

Academia de Studii Economice

Facultatea de Cibernetică, Statistică și Informatică Economică

**Sistem suport de decizie pentru fundamentarea deciziilor pentru o companie de taxi din orașul New York**

Sora Cornel,

Master E-Business,

Anul 2

Cuprins

[1. Introducere 2](#_Toc530443604)

[2. Componența sistemului decizional 3](#_Toc530443605)

[2.1. Subsistemul de date 3](#_Toc530443606)

[2.2. Susbsitemul de modele 11](#_Toc530443607)

[2.3. Subsistemul interfeței cu utilizatorul 14](#_Toc530443608)

# Introducere

Obiectivul acestui proiect este de a construi un model de machine learning care va prezice durata călătoriei pentru o cursă cu o mașină de taxi în orașul New York City, bazându-se pe date din anul 2016 culese pentru firma NYC Yellow Cab.

Pentru a avea o prediciție cât mai corectă și cu o marjă de eroare cât mai mică se vor urma patru faze principale:

* Preprocesare: curățarea și transformarea datelor pentru a putea fi utilizate în analiză.
* Analiza exploratorie: se vor exploata datele pentru a identifica parametri necesari pentru a fi folosiți pentru modelare.
* Implementarea algoritmului: se va utiliza arborele de decizie pentru a genera regresia liniară pe baza căreia se vor putea face previziuni
* Utilizarea modelului pentru a efectua previziuni pentru un set de date de test.

Orașul New York prezintă milioane de curse în fiecare lună care duc la trafic regulat,blocaje ale străzilor sau o scară larăg de evenimente. Astfel, este foarte important pentru firmele de taxi-uri să poată avea o vizibilitate a taxelor și a duratei curselor. Prin predicția duratei unei curse poate ajuta un client să ia decizia dacă să înceapă sau nu călătoria sau totodată poate ajuta firma de taxi să ia decizia care dintre două potențiale curse este mai profitabilă.

În scopul realizării proiectului a fost utilizat limbajul de progrmare pytho, prin intermediul căruia au fost citite și prelucrate datele și a fost construit modelul în urma căruia putem face predicția dorită, iar mai târziu decizii pe baza acesteia.

Datele au fost culese din baza de date oferită de **Google Cloud**, iar acestea au fost salvate într-un fișier de tip excel, având următoarele coloane:

|  |  |
| --- | --- |
| **Nume variabilă** | **Descriere** |
| id | Identificator unic pentru fiecare cursa |
| id\_sofer | Identificator asociat soferului care a preluat cursa |
| data\_pornire | Data la care a fost solicitată cursa |
| data\_sosire | Data la care s-a încheiat cursa |
| nr\_pasageri | Numărul de pasageri din interiorul mașinii |
| longitudine\_pornire | Longitudinea de unde a fost solicitat taxi-ul |
| latitudine\_pornire | Latitudinea de unde a fost solicitat taxi-ul |
| longitudine\_sosire | Longitudinea unde cursa a fost încheiată |
| latitudine\_sosire | Latitudinea unde cursa a fost încheiată |
| cursa\_marcata | Y=cursa a fost înregistrată; N=cursa nu a fost înregistrată |
| durata\_cursa | Durata cursei măsurată în secunde |

*Tabel 1. Descriere coloane date*

# Sistemului decizional pentru firma de taxi

## Subsistemul de date

Mai întai datele vor fi încărcate și curătate într-o formă cât mai utila analizei. Pentru prelucrarea datelor este utilzat limbajul de programare Python. Datele sunt citite dintr-un fișier de tip CSV, excel. În urma citirii datelor se observă că fișierul conține 1048575 de linii.

Pentru a analiza dacă datele conțin anomalii sau dacă necesită modificări au fost calculate: media, dispersia, valoarea minimă și valoarea maximă pentru fiecare coloană:

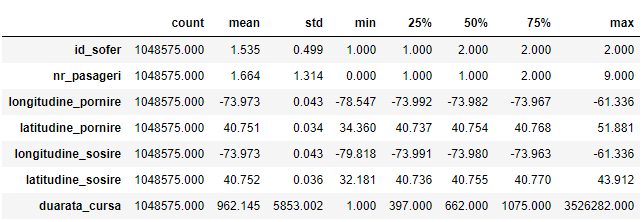


Figura 1 Descrierea datelor folosite

Conform tabelului de mai sus, se poate observa că analiza a fost facută pentru 2 șoferi, având id-ul 1 și respectiv id-ul 2. Numărul mediu de pasageri într-o mașină pentru o singură cursă este de 1.664, ceea ce ne arată că în mare parte din cazuri fost un singur pasager într-o cursă, în mai bine de 50% din curse. Numărul maxim de pasageri este de 9 pasageri într-o cursă, iar numărul minim este de 0, ambele arătând 2 valori anormale. De asemenea, anomalii se observă și în analiza coloanei ce arată durata cursei, fiind valori foarte mari.

Totodată, trebuie analizate și latitudinea și longitudinea punctului de plecare pentru a ne asigura că datele analizate sunt din interiorul orașului New York. Cunoscând că datele geografice despre New York City: latitudinea este între 40.7128 și 40.748817, iar longitudinea este între: -74.0059 și -73.968285, a fost analizată harta de mai jos pentru punctele care sunt în afara orașului New York.



Figura 2 Harta cu punctele din afara orașului New York

Prin urmare, aceste date au fost eliminate, analiza fiind efectuata exclusiv pentru cursele din interiorul orașului.

În urma eliminării anomaliilor s-a obținut tabelul următor:

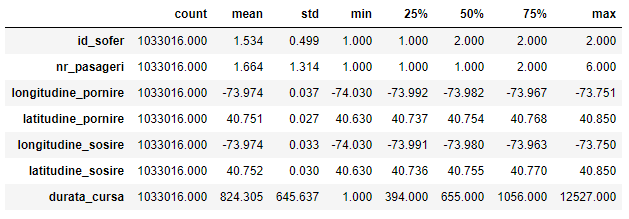


Figura 3 Rezultat în urma eliminării anomaliilor

De asemenea, se poate observa că nu mai sunt locații în afara orașului New York.



Figura 4 Harta cu punctele din afara orașului New York în urma eliminării anomaliilor

Astfel, putem afirma în final că numărul de pasageri maxim a fost de 6, iar numărul minim: 1 singur pasager. Cea mai lungă cursă a fost de 2115 secunde, ceea ce înseamnă 35.25 minute. Valoarea medie pentru durata unei curse este aproximativ 728 de secunde, aproximativ 12 minute.

În urma eliminării anomaliilor, datele sunt în număr de 1033016, ce vor fi utilizate în pașii următori pentru realizarea modelului.

Data plecării, respectiv data sosirii, conțin atât informații despre dată cât și despre ora plecării/sosirii. Mai departe, vom separa această informație în: lună, ziua săptămânii, oră pentru data plecării pentru o analiză mai ușoară. Avem nevoie de aceste informații detalitate deoarece pot influența durata cursei. Nu este necesar acest proces și pentru data plecării deoarece, ziua și ora la care pasagerul este lăsat nu va influenta durata cursei, deoarce durata este înregistrată la final. Astfel, datele cu care vom continua analiza vor avea următoarele coloane: id, id\_sofer, nr\_pasageri, longitudine\_pornire, latitudine\_pornire, longitudine\_sosire, latitudine\_sosire, cursa\_marcata, durta\_cursa, timp\_pornire, timp\_sosire, luna\_pornire, ora\_pornire, ziua\_pornire.

**Analiza exploratorie a datelor**

* **Variabila dependenta**: durata cursei

Variabila ce dorim să fie previzionată este durata cursei. Avem nevoie sa analizăm distribuția pentru a vedea dacă este nevoie de transformări asupra valorilor.

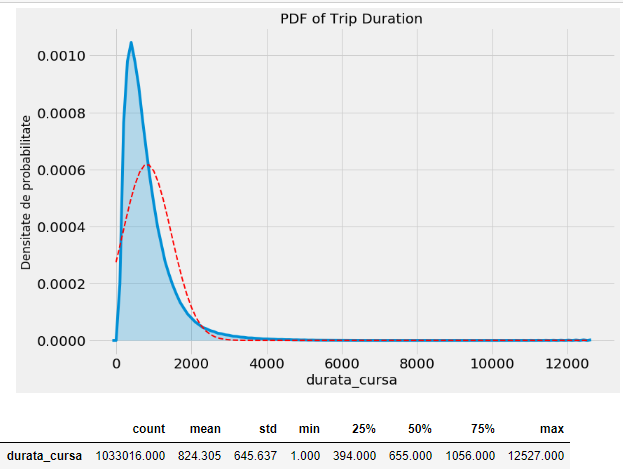


Figura 5 Curba de asimetrie pentru variabila *durata cursei*

Conform graficului prezentat se observă clar că durata\_cursa este foarte înclinată spre dreapta, având valoarea asimetriei este mai mare ca 1.0, ceea ce ne arată că distribuția este înclinată spre dreapta, având foare multe vlori extreme spre dreapta.

Mediana pentru durata cursei este de 655 secunde, aproximativ 11 minute. Totuși, se observa că sunt și curse lungi de 12527 secunde, aproximativ 208 minute, aprox. 3 ore jumătate. Aceste valori mari nu au fost șterse în pasul anterior deoarece acestea sunt totuși între cele 2 valori ale deviației standar ale mediei.

* **Variabilele independete**

Vom analiza celelalte coloane pentru identificarea celor care pot fi utilizate în model pentru a prezice durata cursei. În mod sigur, nu toate coloanele sunt utile pentru realizarea modelului, dar pot oferi informații utile pentru fundamentarea unei decizii. Variabilele selectate vor fi folosite pentru a antrena modelul de machine learning pentru a prezice variabila dependenta, *durata\_cursa*.

1. **Id**: variabila id este un identificator unic al fiecărei curse. Se va analiza cum variază aceata caractistică în timp. În mod clar, această variabilă nu influentează durata cursei, dar ne poate ajuta în analiza comenzilor pe diferite luni.

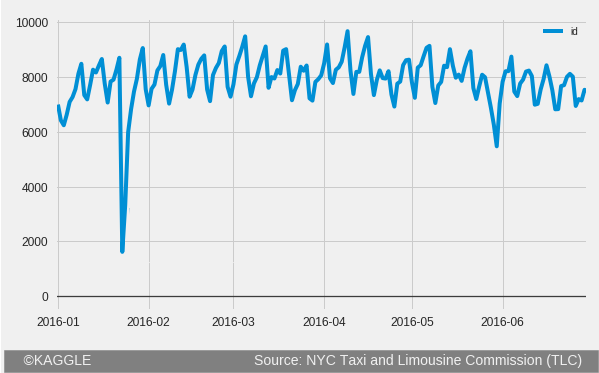


Figura 6 Evoluția comenzilor de taxi în anul 2016

Se poate observa conform graficului că este un declin al valorii la finalul lunii Ianuarie. Ca primă vedere se poate spune că este vorba de sezonalitate, dar declinul este mult prea mare față de celelalte zile de iarnă înainte și după această scădere bruscă, ceea ce ne arată că în acea zi circulația taxi-urilor a fost aproape oprită.

Deși variabila id ne arată informații utile în legătură cu traseele efectuate de firma de taxi, forma actuală a id-ului cursei nu va fi folositoare pentru a prezice durata cursei în algoritm.

1. **Id șofer**: această variabilă este un cod ce indică șoferul asociat fiecărei curse. Vom examina dacă un șofer are cursele mai lungi decât ceilalți.

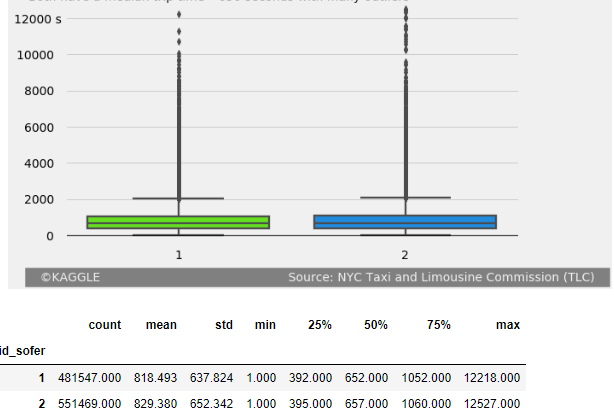


Figura 7 Analiza duratei curselor pentru fiecare șofer

Conform figurii de mai sus se observă că nu sunt diferențe mari între timpul cursei efectuat de cei doi șoferi, observându-se valorile medianelor aproximativ egale: 652 și 657. Această varabilă va fi folosită în modelul pentru prezicerea duratei cursei, având o influență mică asupra duratei cursei. Totuși, dacă ne dorim ca un client final să fie un pasager, este dificil ca acesta să cunoască id-ul șoferului.

1. **cursa\_marcata**: această variabilă indică dacă ultima cursă efectuată este înregistrată în memoria mașinii înainte de a o trimite către producptor deoarece mașina nu a avut conexiune directă către server.

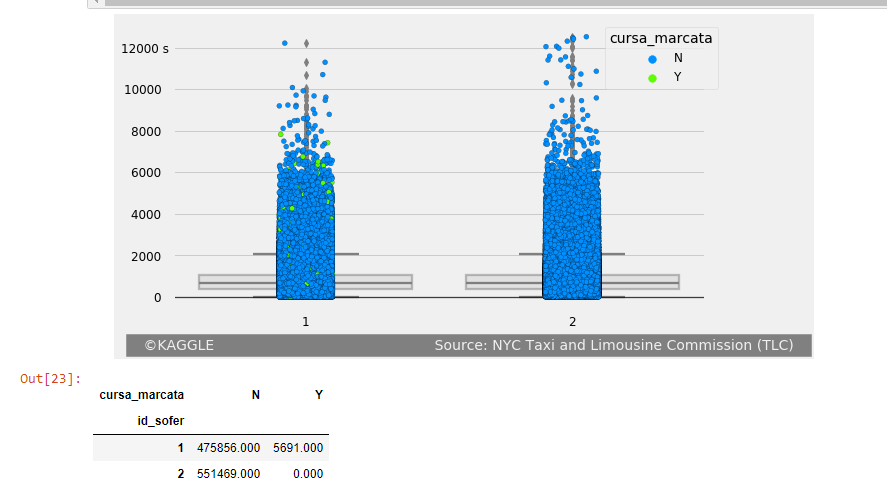


Figura 8 Grafic de tip plot pentru cursele marcate ale șoferilor

Se poate observa că doar șoferul cu id-ul 1 are curse care nu au fost marcate direct pe server. Astfel, această variabilă nu influentează cu mult predicția pentru durata cursei pentru șoferul cu id-ul 2, dar puțin pentru șoferul cu id-ul 1. Totuși, această variabilă nu influentează durata cursei și nu va fi utilizată pentru antrenarea modelului, dar în urma analizei acesteia se observă că între cei doi șoferi există o diferență mare a numărului de comenzi, durata cursei putând fi influențată și de experiența șoferului.

1. **Numărul de pasageri:**această variabilă este numărul de pasageri din mașină introduse de șofer. Prezumția este că acele curse cu mai mulți pasageri durează mai mult din cauza opririlor. Vom analiza dacă această variabilă are un impact și asupra datelor curente.

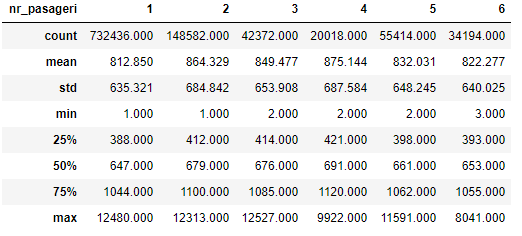


Figura 9 Descrierea variabilei: număr pasageri

Conform tabelului rezultat, se observă că nuumărul de pasageri nu are o implicație foarte mare asupra duratei cursei, mediana variind foarte puțin. Totuși, această variabilă va fi folosită în model, ținând cont de ipoteză: pot apărea mai multe opriri ceea ce duce la creșterea timpului.

1. **Durata cursei în funcție de dată și oră.** Această variabilă influențează în mod clar durata cursei, traficul variind în funcție de zi și oră. Mai departe vom analiza dacă durata cursei durează mai mult cu adevărat în anumite zile față de celelalte.

**Durata cursei în funcție de zi și oră**

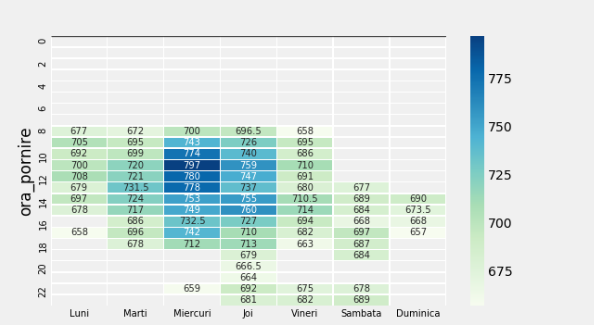


Figura 10 Durata cursei în functție de zi și oră

Așa cum a fost precizat mai sus mediana pentru durata cursei este de 655 secunde, iar zilele momentele din săptămână care tind să depășească această valoare sunt:

* luni – joi, între orele 8-18, reprezentând orele de muncă;
* joi – sâmbătă, între orele 18-24
* sâmbătă – duminică, între orele 24-1
* duminică, între orele 14-16

**Durata cursei în funcție de lună**

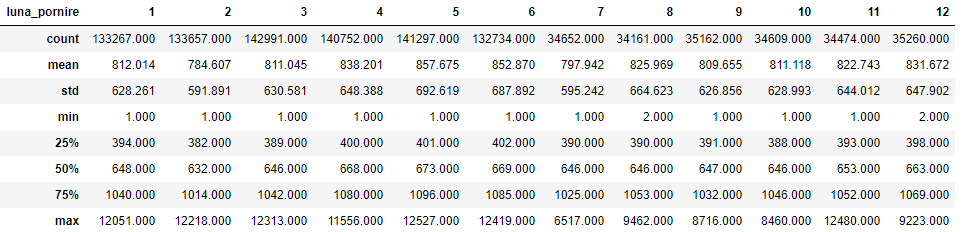


Figura 11 Durata cursei în funcție de lună

Se observă că mediana cu valoarea cea mai mare este în luna Mai. Totuși, se observă că nu variază foarte mult de la o lună la alta.Ceea ce ne arată că luna nu va fi neapărat un ajutor în predicția pentru variabila: durata\_cursa. Se poate afirma astfel, că luna are un impact minimal.

**Grafic cu punctele de plecare și sosire în orașul New York**

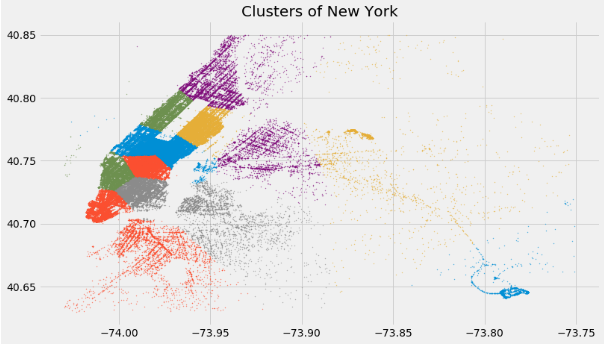
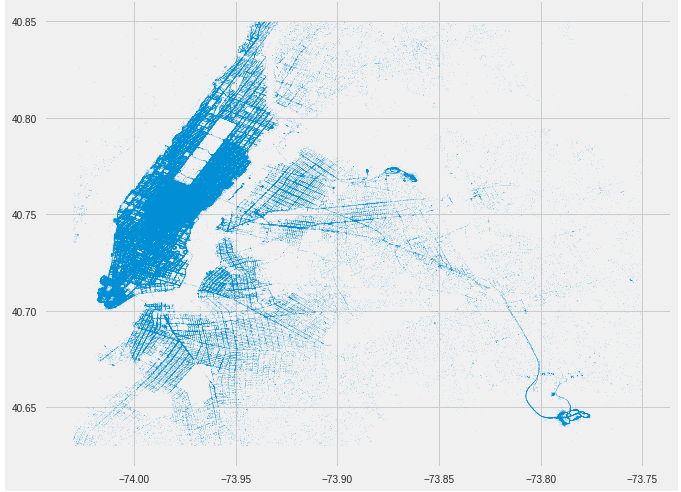


Figura 12 Grafic puncte plecare și sosire în orașul New York Figura 13 Cluster - taxi New York

În urma clusterizării se poate observa faptul că datele sunt neliniare, fiind influențate de traficul din fiecare zonă, ce nu respectă mereu un tipar anume. Astfel, prin neliniaritatea datelor vom avea nevoie de un model de regresie specific.

**Relația dintre variabile**

În urma exploatării, vom analiza în continuare corelația dintre date. Dacă coeficientul de corelație dintre variabilele independente este mai mare de 0.7 putem renunța la acele variabile, prezentând informație redundantă.

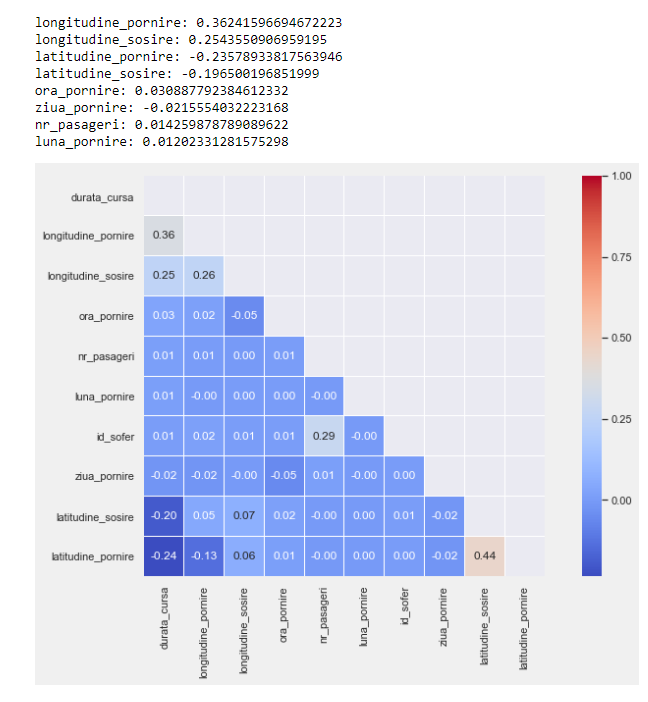


Figura 14 Corelația dintre variabile

Prin urmare, coeficientul de corelație ale coordantelor: latitudine\_plecare, longitudine\_plecare, latitudine\_sosire, longitudine\_sosire indică o relație liniară între ele. Totuși, coeficientul de corelație este mai mic de 0.5 și nu este suficient de mare încât să ștergem aceasta valoare.

## Susbsitemul de modele

Conform analizei efectuate mai sus, coloanele ce vor fi folosite pentru realizarea modelului din datele inițiale sunt: **id\_sofer**, **nr\_pasageri**, **longitudine\_pornire**, **latitudine\_pornire, longitudine\_sosire**, **latitudine\_sosire**, **luna\_pornire**, **ora\_pornire**, **ziua\_pornire** ca variabile independente, iar variabila dependeta este **durata\_cursa**.

Prin urmare, variabilele au fost alese din coloanele prezente in excel, anomaliile ce puteau duce la rezultate eronate au fost eliminate, astfel modelul fiind gata pentru implementarea modelul arborelui de regresie.

A fost ales algoritmul arobre de regresie deoarece conform analizei datelor modelul este unul non-liniar. Arborele de regresie este un model de estimare (în acest caz) folosit în situația în care se dorește o prognoză pentru o variabilă dependentă.

Pentru a antrena modelul, arborele de regresie, avem nevoie de un set de date de antrenament și un set de date de test. Pe baza datelor de antrenament se va construi arborele de decizie, iar pe baza acestuia se poate face prognoza datelor de test și astfel se pot determina metricele acestui algoritm. Astfel, vom extrage din datele inițiale 80% pentru a antrena modelul, iar 20% pentru a testa arborele de decizie rezultat. Pentru a selecta datele este folosit în continuare limbajul Python, selectând 80% din date aleator prin funcția *train\_test\_split*.

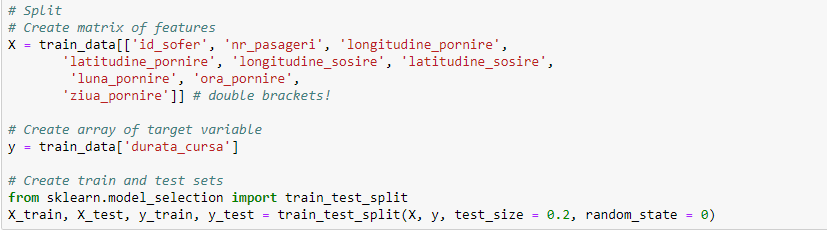


Figura 15 Separarea datelor - date antrenamet & date test

Descriere date:

* train\_data reprezintă datele analizate și modificate
* X reprezintă variabilele independente
* y reprezintă variabila dependentă
* X\_train, y\_train reprezintă datele de antrenament
* X\_test, y\_test reprezintă datele pentru testarea rezultatelor

Mai departe, modelul de arobre de decizie este antrenat, iar în urma acestei etape putem face predicții pe baza lui. În acest scop a fost folosit DecisionTreeRegressor din biblioteca sklearn în limbajul Python. Acest pachet prezintă clasa modelul *tree* cu metodele: *fit* pentru antrenarea modelului și metoda *predit* pentru a realiza predicția pe baza arborelui de decizie.

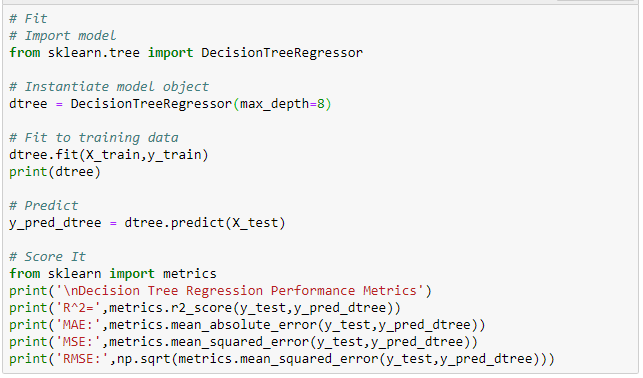


Figura 16 Antrenare arbore de decizie

Paramtrii cu care a fost apelată metoda pentru arborele de decizie sunt următorii:

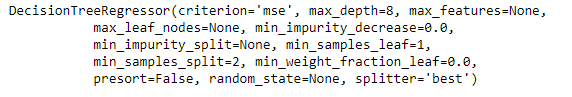


Figura 17 Parametri pentru arbore de decizie

Astfel, calculând metricele pentru date, comparând predicția dată de model cu rezultatele deja cunoscute, am determinat coeficientul de determinare (R^2), media absoluta a erorilor.

În urma analizei, am obținut R^2 = 0.62122, ceea ce ne arată că 62.12% din variația duratei cursei este explicată prin variația simultană a șoferului, a numărului de pasageri, a punctului de pornire, a punctului de oprire și a momentului pornirii (luna, ora, ziua).

Totuși acest model poate fi îmbunătățit pentru rezultate mai bune. Prin urmare, vom adăuga încă o variabilă ce poate ajuta în mod evident modelul, distanța dintre cele două puncte. Latitudinea și longitudinea punctelor de plecare și sosire oferă informații referitoare și la trafic pe baza datelor din trecut, și totodată și pentru distanță. Dar, pentru a ajuta modelul prin a da predicții cât mai realiste, vom calcula distanța dintre cele două puncte.

În urma modificărilor, se observă clar o îmbunătățire a modelului deoarece R^2 este în valoare de 0.6981, 69.81% din variația duratei cursei este explicată prin variația simultană a șoferului, a numărului de pasageri, a punctului de pornire, a punctului de oprire, a momentului pornirii (luna, ora, ziua) și a distanței. Eroarea, RMSE, este în valoare de 353.03 secunode, aproximativ 5.88 minute.

Totodată un model care se bazeaza pe arbori de decizie este *Random Forest*. Cum traficul este grupat și agregat în funcție de densitate pe diferite locații, locația cursei are în mod clar impact asupra duratei cursei. Random Forest este un algoritm ce poate da predicții pentru modele nonliniare. Acesta însumează mai mulți arbori de decizie construiți pe baza datelor de antrenament în loc de a utiliza un sigur arbore de decizie, imbunătățind astfel acuratețea predicției. Fiecare dintre acești arbori de decizie au ca scop împarțirea spațiului predictor, adică setul de variabile independente (x1, x2, ..., xn) în distincte regiuni (R1, R2, ..., Rj) care minimizează suma reziduurilor.

Astfel, acest model a fost utilizat din biblioteca *sklearn* în limbajul de programare Python.



Figura 18 Antrenare Model Random Forest

Apelul, este similar celui pentru arborele de decizie. Mai întai a fost antrenat modelul cu 80% din datele inițiale, iar restul de 20% au fost folosite pentru testare. În urma testării s-au obținut rezultatele: R^2 = 0,7986 ceea ce înseamnă că 79.86% din variația duratei cursei este explicată prin variația simultană a șoferului, a numărului de pasageri, a punctului de pornire, a punctului de oprire, a momentului pornirii (luna, ora, ziua) și a distanței. RMSE este în valoare de 302.76 secunde, aproximativ 5 minute.

În continuare, pentru a ne folosi de modelul rezultat, îl vom testa, citind dintr-un fișier de tip csv date ce conțin informații referitoare doar la variabilele independente, lăsând modelul să prezică variabila dependentă, durata cursei. Se presupune că datele într-o aplicație reală sunt trimise de o aplicație externă.

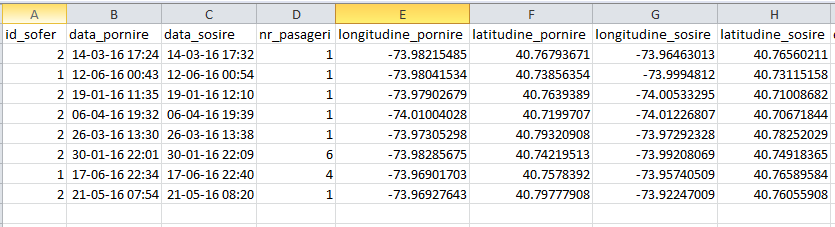


Figura 19 Tabel date test

Pentru datele din figură rezultatele predicției sunt:

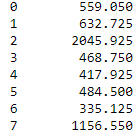


Figura 20 Rezultat predicție

Concluzionând, modelul Random Forest a prezentat rezultate mai bune decât regresia folosind modelul arbore de decizie (arbore de regresie). Totuși, ambii algoritmi ne ajută în estimarea unei valori a duratei cursei în funcție de parametrii prezentați, iar în urma acestei predicții atât firma de taxi, cât și clientul pot lua o decizie: dacă se merită cursa cu taxi-ul, ce client este mai rentabil, viteza și atenția cu care să circule șoferul și totodată prețul cursei poate fi determinat în funcție de distanță și timp. Pentru a îmbuntăți acuratețea predicției mai multe variabile trebuie luate în considerare și modelate.

## Subsistemul interfeței cu utilizatorul

În interfață utilizatorul, în această situație compania de taxi are posibilitatea de a urmări informații statistice prezentate anterior, precum: curba de asimetrie, harta cu toate cursele efectuate, durata curselor în funcție de zi, grafic ce conține cursele din orașul New York grupate pe cartiere. Poate vedea cursele efectuate și sub formă de hartă pentru a identifica locațiile exacte.

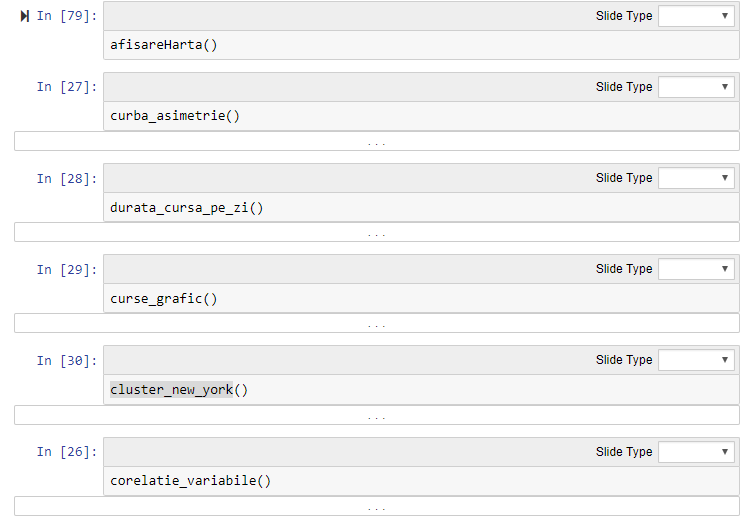


Figura 21 Opțiuni afișare grafice

Pe lângă informțiile statistice, utilizatorul are posibilitatea de a introduce un fișier de tip csv pentru a fi citit (denumirea acestuia). În urma citirii acestui fișier, se calculeaza distanța medie dintre punctele introduse de utilizator.



Figura 22 Opțiune citire fișier pentru predicție

Fișierul de tip CSV are structura prezentată mai sus. Astfel, acesta poate fi importat manual de o persoană de la firma de taxi și să ia decizii pe baza acestuia, sau să fie importat de un sistem automat care să comunice cu această aplicație și cu o aplicație pe dispozitivul șoferilor.

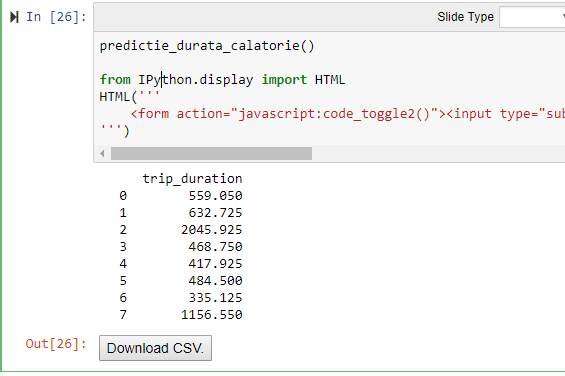


Figura 23 Rezultat predicție

În urma realizării predicției, utilizatorul are posibilitatea de a salva rezultatele într-un fișier de tip CSV, ce va fi descărcat pe calculator pentru o evidență a acestora.

# Concluzii

Concluzionând, obiectivul acestui proiect a fost atins, acela de a construi un model de machine learning care să prezică durata călătoriei pentru o cursă cu o mașină de taxi în orașul New York City, bazându-se pe date din anul 2016 culese pentru firma NYC Yellow Cab. Astfel, au fost comparate două modele pentru date neliniare pentru a avea o predicție cât mai buna, iar modelul ales în final este *Random Forest*.

În urma analizei prezentate, s-au determinat variabilele independete care pot influența cu adevărat variabila dependetă durata cursei din datele inițiale. În urma eliminării anomaliilor, datele au fost pregătite pentru a putea fi realizat modelul *Random Forest*.

Orașul New York prezintă milioane de curse în fiecare lună care duc la trafic regulat, blocaje ale străzilor sau o scară larăg de evenimente. Astfel, este foarte important pentru firmele de taxi-uri să poată avea o vizibilitate a taxelor și a duratei curselor. Prin predicția duratei unei curse poate ajuta un client să ia decizia dacă să înceapă sau nu călătoria sau totodată poate ajuta firma de taxi să ia decizia care dintre două potențiale curse este mai profitabilă. De asemenea, predicția referitoare la timpul călătoriei poate determina totodată un preț estimativ al acesteia și poate atrage mai mulți clienți în timp.

Astfel, este necesară estimarea unei valori a duratei cursei în funcție de parametrii prezentați, iar în urma acestei predicții atât firma de taxi, cât și clientul pot lua o decizie: dacă se merită cursa cu taxi-ul, ce client este mai rentabil, viteza și atenția cu care să circule șoferul și totodată prețul cursei poate fi determinat în funcție de distanță și timp.

# Bibliografie

1. Kelareva, Elena. "Predicting the Future with Google Maps APIs." Web blog post. Geo Developers Blog, [https://maps-apis.googleblog.com/2015/11/predicting-future-with-google-maps-apis.html Accessed 15 Dec. 2016](https://maps-apis.googleblog.com/2015/11/predicting-future-with-google-maps-apis.html%20Accessed%2015%20Dec.%202016).
2. Vanajakshi, L., S. C. Subramanian, and R. Sivanandan. "Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses." IET intelligent transport systems 3.1 (2009): 1-9..
3. <https://console.cloud.google.com/marketplace/details/city-of-new-york/nyc-tlc-trips?filter=solution-type:dataset&filter=category:encyclopedic>
4. <http://cs229.stanford.edu/proj2016/report/AntoniadesFadaviFobaAmonJuniorNewYorkCityCabPricing-report.pdf?fbclid=IwAR3iOURhw8U5KlrN8hdKaC6bPPNJ8vpo33mrbdYioxN559YNLSUfvmzRquA>
5. James, G, D Witten, T Hastie, and R Tibshirani. An introduction to statistical learning. Vol. 6. , New York, Springer., 2013
6. Friedman, J, T Hastie, and R Tibshirani. The elements of statistical learning. Vol. 1. , Berlin, Springer, 2001.