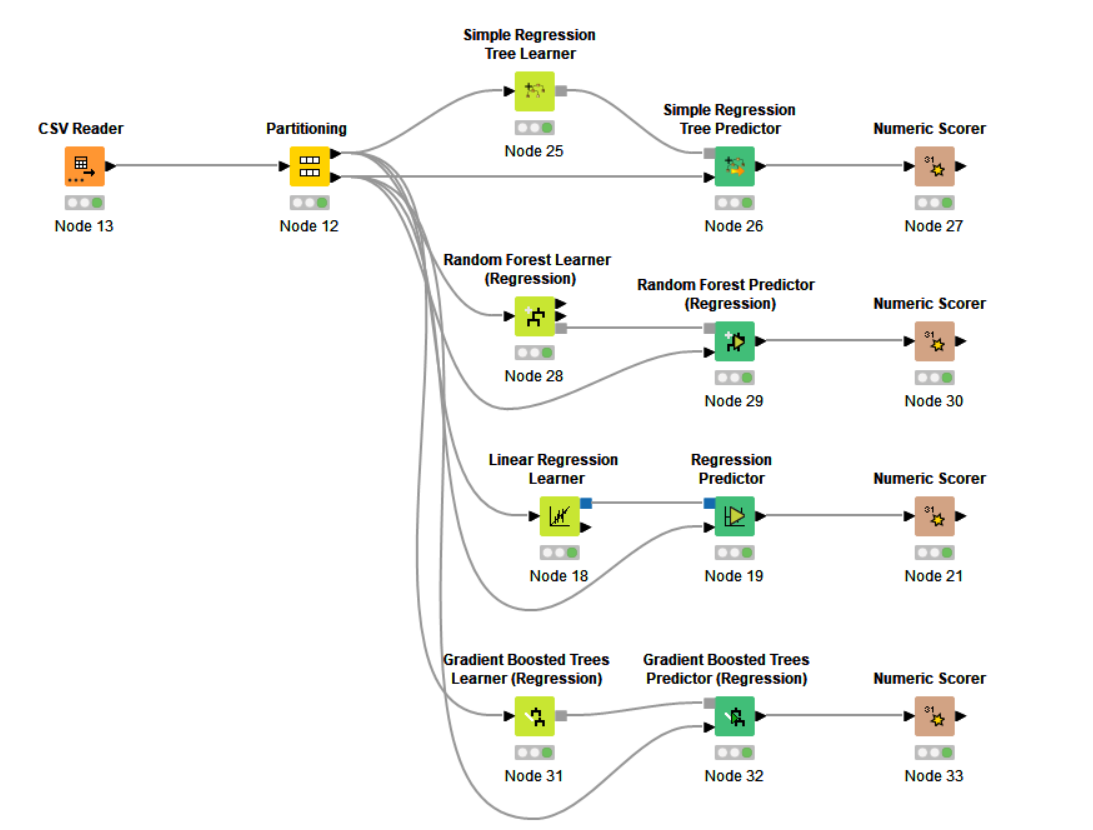
**Figura 15.** Predicțiile pentru

Următoare aplicație pe care am folosită pentru a evalua și a creare model pentru datele noastre a fost excell-ul aici am creat o regresie și anume ANOVA, rezultatele le putem observa în figura 16.

****

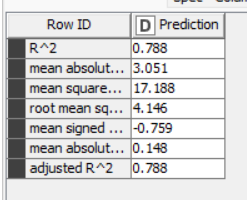
**Figura 16.** Rezultatele din excell

Următoare aplicație care a fost folosită pentru a crea modele a fost KNIME, în figura 17 sunt prezentate flow-urile pentru modelele noastre: Simple Regression Tree Predictions, Random Forest Prediction, Linera Regression și Gradient Booster Regression.

****

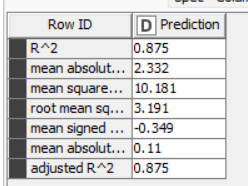
**Figura 17.** Flow-ul pentru datele noastre

În figura 18, este prezentat datele pentru prima regresie și anume regresia „Simpe Regression Tree”, după cum putem observa avem un rezultat de 0.788 pentru R2.



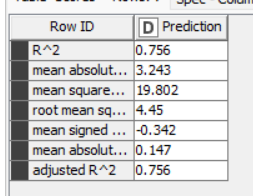
**Figura 18.** Simpe Regression Tree

În figura 19, este prezentat datele pentru prima regresie și anume regresia „Random Forest Regression”, după cum putem observa avem un rezultat de 0.87 pentru R2.



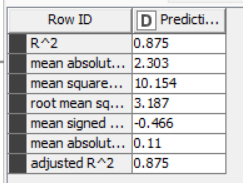
**Figura 19.** Random Forest Regression

În figura 20, este prezentat datele pentru prima regresie și anume regresia „Linear Regression”, după cum putem observa avem un rezultat de 0.75 pentru R2.



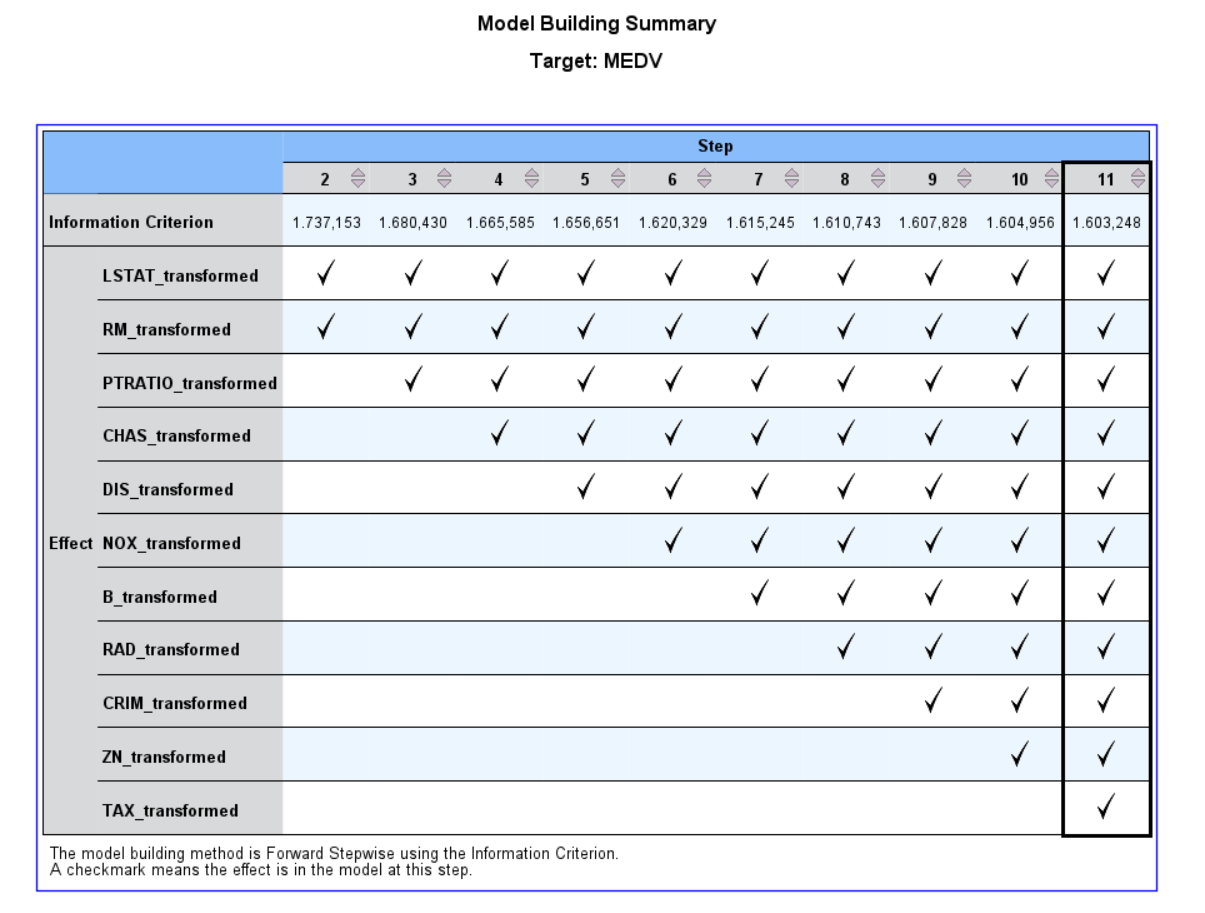
**Figura 20.** Linear Regression

În figura 20, este prezentat datele pentru prima regresie și anume regresia „Gradient Booster Trees Regression”, după cum putem observa avem un rezultat de 0.87 pentru R2.

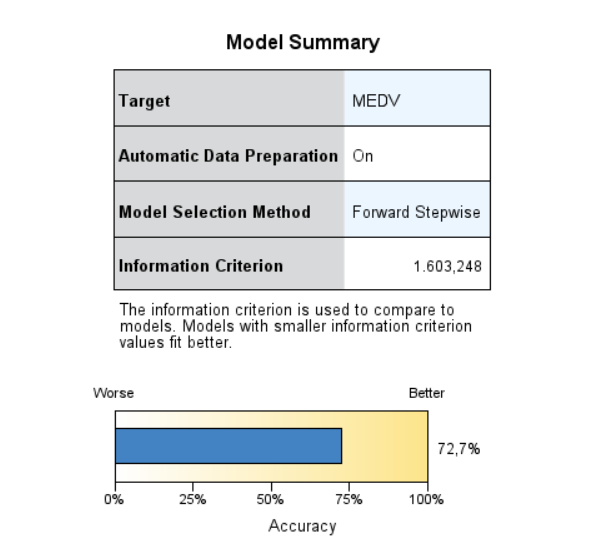


**Figura 20.** Gradient Booster Trees Regression

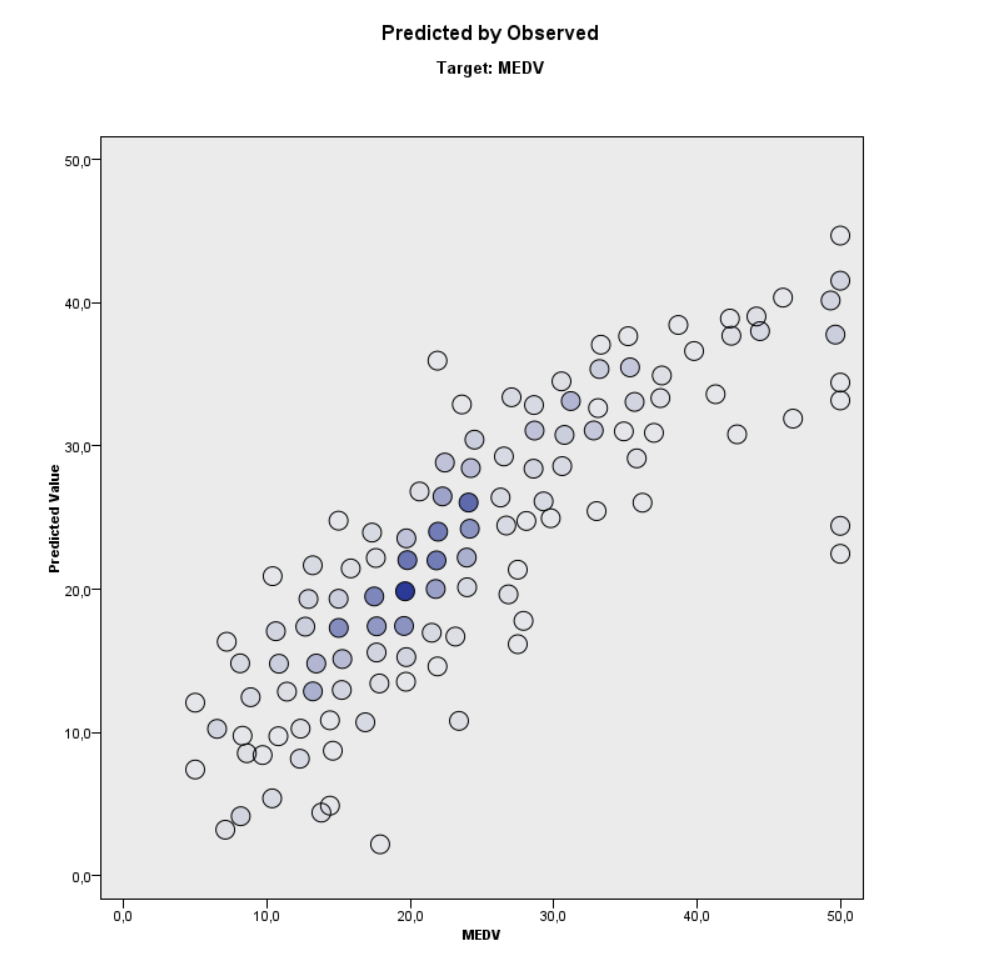
Următoare aplicație care a fost utilizată pentru a IBM SPSS, unde am făcut noi regresii pentru același set de date. În figura 21, este prezentat Regressia Forward Step, iar în figura 22, putem observa acurateţea modelului și în figura 23 este însăși predicția făcută după modelul dat, acest model a fost realizat pe întreg data set.



**Figura 21.** Modelul a fost creat cu ajutor metodei Forward Stepwise

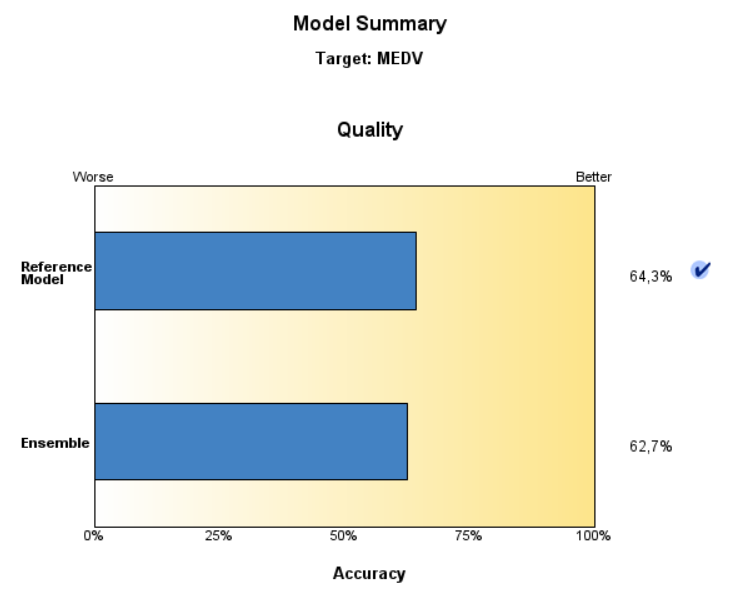


**Figura 22.** Acuratețea modelului



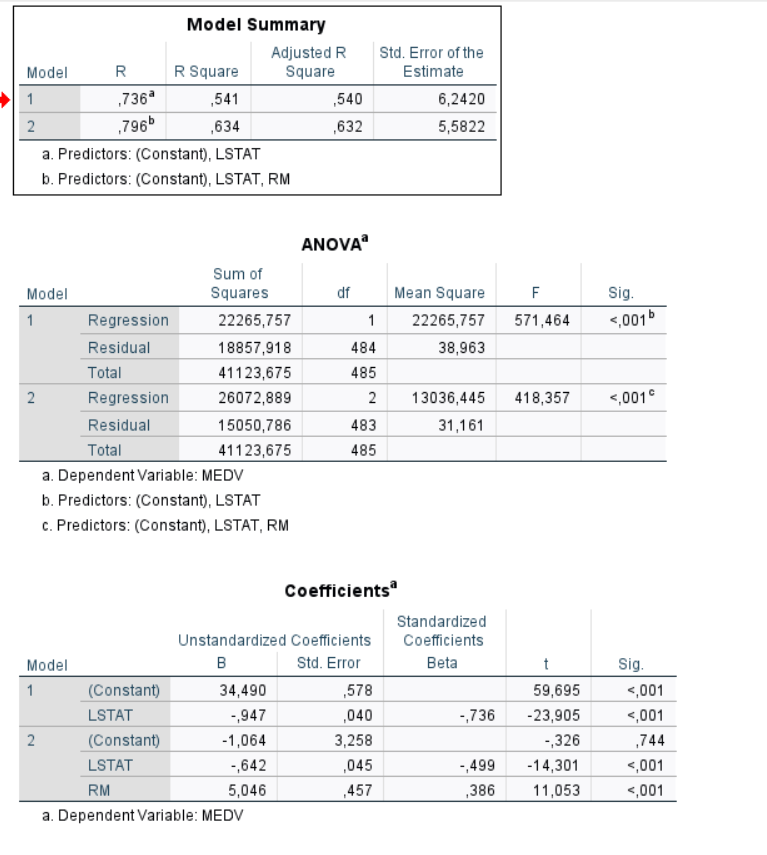
**Figura 23.** Predicţiile cu ajutor modelului creat.

În cele din urmă a fost selectat o altă metoda Regresii Liniară și a anume cu caracteristicele de Auto Linear Modeling pe datele care aveau o corelație mai înaltă, figura 24.



**Figura 24.**  Automatic Linear Modeling

Și ca o metodă opțională a fost utilizată Regressi Anova, unde rezultatele pe putem observa în figura 25.



**Figura 25.** Regresia Anova

CONCLUZIA

În urma efectuări acestui laborator am făcut cunoștință cu analiza și vizualizarea datelor ceea ce este ceva nou și sper să mă aprofundez mai mult în ceea ce urmează la sigur pot spune că acest laborator m-a inițiat în vizualizare datelor.

BIBLOGRAFIE

1. Iris Dataset [online]. [accesat la data de 5.10.2022] Disponibil: <https://www.geeksforgeeks.org/python-basics-of-pandas-using-iris-dataset/>