

***Academia de Studii Economice***

***Facultatea de Cibernetică, Statistică și Informatică Economică***

**Sistem suport de decizie pentru fundamentarea deciziilor pentru o companie de taxi**

**Sora Cornel**

Cuprins

[1. Introducere 2](#_Toc530262103)

[2. Componența sistemului decizional 3](#_Toc530262104)

[2.1 Subsistemul de date 4](#_Toc530262105)

[2.2 Subsistemul de modele 8](#_Toc530262106)

[2.3 Subsistemul interfata cu utilizatorul 26](#_Toc530262107)

[3. Concluzii 26](#_Toc530262108)

# 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 2018 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:

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

**Informații despre arhitectura proiectului**

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.

**Informații despre date**

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

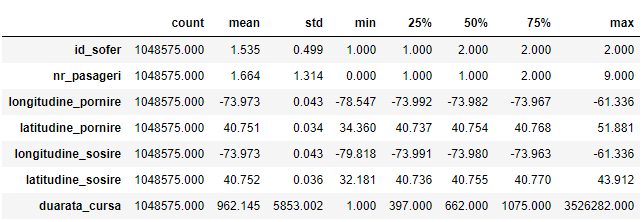
|  |  |
| --- | --- |
| **Nume variabilă** | **Descriere** |
| id | Identificator unic pentru fiecare cursa |
| id\_sofer | Identificator asociat soferului care a preluat cursa |
| data\_pornire | Data la care a fost solicitată cursa |
| data\_sosire | Data la care s-a încheiat cursa |
| nr\_pasageri | 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 |

# Componenta sistemului decizional

## 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:



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.

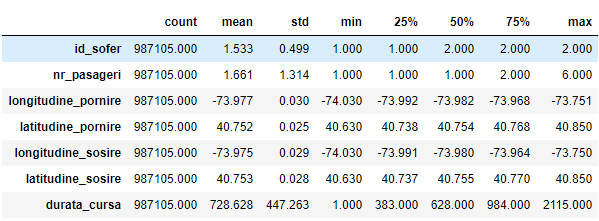


Cunoscând că datele geografice despre New York City:

* Latitude is between 40.7128 and 40.748817
* Longitude is between - 74.0059 and - 73.968285

Se observa astefel ca foarte multe date sunt eronate.

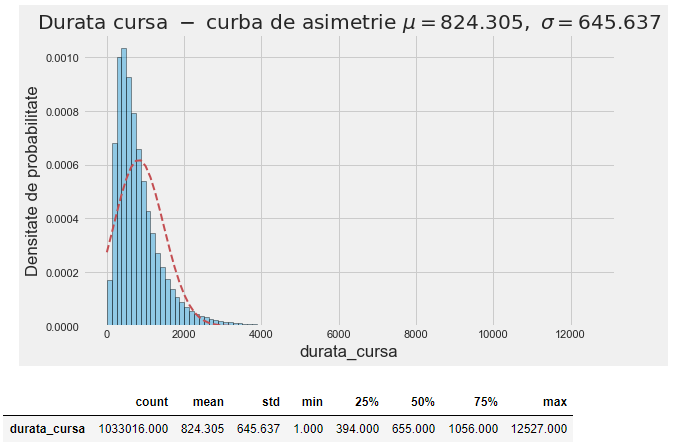
Astfel, datele vor fi modificate, pentru a elimina anomaliile prezentate.



In urma eliminarii anomaliilor, datele sunt in numar de 1033016, ce vor fi utilizate mai departe pentru realizarea modelului.

Mai departe vom avea nevoie de valorile detaliate referitoare la data plecării, separând ora, ziua, luna

The target variable we are trying to predict is trip\_duration. I'll need to examine its distribution to see if there are transformations that need to be applied.



There is a clear indication that trip\_duration is highly skewed to the right based on two key signals:

* Skewness value > 1.0
* Long right tail

The median trip\_duration is only 655 seconds (~11 min). However, there are trip\_duration observations as high as 11,411 sec (3.12 hours) that were not removed previously as they are still within two standard deviations of the mean (our specified outlier cutoff). As a result, these high trip\_duration observations are skewing the distribution to the right.

Thus, applying the log transformation to trip\_duration will normalize its distribution and reduce the influence of these high observations in the right tail.