Preturile masinilor second-hand din Belarus in functie de caracteristicile acestora

Proiect realizat de: Ciobota George-Marian, george.ciobota@stud.ubbcluj.ro

Cadis Andreea, [andreea.cadis1@stud.ubbcluj.ro](mailto:george.ciobota@stud.ubbcluj.ro) , IE, An 3, Gr 1

 **1.Introducere:**

In introducere va vom furniza niste informatii despre dataset-ul ales si va vom prezenta intrebariile pe care noi le-am considerat relevante pentru efectuarea de acestui studiu.

Noi am efectuat studiul pe un set de date care reprezinta piata de masini second-hand din Belarus de la inceputul anului 2019. Informatiile pe are le avem in acest set de date ne sunt foarte folositoare pentru efectuarea de rapoarte, aceestea fiind marca, modelul masinii, preturile in USD, anul fabricarii, conditia masinii la momentul vanzarii, kilometrajul, tipul de combustibil al masinii, volumul motorului, etc.

Prin acest studiu noi am vrut sa analizam impactul mai multor caracteristici al masinilor puse spre vanzare asupra pretului acestora. Caracteristicile pe care noi le-am considerat cele mai relevante si pe care le-am analizat au fost anului in care a fost produsa masina, kilometrajul acesteia, tipul de combustibil folosit de aceasta si volumului motorului de care dispune masina. Noi consideram ca aceste caracteristici sunt importante pentru piata de masini second-hand si sunt de relevante atat pentru cumparatori cand vor sa achizitioneze o masina, cat si pentru vanzatorii acestora.

Prin intermediul predictiilor noastre, am dorit sa raspundem la cateva intrebari de cercetare cum ar fii: Cum influenteaza anul productiei pretul unei masini? Care este relatia dintre kilometraj si pret? Ce relevanta are tipul de combustibil asupra pretului? Care este relatia dintre volumul motorului cu pretul unei masini?

**2.Setul de date:**

Setul de date pe care noi l-am folosit pentru acest studiu reprezinta piata de masini second-hand din Belarus. Noi am preluat acest set de date de pe site-ul <https://www.kaggle.com/> iar in acest dataset regasim informatii importante referitoare la masinile disponibile spre vanzare in Belarus la inceputul anului 2019.

Coloanele din dataset-ul nostru sunt:

**make**: reprezinta firma masinii.

**model**: reprezinta modelul masinii.

**priceUSD**: reprezinta pretul masinii in USD

**year**: indica anul fabricatiei al masinii.

**condition**: starea masinii (cu kilometraj, pentru piese, etc.).

**mileage.kilometers.**: ne indica kilometrajul masinii.

**fuel\_type**: indica tipul de combustibil utilizat de masina (electro, benzina, motorina).

**volume.cm3.**: specifica volumul motorului.

**color**: denumeste culoarea masinii.

**transmission**: indica tipul de transmisie (manuala, automata).

**drive\_unit**: specifica sistemul de tractiune al masinii.

**segment:** reprezinta o caracteristica colectata manual, care indica segmentul de piata al masinii.

Pe baza acestor informatii pe care le avem in dataset putem obtine informatii valoroase despre piata de masini second-hand din Belarus, cum ar fi preturile medii in functie de anul de productie, relatia dintre kilometraj si pret, influenta tipului de combustibil asupra pretului etc. Cu ajutorul acestor caracteristici vom putea raspunde intrebarilor noastre de cercetare si sa intelegem mai bine dinamica pietei auto second-hand in Belarus.

**3.Rezultate si discutii**

**3.1.Metode de analiza**

Ca si metode de analiza noi am folosit **Regresia liniara** si **Arborii de decizie.**

Am utilizat regresia liniara pentru a observa mai bine relatia dintre variabilele independente si pretul masinilor din piata auto. Am aplicat regresie liniara pentru variabilele "year" (anul de productie), "mileage.kilometers." (kilometrajul) si "fuel\_type" (tipul de combustibil) in raport cu variabila dependenta "priceUSD" (pretul masinii). Am folosit metoda arborilor de decizie pentru a prezice pretul masinilor (ArborePret) si a anului masinilor (ArboreAn) si am creat doi arbori care au prezentat predictii pe baza combinatiilor variabilelor independente. Acestea au oferit o perspectiva asupra importantei variabilelor in predictia pretului si a anului masinilor.

Am mai utilizat arborele de decizie pentru predictia kilometrajului masinilor (ArboreTipComb) care a oferit predictii pe baza anului de productie. Acesta ne-a indicat cum anul productiei unei masini poate influenta kilometrajul acesteia.

**3.1.Setarile testate**

- Pentru regresia liniara, am utilizat modelul de baza, care includea doar variabila independenta de interes si variabila dependenta. Am folosit regresia liniara pe toate variabilele din dataset-ul nostru pentru a observa cum ele influenteaza pretul masinilor.

- Pentru arborii de decizie, nu am modificat setarile functiei rpart(), deci am folosit setarile implicite ale acesteia in R. Din aceasta cauza am creat arborii de decizie cu setarile prestabilite.

**3.2.Strategia de validare**

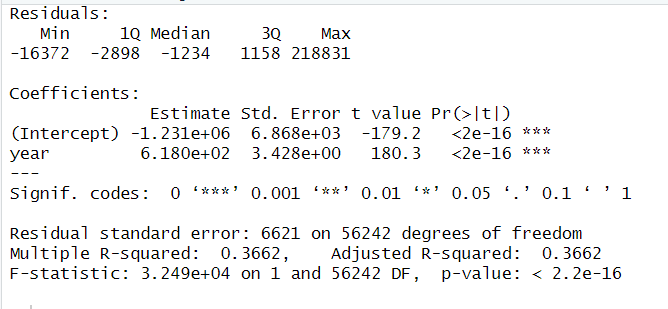
Validarea modelelor create de noi am facut-o pe baza analizei regresiei liniare si a arborilor de decizie, nu am aplicat o strategie specifica de validare.

**3.3.Rezultatele obtinute**

* Noi am vrut sa investigam influenta mai multor caracteristici ale masinilor asupra preturilor acestora, astfel am inceput prin a realiza o regresie liniara intre anul si pretul masinilor pentru a observa ce influenta are anul aparitiei asupra pretului. Pentru a obtine rezultate relevante in cercetarea noastra am utilizat metoda regresiei liniare:



Care a generat urmatoarele rezultate:



Din rezultatele obtinute in urma regresiei liniare “lm\_year” noi putem observa ca intre variabila independeta “year” si variabila independenta “priceUSD” exista o relatie puternica. Constatam ca variabila year are o influenta pozitiva asupra pretului masinii.

- Datorita coeficientului “year” de 6.180e+02 noi observam ca media pretului va creste cu 618 USD pentru fiecare unitate cu care creste anul.

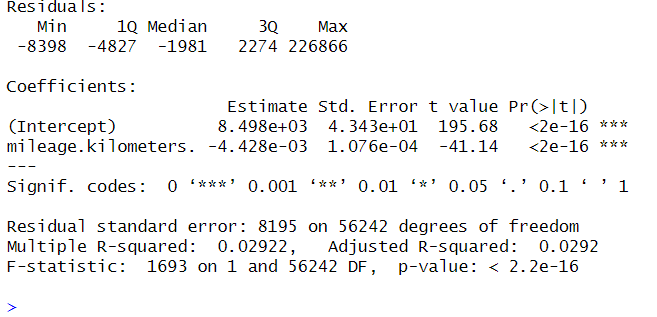
- Mai apoi datorita coeficientului “R-squared” de 0.3662 noi constatam ca 36.62% din variabilitatea pretului poate fii influentata de anul masinii

* Dupa ce am analizat impactul anului aparitiei asupra pretului masinii, noi am continuat cercetarea noastra analizand si relatia dintre kilometraj respectiv tipul de combustibil si pretul masinii pentru a putea avea impactul pe care cele doua caracteristici il au asupra pretului. Pentru a efectua aceasta cercetare noi am procedat la fel ca si in cazul impactului anului asupra pretului, realizand doua regresii liniare.

Mai intai o sa analizam regresia liniara “lm\_mileage” care analizeaza impactul kilometrajului asupra pretului:



Iar rezulatele obtinute sunt:



Din rezultatele obtinute de noi observam ca kilometrajul are intr-adevar o influenta negativa asupra pretului masinii, dar nu o influenta foarte mare, cum era in cazul anului.

Aceasta concluzie o putem desprinde pe baza rezultatelor regresiei liniare:

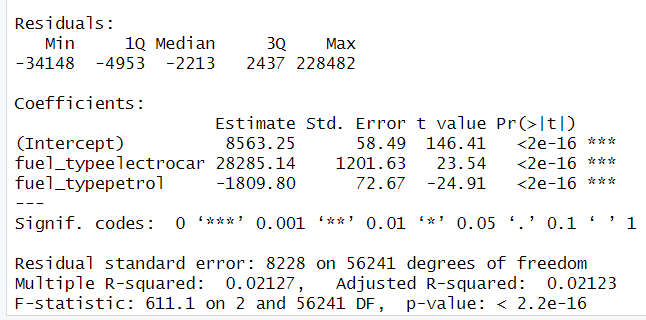
- Coeficientul “R-Squared” este doar de doar de 0.02922 indicand ca doar aproximativ 2.92% din variabilitatea pretului poate fi motivata de kilometrii masinii.

- Coeficientul “mileage.kilometers”: de -4.428e-03, ne sugereaza ca pretul mediu al masinilor scade cu aproximativ 0.004428 USD pentru fiecare kilometru in plus al masinii.

Mai apoi am continuat cu analiza regresiei liniare “lm\_fuel\_type” care analizeaza impactul tipului de combustibil asupra pretului:



Si am obtinut rezultatele:



Din rezultatele obtinute putem constata faptul ca exista diferente importante intre preturile masinilor in functie de tipul de combustibil. Masinile electrice au in medie preturi mai mari decat cele pe petrol.

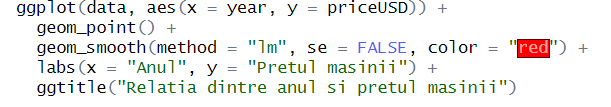
- Variabila “fuel\_typeelectrocar” este de 28285.14 si ne indica diferenta dintre pretul mediu al masinilor electrice si al celor pe petrol sau diesel.

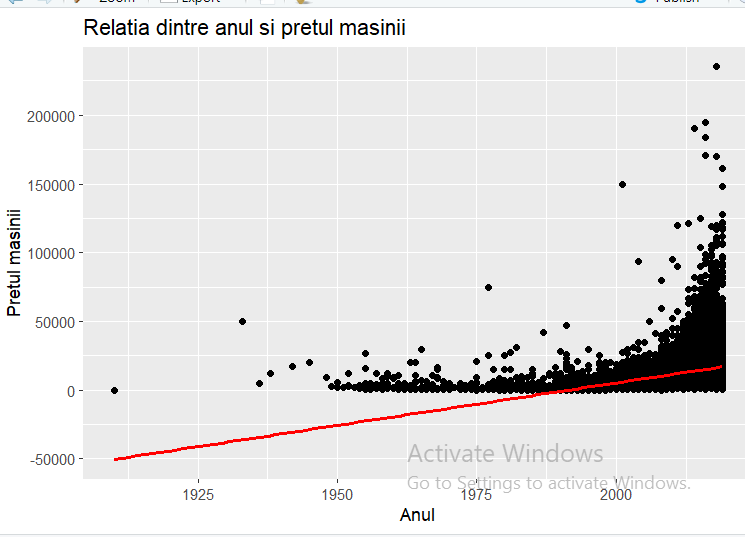
- Variabila “fuel\_typepetrol” este de -1809.80 si ne indica diferenta dintre pretul mediu al masinilor pe petrol si al masinilor electrice sau diesel

- Coeficientul “R-squared” este doar de 2.13% iar de aici constatam ca "fuel\_type" influenteaza foarte putin variabilitatea preturilor.

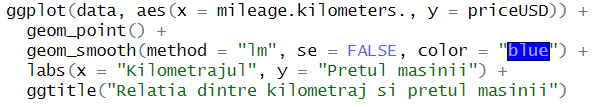
* Cele trei regresii liniare pe care le-am realizat sunt foarte utile atat pentru cumparatorii de masini cat si pentru vanzatorii de masini. Aceste regresii liniare furnizeaza informatii importante atat pentru cumparatorii de masini cat si pentru vanzatori. Cumparatorii vor stii daca pretul masinii este unul profitabil pentru ei pe baza anului, kilometrajului si al tipului de combustibil folosit de masina, iar vanzatorii vor stii cu ce pret sa isi puna la vanzare masina in functie de caracteristicile pe care aceasta le are.
* Pentru o vizualizare mai buna a rezulatelor regresiilor noastre am realizat niste grafice de dispersie. Ele ne vor ajuta la o intelegere a relatiilor si tendintelor dintre variabile.

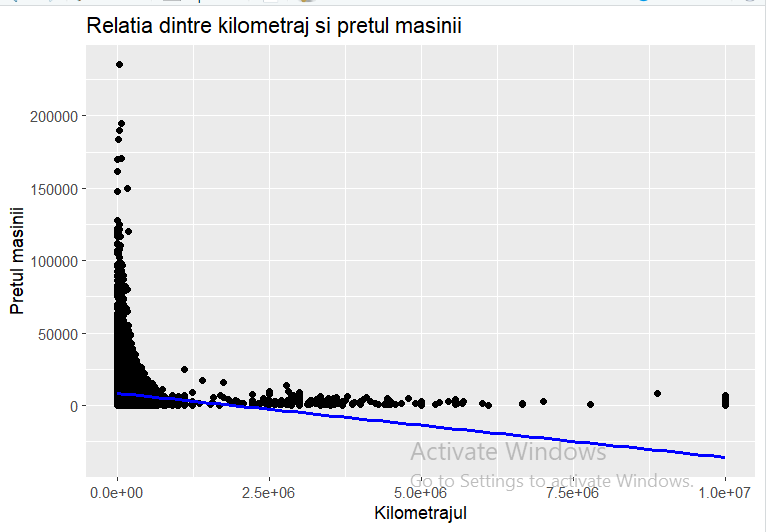
- Primul grafic de dispersie reprezinta relatia dintre anul productiei masinii si pretul acesteia, linia rosie ne ofera o estimare a relatiei dintre cele doua variabile.



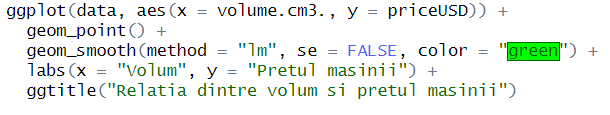


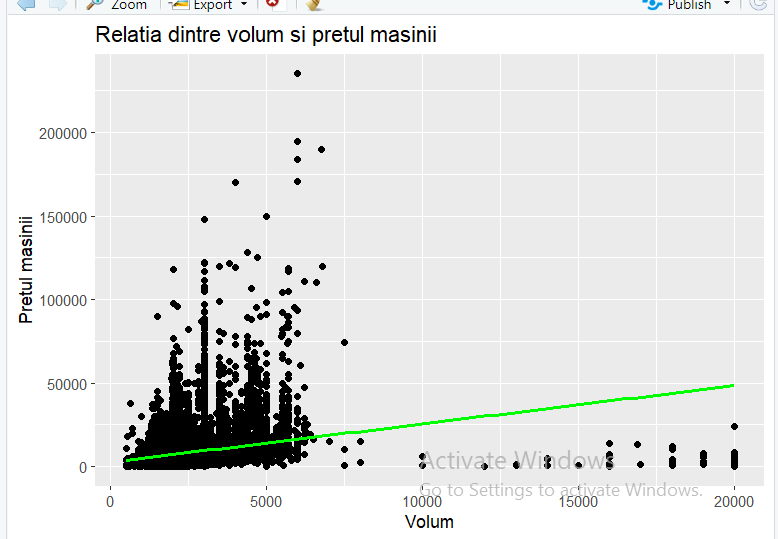
- Al doilea grafic de dispersie reprezinta relatia dintre kilometrajul masinii si pretul acesteia, linia albastra ne ofera o estimare a relatiei dintre cele doua variabile.





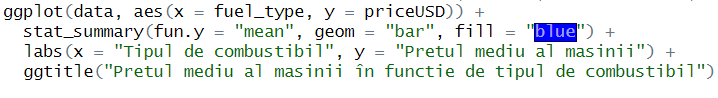
- Al treilea grafic de dispersie reprezinta relatia dintre volumul motorului masinii si pretul acesteia, linia verde ne ofera o estimare a relatiei dintre cele doua variabile.

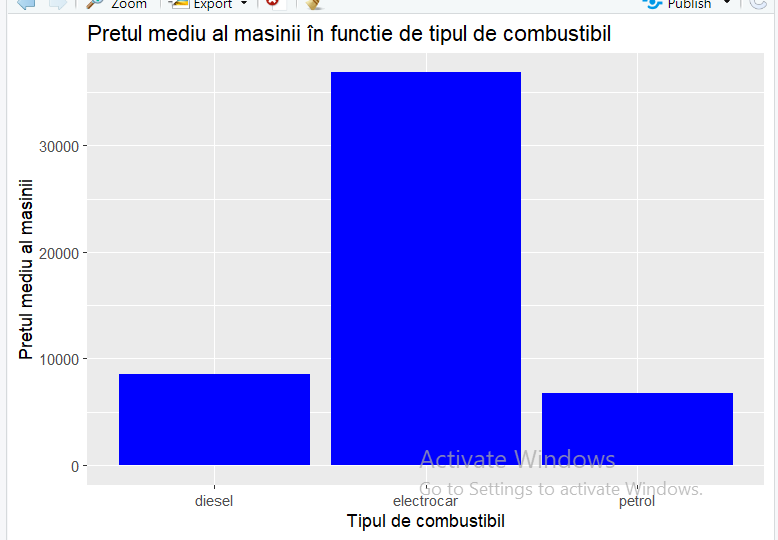




* Primele doua grafice ne confirma analiza primelor doua regresii liniare. Ultimul grafic ne reprezinta ca, cu cat volumul motorului creste cu atat este mai mare si pretul masinii.

- Pentru a vizualiza mai bine diferentele de pret in functie de tipul de combustibil, am realizat un grafic de bare pe care putem observa ca pretul cel mai mare il au masinile cu “fuel\_type”=“electrocar” si cele mai mici preturi le au cele cu “fuel\_type”=”petrol”.

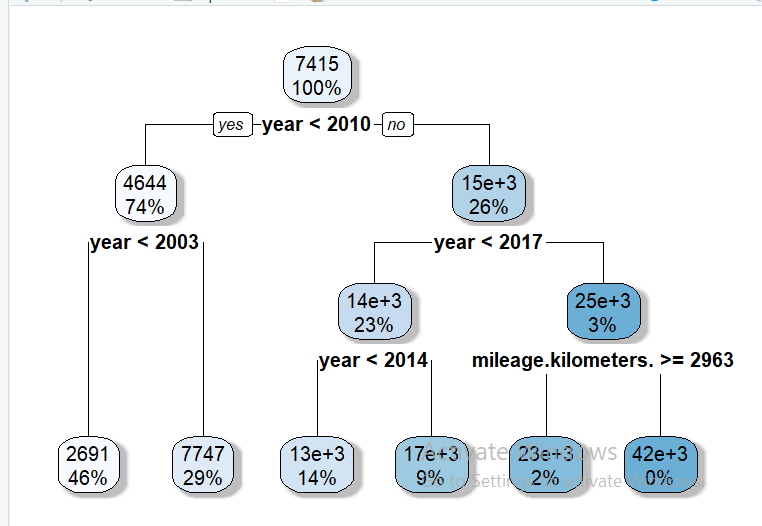




* Am continuat analiza noastra prin crearea unor arbori decizionali care ne vor ajuta la predictia anumitor variabile independente in functie de alte variabile independente prin crearea unor criterii de decizie a variabilelor independente furnizate de noi pentru predictie.

Primul arbore pe care noi am ales sa il construim a fost arborele de decizie pentru predictia pretului masinii in functie de anul productiei si al kilometrajului.



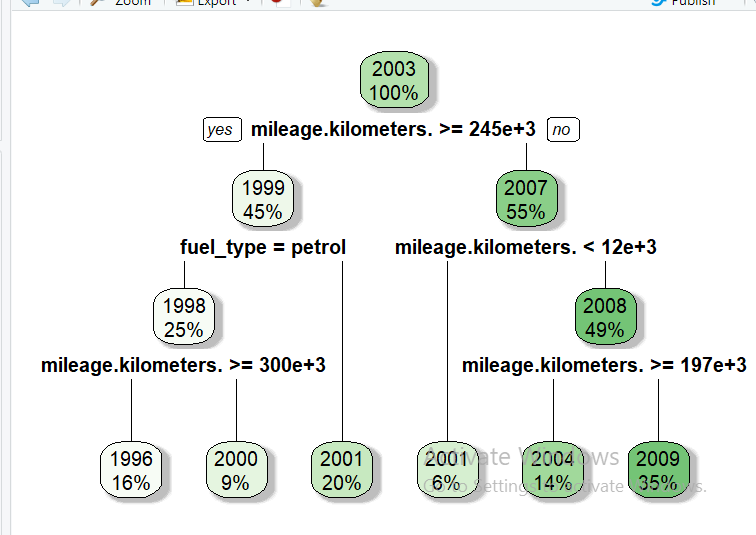


Din acest arbore rezultat noi putem observa predictii ale pretului in USD pe baza unor reguli stabiie in functie de kilometrajul respectiv anul productiei masinii. La radacina arborelui avem o valoare stabilita pentru separarea datelor in doua ramuri. Urmeaza ca fiecrui nod sa ii fie atribuit o regula dupa care se va face urmatoarea separare a acelui nod. De exemplu daca year<2010, nodul din stanga lui(copilul nodului root) ne spune ca 74% din ani sunt <2010 iar media pretului masinilor cu year<2010 este de 4644 lei aproximativ. Mai apoi acest nod va pune o regula year<2003, iar pe baza acestei reguli, daca year<2003 se merge pe nodul din stanga iar daca nu se merge pe nodul din dreapta, astfel prezicand valoriile medii a preturilor in functie de variabila independenta year. In parte din stanga a arborelui putem observa predictii ale pretului si pe baza unor reguli care implica variabila independent a mileage.kilometers.

Astfel, arborii nostrii de predictie cuprind toate cazuri si predictiile posibile in functie de variabilele independente.

* Pe acelasi concept si aceasi metoda a arborilor de decizie am creat urmatorul nostru arbore care reprezinta predictia anului masinii pe baza tipului de combustibil si a kilometrajului.

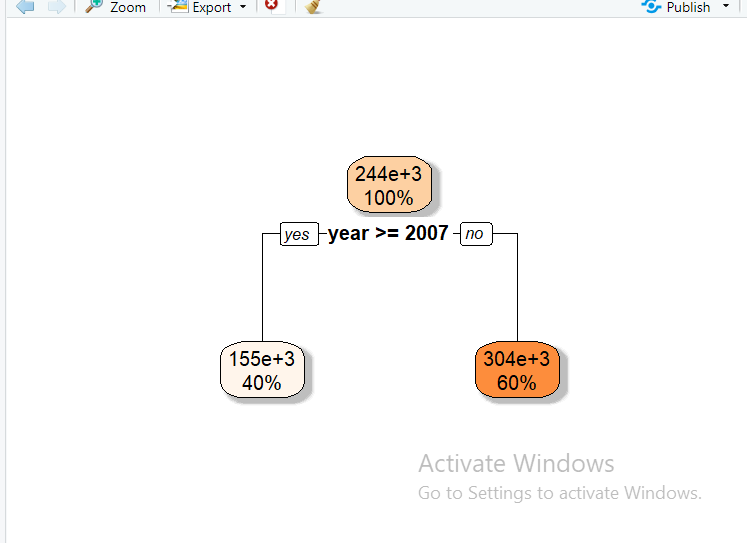




Pe acelasi model ca arborele anterior, acest arbore de predictie cuprinde toate cazuri si predictiile posibile pentru anul masinii in functie de kilometrajul si tipul de combustibil al masinii.

* Ultimul nostru arbore de decizie a fost folosit pentru predictia kilometrajului in functia anului productiei acesteia:





Din acest arbore prezicem a daca anul este >= 2007 atunci kilometrajul este in medie de 155.000km iar in aceasta categorie se afla 40% dintre masini, iar daca anul este < 2007 atunci kilometrajul in medie este de 304.000km si in aceasta categorie se afla 60% dintre masini.

* Arborii de decizie construiti de noi sunt foarte importanti pentru cumparatorii si vanzatorii de masini, deoarece ei realizaeaza niste predictii pe baza a diferite variabile independente furnizate de noi. Acestea ajuta la luarea de decizii atunci cand ei analizeaza diferite caracteristici a masinilor, cum ar fi stabilirea pretului in functie de anul productiei si kilometraj, prezicerea anului de productie in functie de kilometraj si tipul de combustibil si estimarea kilometrajului in functie de anul aparitiei.

1. **Concluzie**

In concluzie, noi am abordat mai multe intrebari de cercetare la care am gasit raspunsuri prin utilizarea regresiei liniare si a arborilor de decizie. Cele trei regresii liniare pe care le-am realizat ne-au ajutat sa investigam relatia dintre anul de productie si pretul, kilometrajul si pretul, dar si dintre volumul motorului si pretul. Rezultatele regresiei liniare ne-au specificat importanta acestor variabile asupra pretului masinilor si au permis realizarea de predictii.

Pe de alta parte, arborii de decizie ne-au ajutat sa raspundem la intrebari precum relevanta anului aparitiei si a kilometrajului asupra pretului masinii, impactul tipului de combustibil si kilometrajului asupra anului masinii si relatia dintre anul de productie si kilometraj. Rezultatele noastre si raspunsurile la intrebarile de cercetare vor aduce o mare contributie la predictiile din domeniul masinilor second-hand din Belarus si vor ajuta cumparatorii si vanzatorii sa ia decizile corecte in acest domeniu.