**LABORATOR – TEHNICI DE DATA MINING – *CLASIFICARE – ARBORI DE DECIZIE***

|  |
| --- |
| **Nume student: Chanchian Armin Andrei**  **Adresa email: armin.chanchian@s.utm.ro**  **Grupa: 301**  **Data: 19.03.2024** |

# FISA DE LUCRU: CLASIFICARE – ARBORI DE DECIZIE

## Problema

Sa se construiasca un model de clasificare care sa estimeze care sunt categoriile de pasageri care au supravietuit scufundarii vasului Titanic.

### Descrierea problemei

Scufundarea Titanicului este una dintre cele mai faimoase evenimente de acest gen din istorie.

La 15 aprilie 1912, în timpul călătoriei sale inaugurale, vasul de croaziera Titanic, considerat "de nescufundat", s-a scufundat după ce s-a ciocnit cu un iceberg. Din păcate, nu au existat suficiente bărci de salvare pentru toată lumea la bord, ceea ce a dus la moartea a 1502 din 2224 de pasageri și echipaj.

Deși se pare ca unele persoane au fost mai norocoase decat altele si au supravietuit, se pare că unele grupuri de oameni aveau mai multe șanse să supraviețuiască decât altele.

Construiți un model predictiv de clasificare care să răspundă la întrebarea: "ce fel de oameni au mai multe șanse să supraviețuiască folosind datele pasagerilor (adică numele, vârsta, sexul, clasa socio-economică etc.) ce se regasesc in colectia de date titanic.csv.

De la <https://www.kaggle.com/c/titanic>

### Atribute: 12 atribute (Survived, Name, PassengerID) descriere mai jos

Atributul tinta este Survived (daca persoana a supravietuit sau nu. 0 = No, 1= Yes)

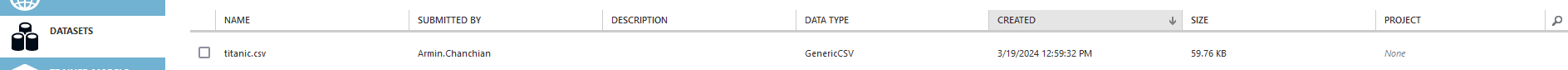
A screenshot of a computer

Description automatically generated

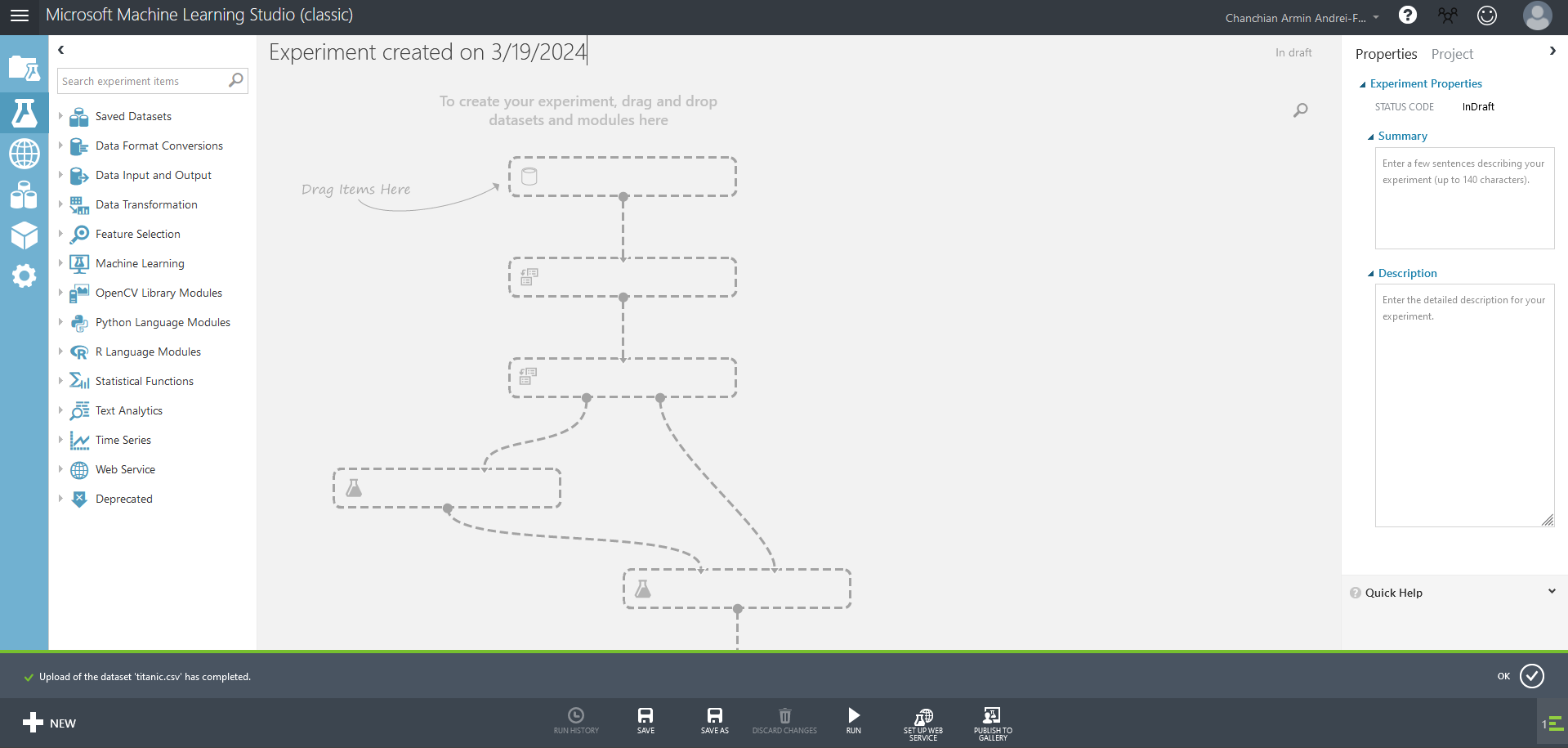
A screenshot of a white page

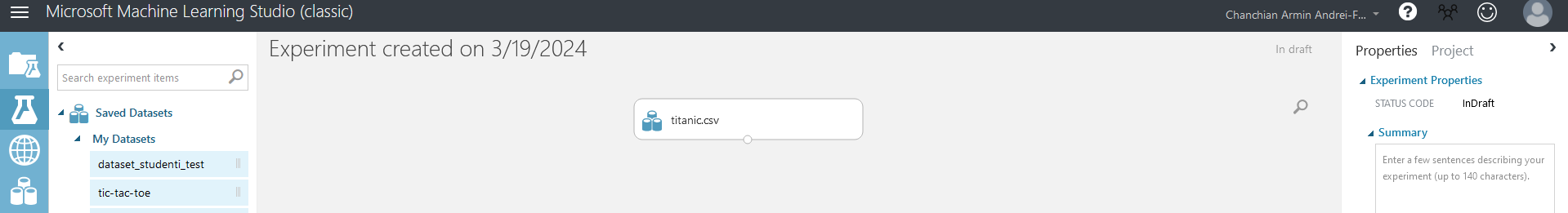
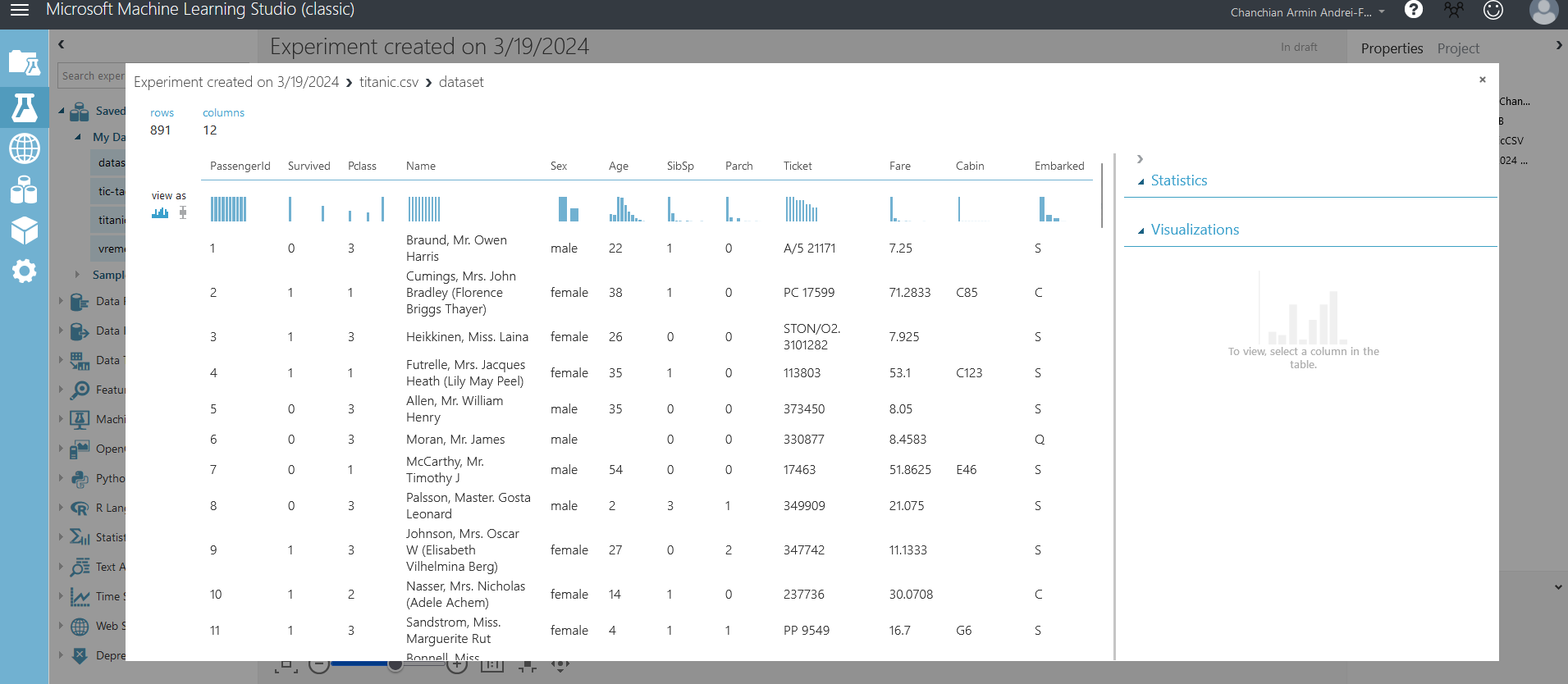
Description automatically generated

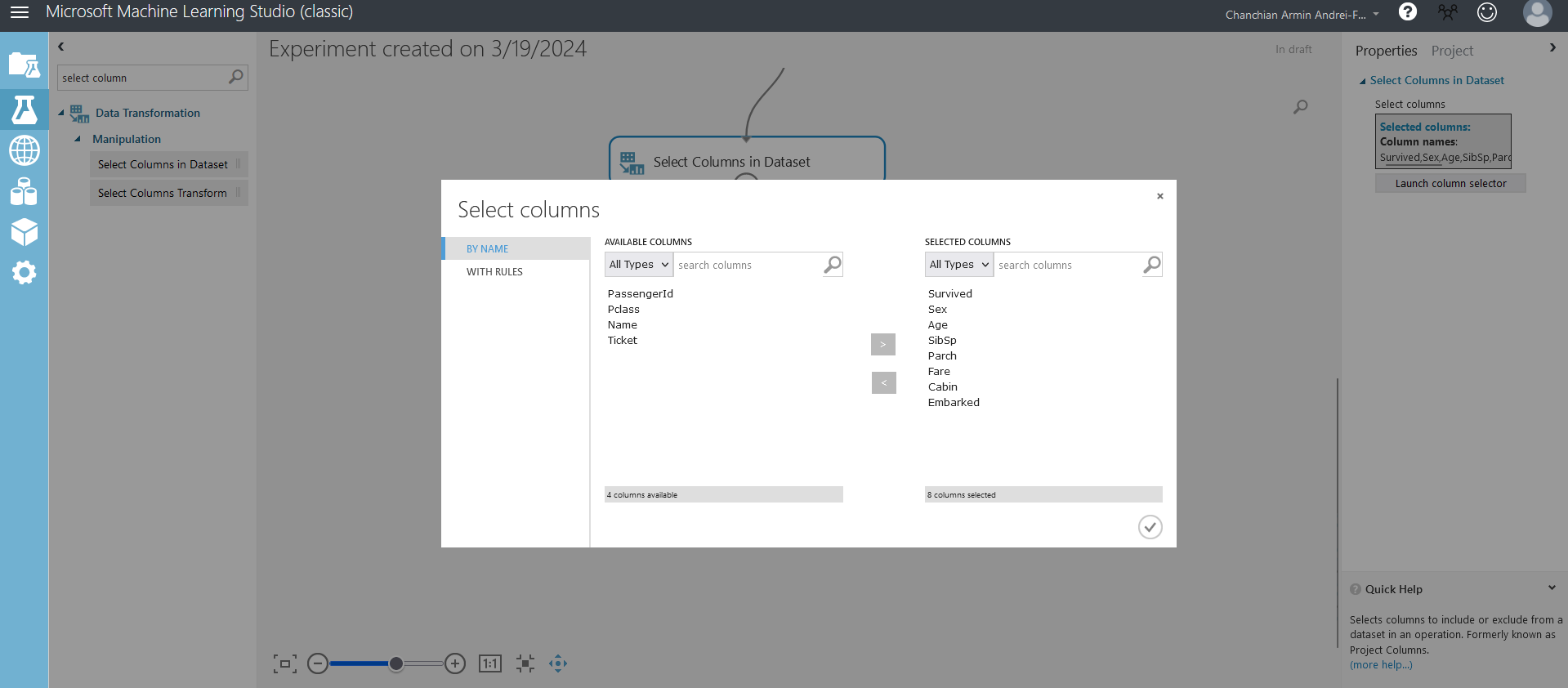
### Pasi de urmat

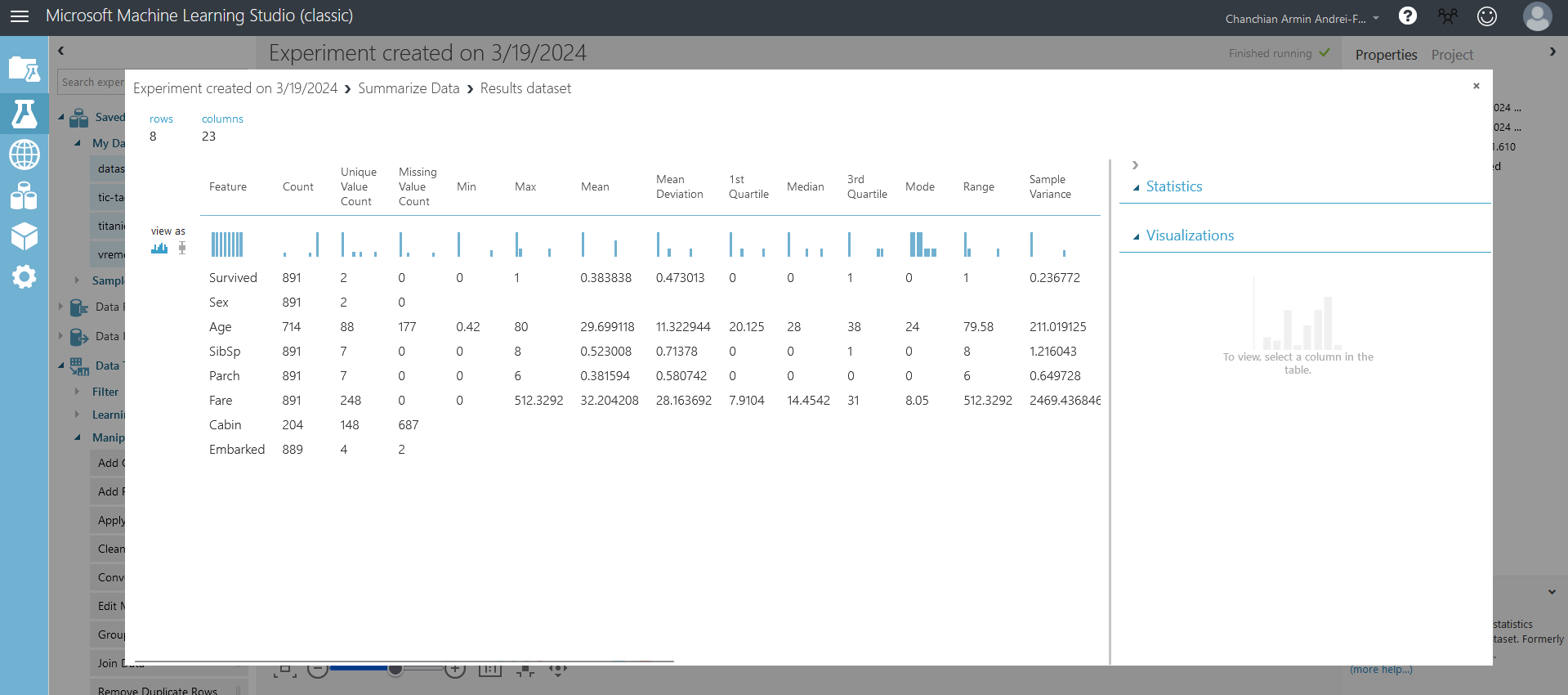
1. Sa se incarce colectia de date titanic.csv in Azure ML 

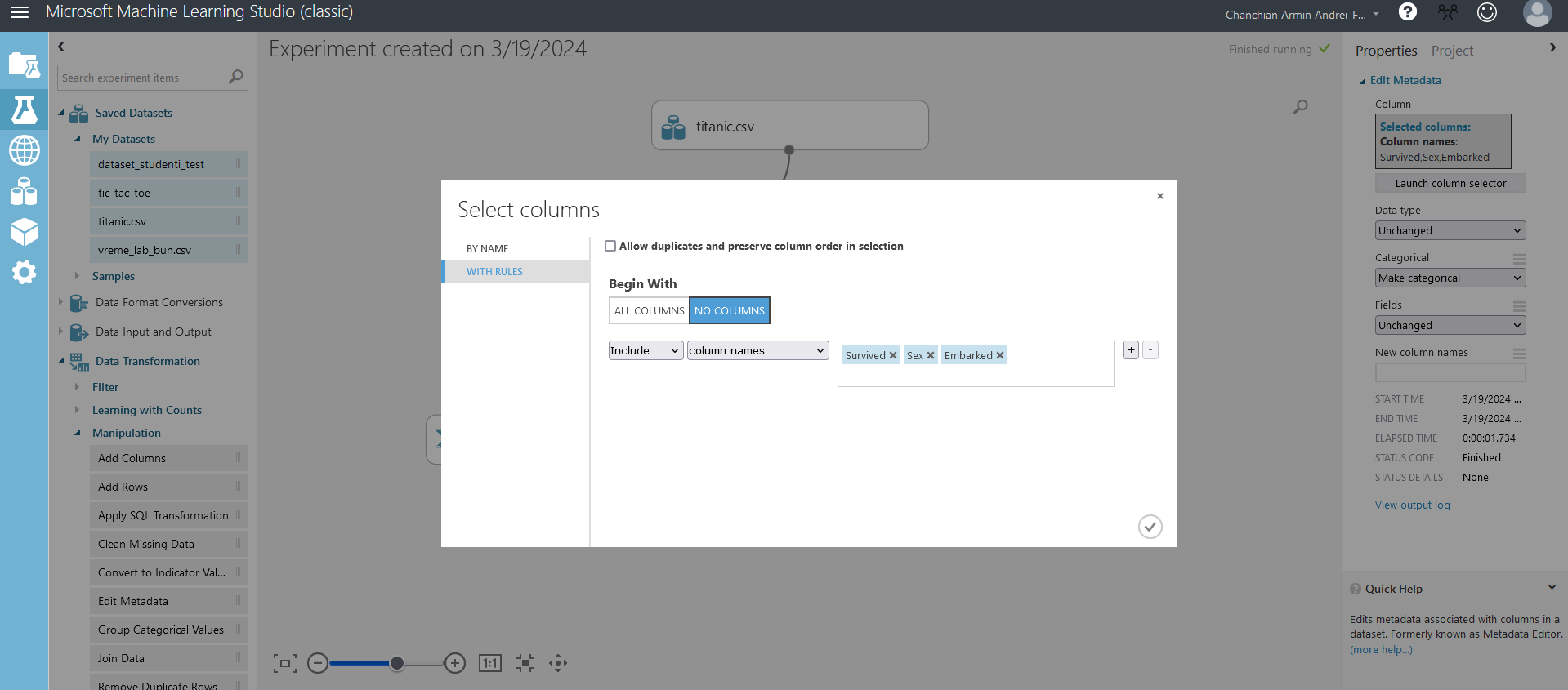
1. Se deschide un experiment nou.

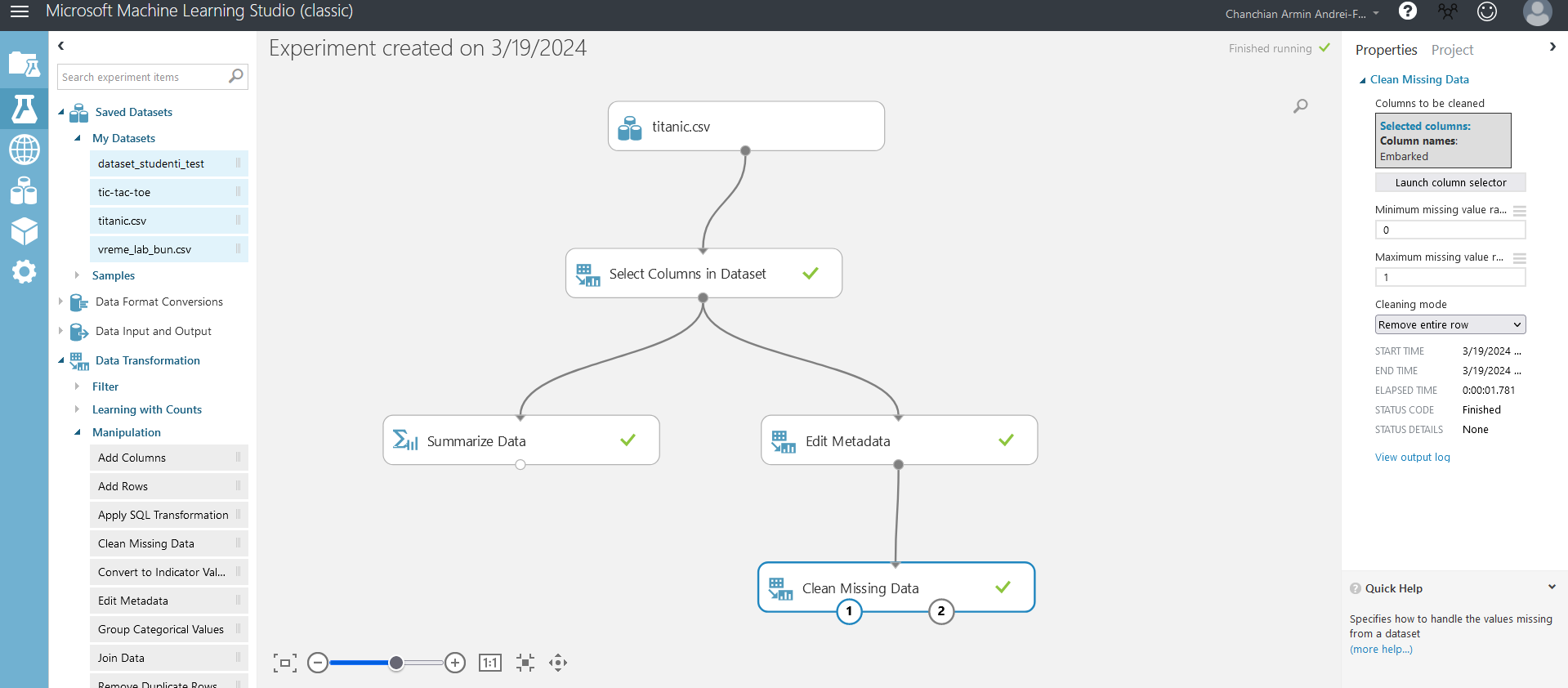


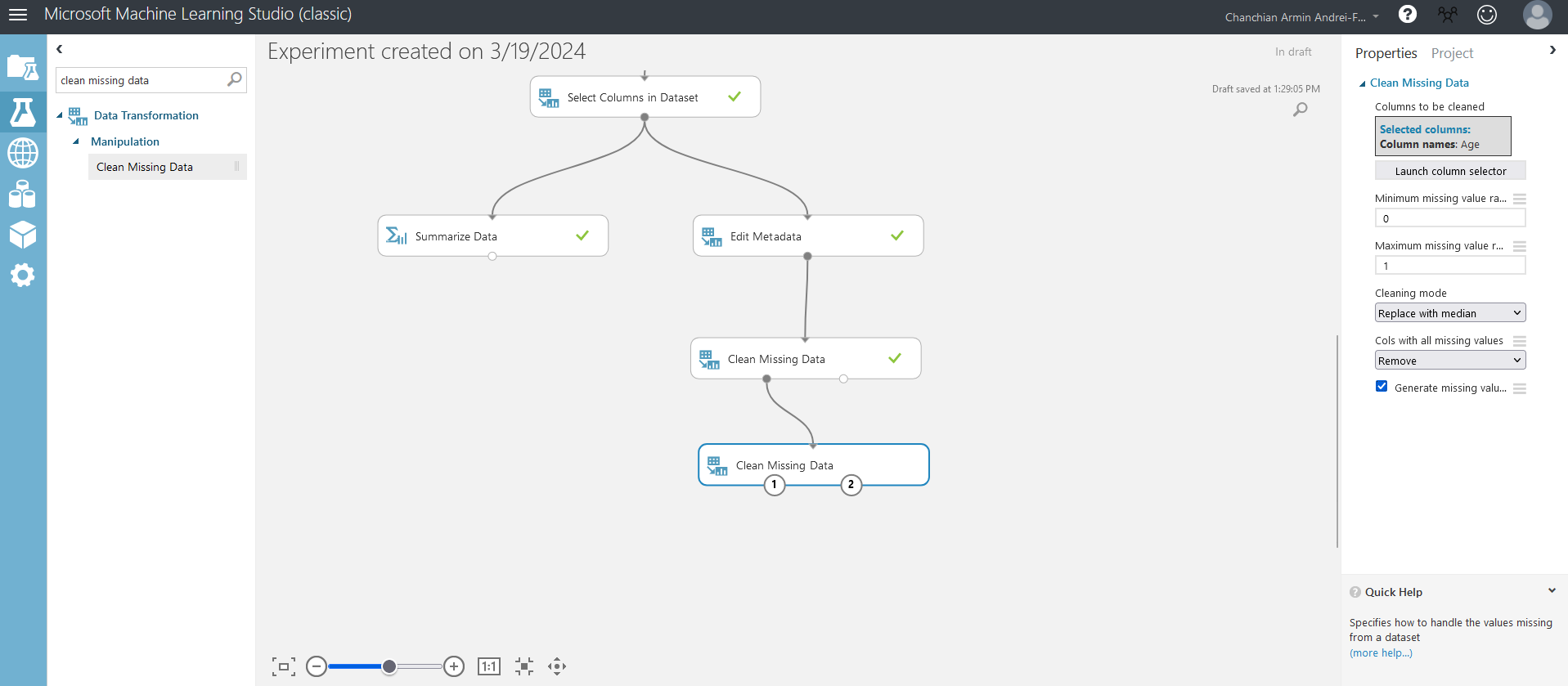
1. Se incarca colectia de date in experiment.
2. Vizualizare
3. Excludem coloanele Name, PassengerID, Ticket, Cabin

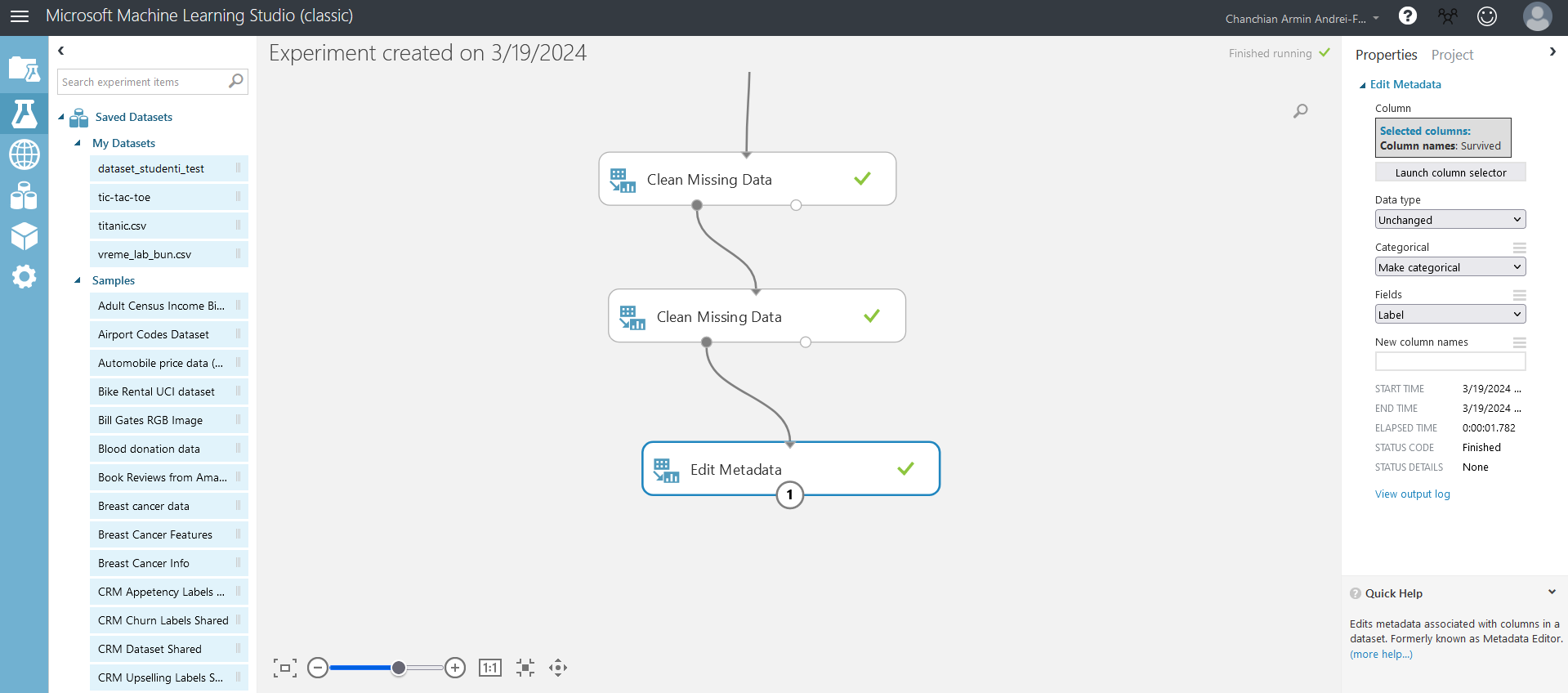
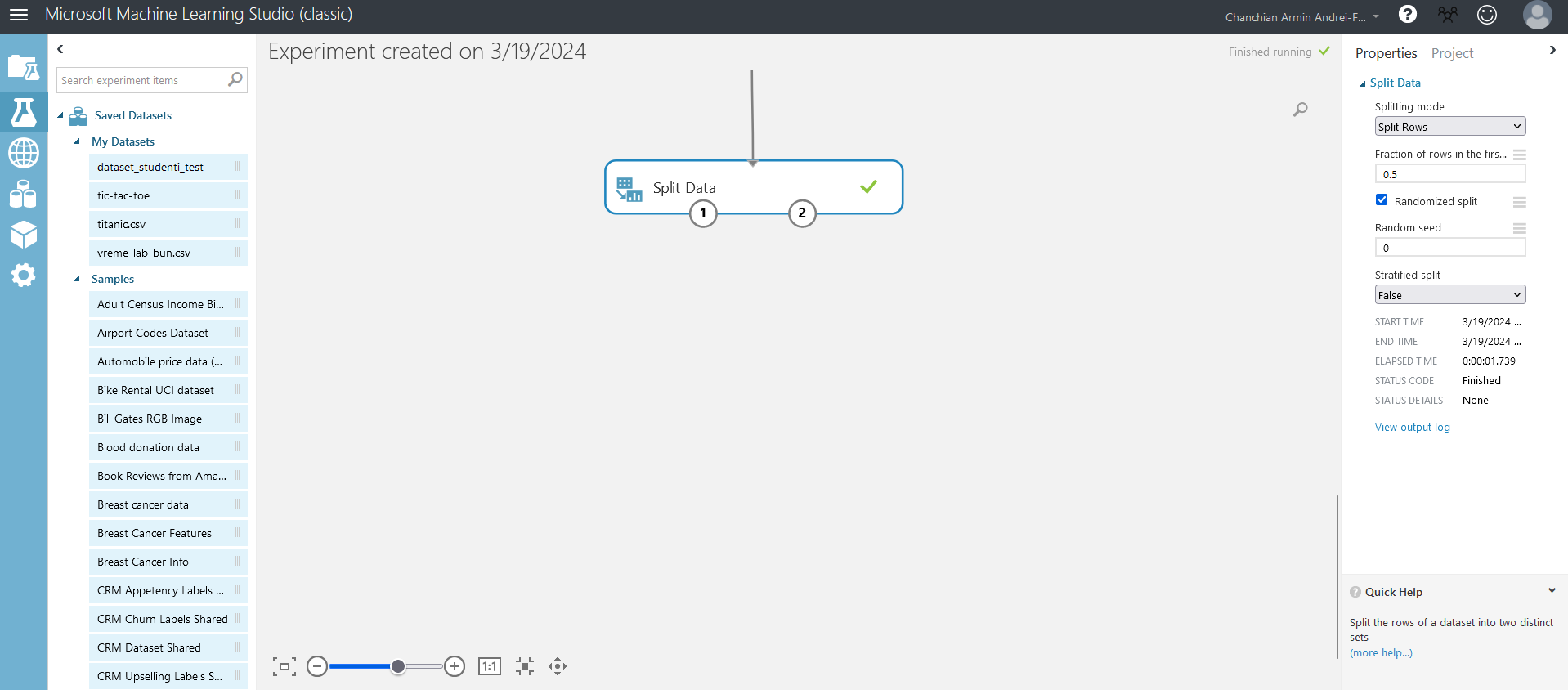


1. Analizam colectia cu Summarize data
2. Atributele care sunt de tip string se transforma in Categorical cu Edit Metadata

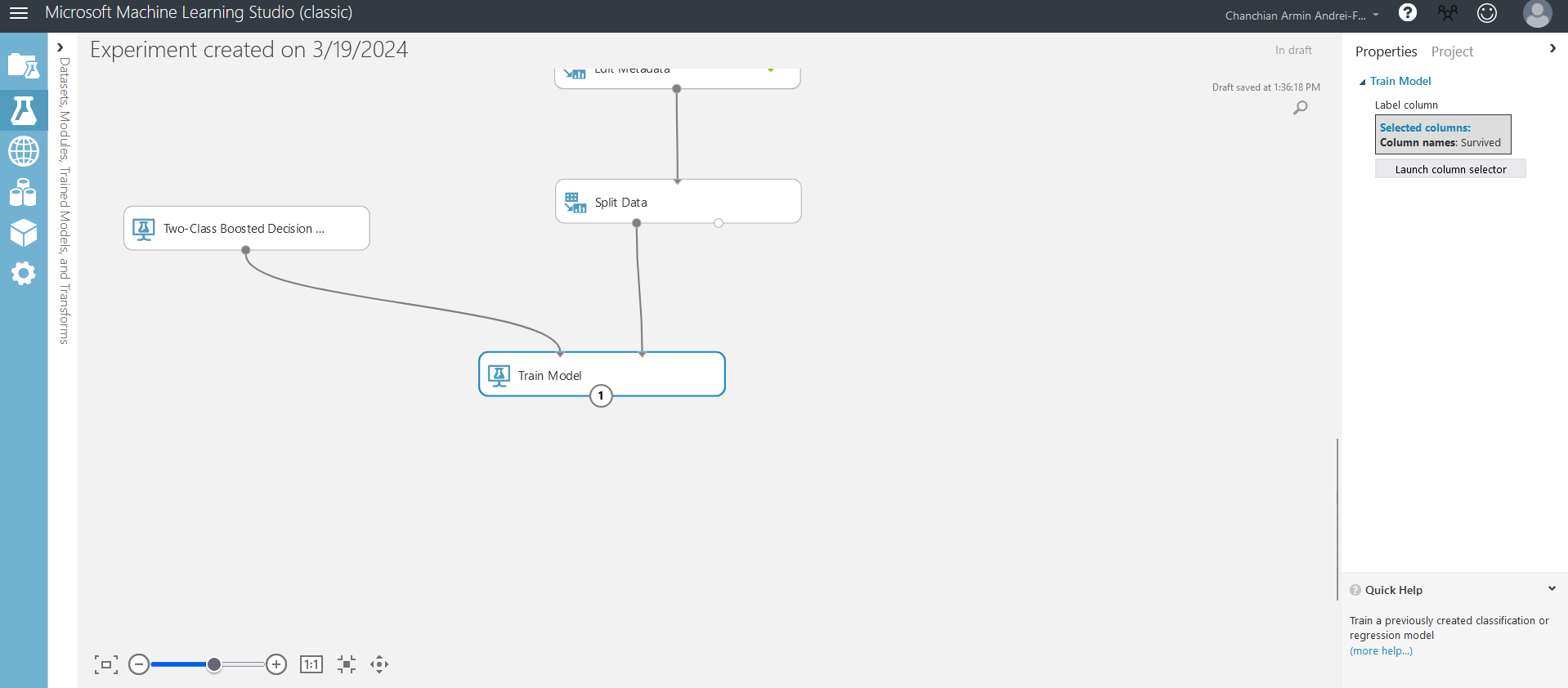


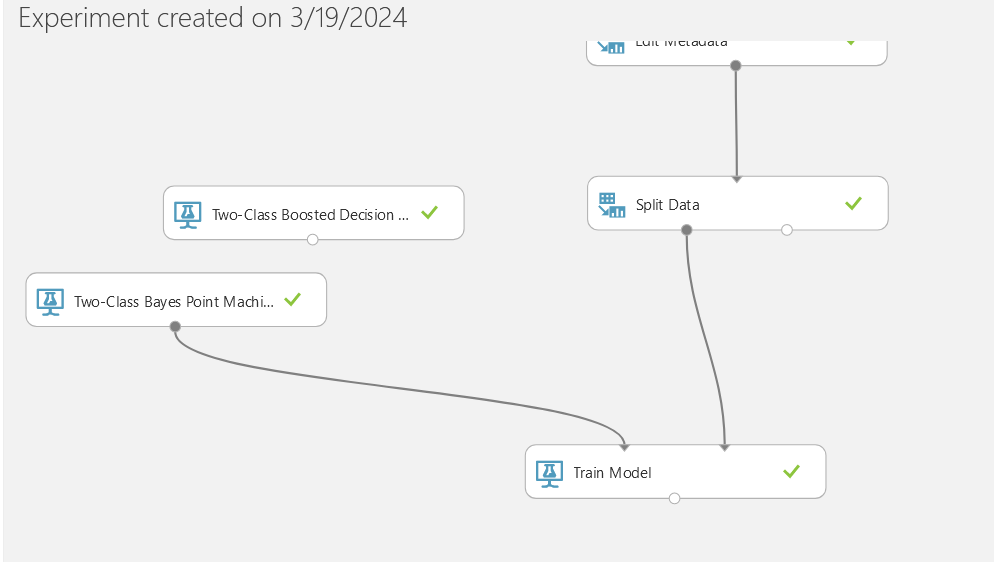
1. Identificam datele lipsa
2. Tratam datele lipsa, folosind Clean Missing Data: fie prin inlocuire cu media, fie stergem exemplele

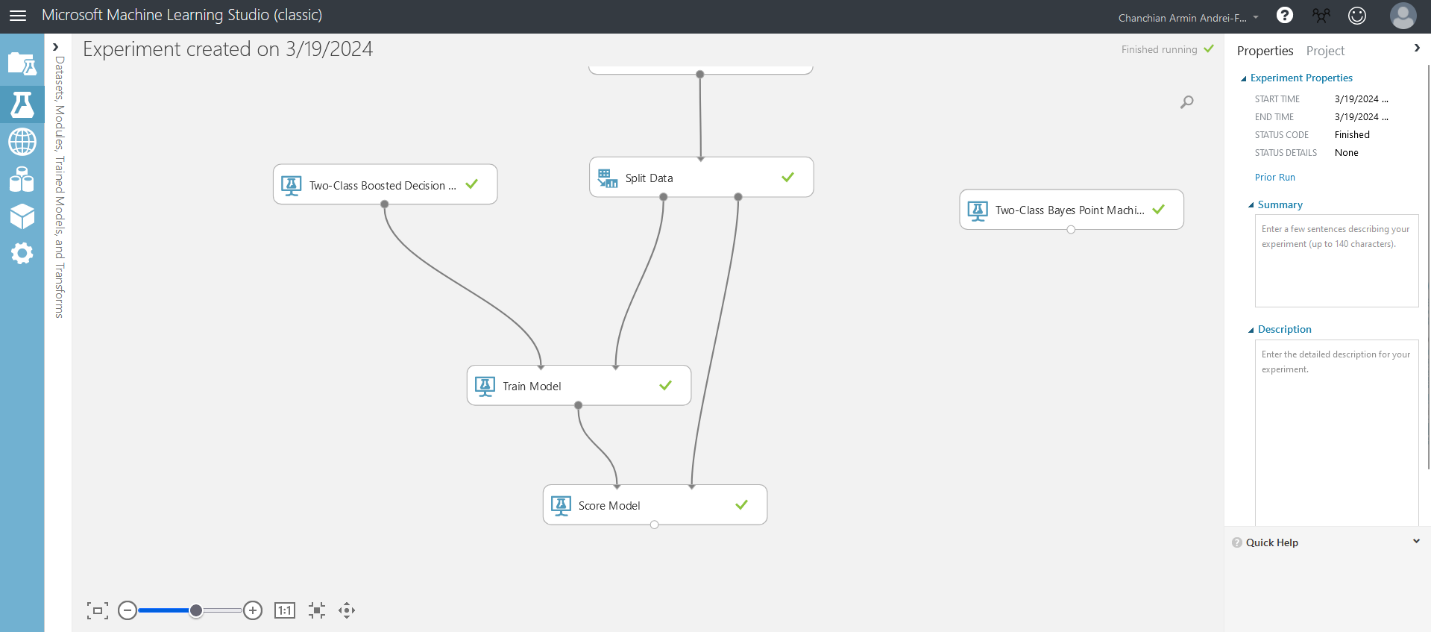


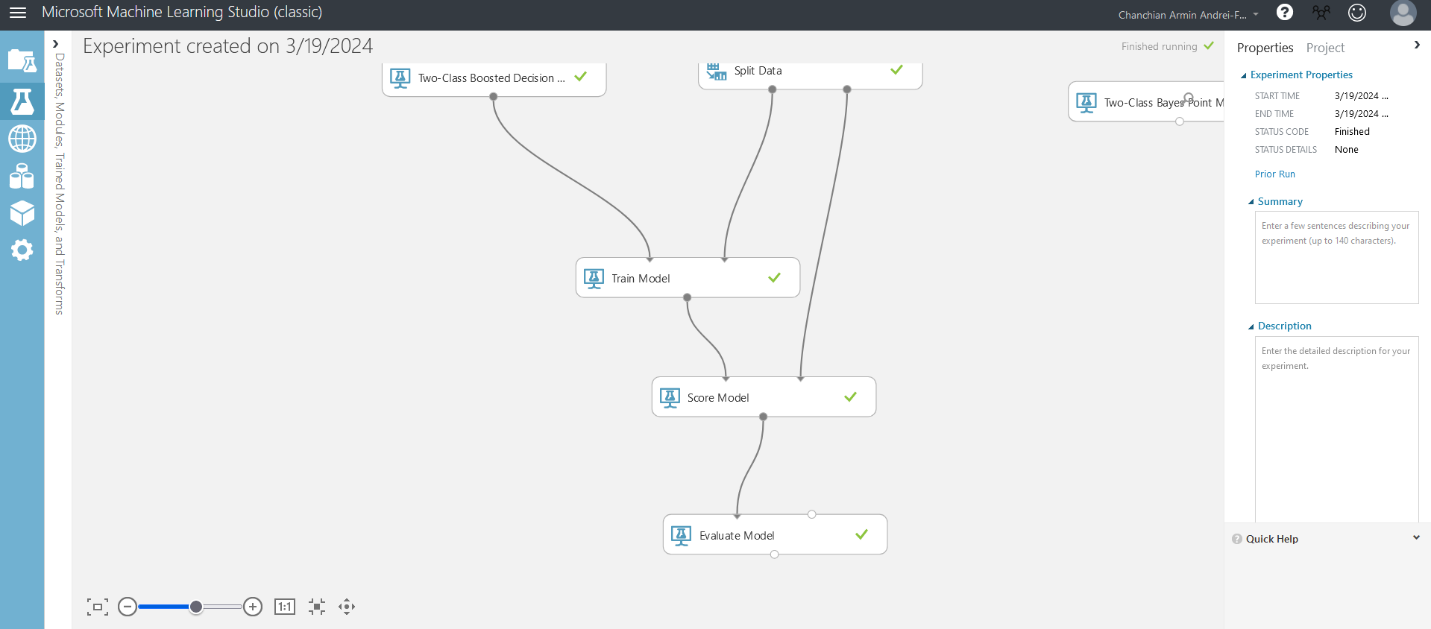
1. Cu Edit Metadata: definim atributul Survived cu Label
2. Split data pentru a selecta mt training si mt testare
3. Train Model. Aplicam algoritmii:

* Two class Boosted Decision Tree pt diferite valori ale parametrilor

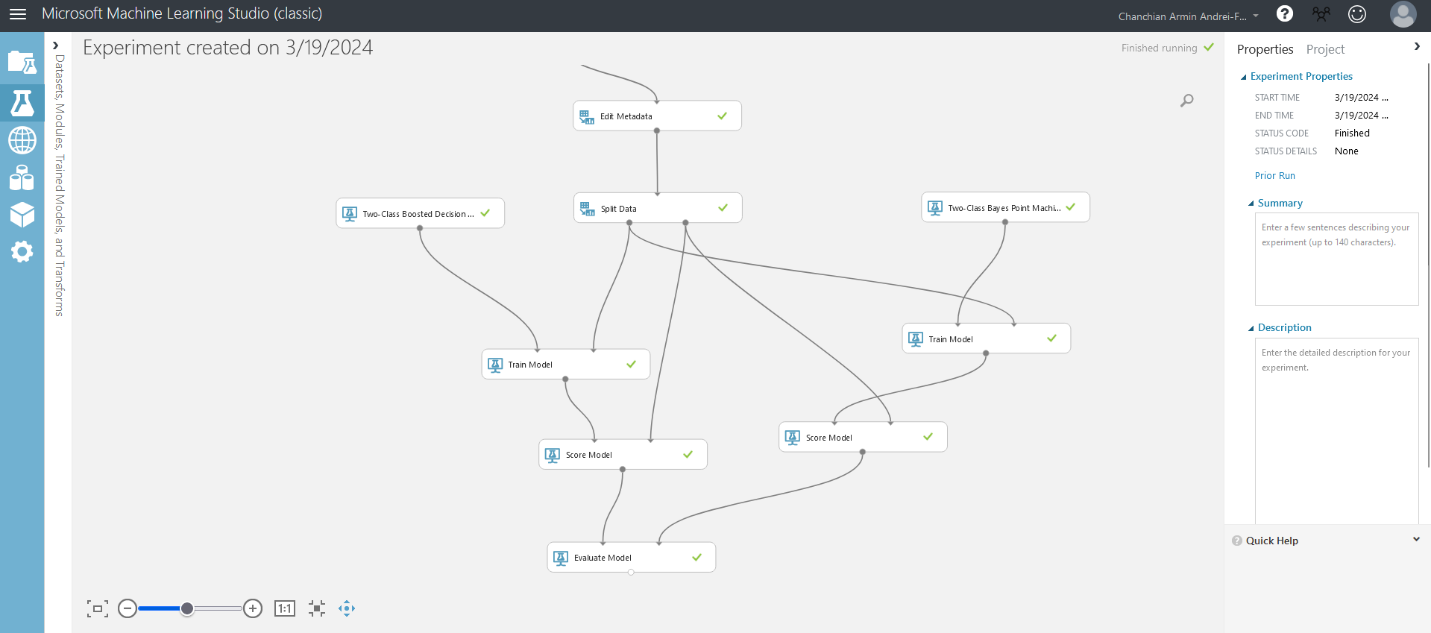


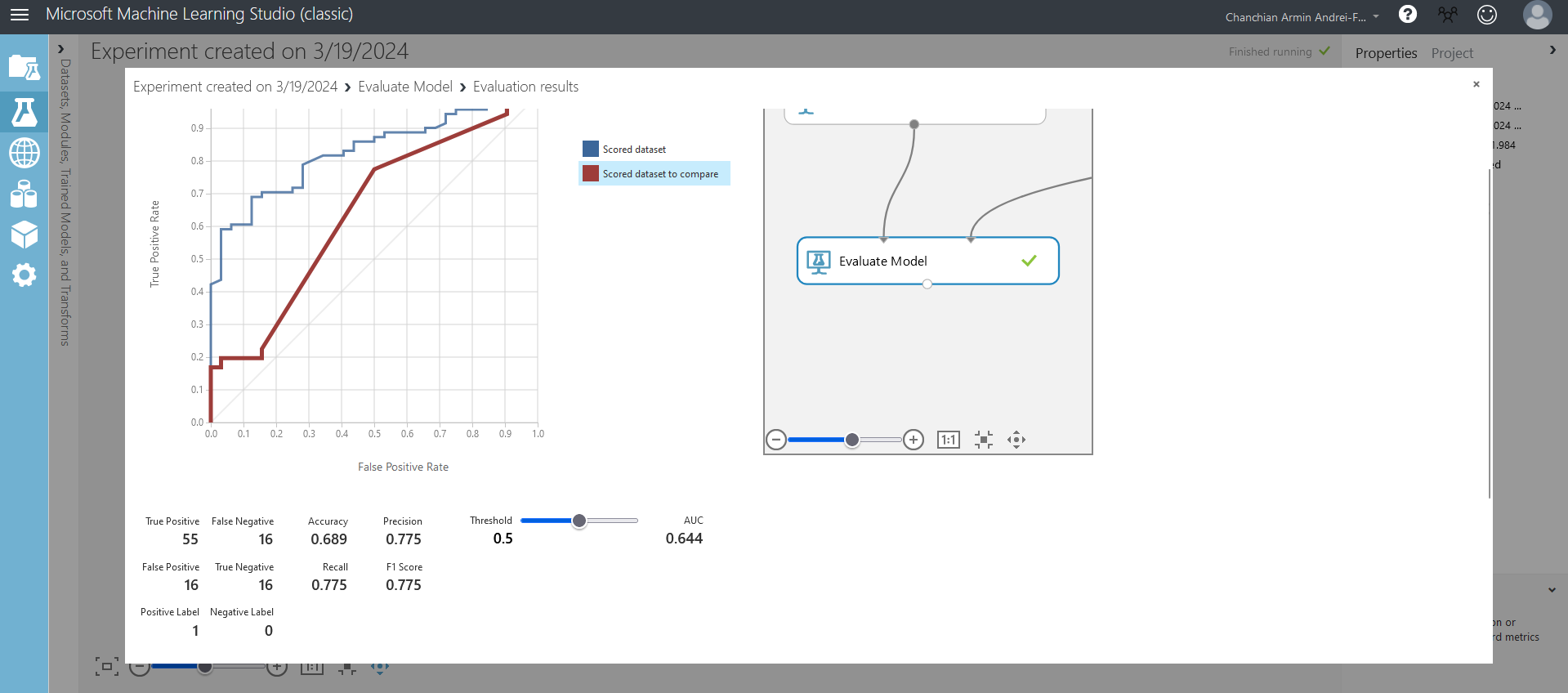
* Two-class Bayes point Machine

1. Score Model
2. Evaluate Model



1. Afisati matricea de confuzie

--



1. Completati tabelul

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritm | Procent mt training | Diversi parametri | Acuratete | Precizie |
| Two class Boosted Decision Tree | 0.5 | - | 0.75 | 0.82 |
| Two-class Bayes point Machine | 0.5 | - | 0.68 | 0.77 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |