În domeniul sănătății, evaluarea calității îngrijirii pacienților este crucială pentru îmbunătățirea serviciilor medicale. Acest raport analizează un set de modele de învățare automată dezvoltate pentru a prezice și clasifica calitatea îngrijirii pacienților pe baza datelor disponibile. Modelele folosite includ Random Forest Regressor, Random Forest Classifier, SVM (Support Vector Machine) Classifier, Decision Tree Regressor și Classifier, precum și o rețea neurală.

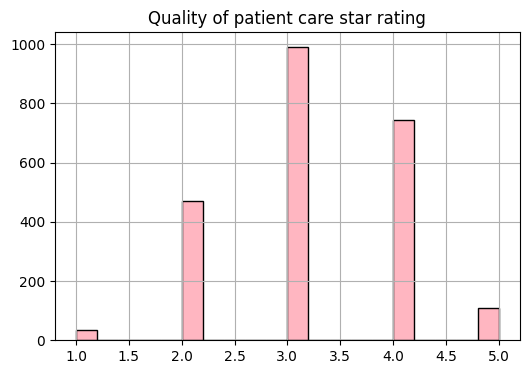
Setul de date conține informații referitoare la agenția care oferă servicii de asistență la domiciliu iar variabila target este calitatea îngrijirii pacientului (‚*Quality of patient care star rating’*). Acest atribut este definit pe o scară de la 1 la 5, în incremente de 0,5, unde 5 definește o satisfacție totală a pacientului în ceea ce privește sistemul de îngrijire la domiciliu.

# Preprocesarea datelor

Trebuie să identificăm și să selectăm cele mai influente atribute dintr-un set de date referitoare la serviciile de sănătate la domiciliu. Procesul de analiză include preprocesarea datelor pentru eliminarea duplicatelor și a informațiilor irelevante, analiza exploratorie pentru a înțelege distribuția și statisticele principale ale atributelor și, în final, utilizarea analizei de componente principale (PCA) pentru selecția celor mai semnificative atribute.

Într-o primă etapă, am efectuat o serie de operațiuni de preprocesare a datelor pentru a le pregăti pentru analiză. Aceasta a inclus eliminarea duplicatelor, eliminarea atributelor cu o cantitate semnificativă de date lipsă și transformarea atributelor cu valori categorice în format numeric. Am aplicat, de asemenea, o tehnică de eliminare a outlier-ilor bazată pe interquartile range (IQR) pentru a asigura consistența datelor.

Următorul pas a constat în analiza exploratorie a setului de date preprocesat. Am calculat valorile medii și mediane pentru a obține o perspectivă asupra tendințelor centrale ale atributelor numerice. De asemenea, am realizat histograme pentru fiecare atribut numeric pentru a vizualiza distribuția datelor și pentru a identifica eventuale pattern-uri sau deviații. Putem observa distribuția oarecum simetrică a atributului-target:

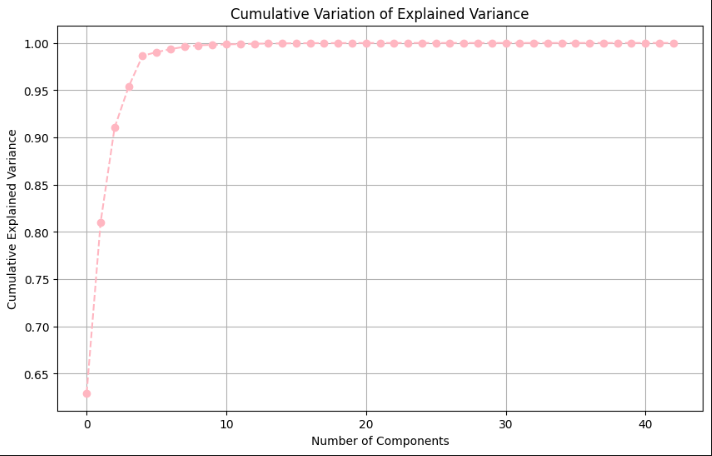


Pentru a identifica cele mai influente atribute, am aplicat analiza de componente principale (PCA), care reprezintă o tehnică utilizată în analiza datelor și în învățarea automată pentru reducerea dimensionalității seturilor de date, păstrând totuși cât mai multe informații relevante posibile. Scopul principal al PCA este să identifice combinații liniare de variabile (componente principale) care explică cea mai mare parte a variației în date.

PCA identifică direcțiile (componentele principale) în care variația datelor este maximă. Aceste componente sunt un set de vectori proprii ale matricei de covarianță a datelor. După care, aceste componente sunt ordonate în funcție de valoarea propriilor valori corespunzătoare. Cu cât o valoare proprie este mai mare, cu atât componenta principală asociată explică mai multă variație în date. Selecționând primele *k* componente principale, unde *k* este un număr mai mic decât dimensiunea inițială a setului de date, se obține o reprezentare mai compactă a datelor. Prin analiza valorilor proprii și a vectorilor proprii, PCA ajută la identificarea și selecția atributelor cu cea mai mare contribuție la variația datelor, îmbunătățind performanța modelelor predictive, eliminând redundanța și zgomotul din setul de date.

După aplicarea analizei de componente principale (PCA), am evaluat *Cumulative Variation of Explained Variance* pentru a determina numărul optim de componente principale de reținut în setul nostru de date. Această metrică este crucială pentru a alege câtă informație dorim să reținem din setul de date inițial și pentru a identifica atributele cele mai semnificative, anume am identificat numărul de componente principale care explică o proporție semnificativă a varianței inițiale, în acest caz, 95%:

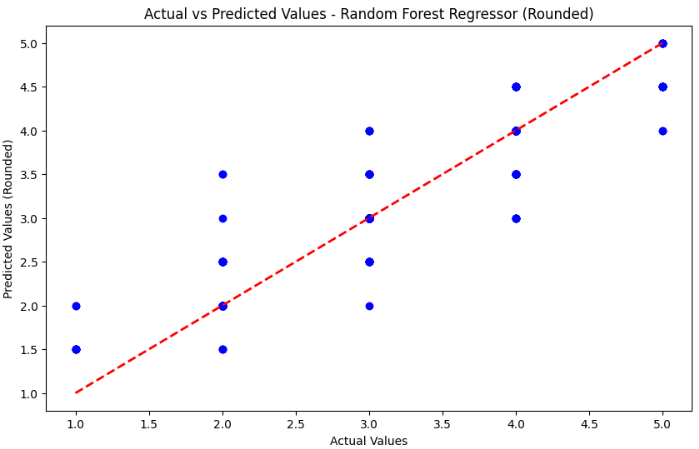
Această tehnică de reducere a dimensionalității ne-a permis să selectăm un subset de atribute care capturează cea mai mare parte a variației în date. Am determinat numărul optim de componente pe baza variației cumulate și am reantrenat modelul PCA cu acest număr optim. Atributele selectate au fost apoi afișate, oferind o înțelegere a contribuției fiecărui atribut la variația totală. În imaginile următoare se poate observa diferența între setul de date inițial și cel prelucrat, fiind rămase doar 43 atribute din cele 70.



# Metode de predicție

## Random Forest Regressor

Modelul Random Forest Regressor este o tehnică de regresie care se bazează pe ideea de a construi un ansamblu de arbori de decizie și de a face o medie a predicțiilor lor. În implementare, setul de date a fost împărțit într-un set de antrenare și un set de testare, iar modelul a fost antrenat pe datele de antrenare utilizând 100 de arbori de decizie. Datele au fost rotunjite pentru a îmbunătăți interpretarea rezultatelor. Metrici precum Mean Absolute Error (MAE) și R2 Score au fost utilizate pentru evaluarea performanței.

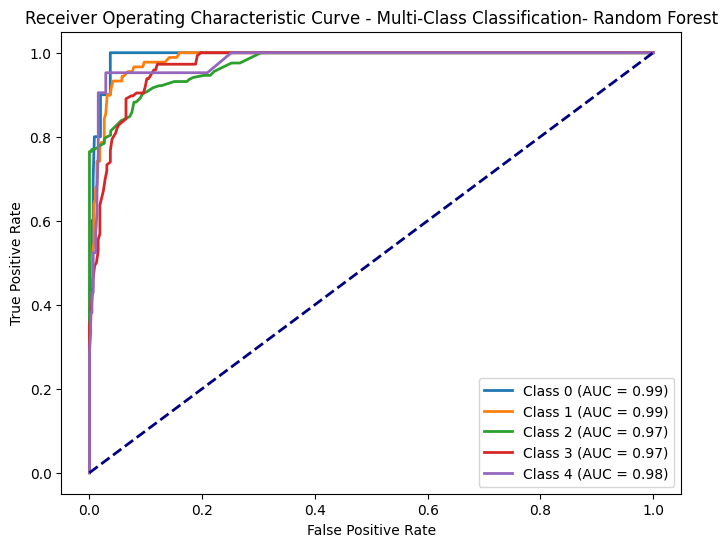


## Random Forest Classifier & Multiclass Classifier

Modelul Random Forest Classifier este o variantă de clasificator ce se bazează pe același principiu al ansamblului de arbori de decizie. În acest caz, obiectivul a fost clasificarea calității îngrijirilor medicale în două categorii: sub sau peste 3.5 stele. Curba ROC a fost folosită pentru evaluarea performanței și s-au prezentat de asemenea acuratețea, matricea de confuzie și raportul de clasificare.

Curba ROC (*Receiver Operating Characteristic*) este o representare grafică a performanțelor unui model de clasificare binară în funcție de pragul de decizie utilizat. Este utilizată pentru a evalua capacitatea unui model de a distinge între două clase, de exemplu, între clasa pozitivă și cea negativă și se construiește pe baza ratelor de adevărat pozitiv (*True Positive Rate, TPR*) și a ratei de fals pozitiv (*False Positive Rate, FPR*) la diferite valori ale pragului de decizie. Aceste rate sunt calculate astfel:

Prin modificarea pragului de decizie, se obțin diferite puncte de pe curba ROC, iar aceasta ilustrează cum se schimbă compromisul între TPR și FPR. Ideal, vrem să maximizăm TPR și să minimizăm FPR. Modelul Random Forest Classifier pentru clasificare multiclase este o extensie a versiunii binare pentru a trata mai multe clase de calitate a îngrijirilor medicale.



## Support Vector Machine Classifier & Multiclass Classifier

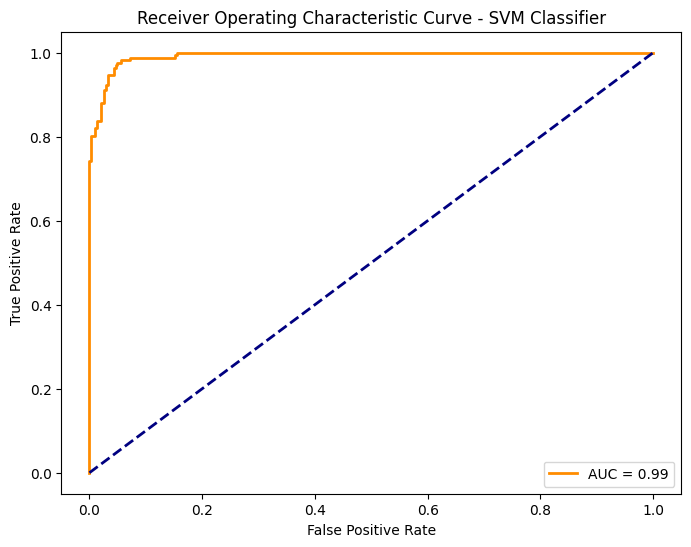
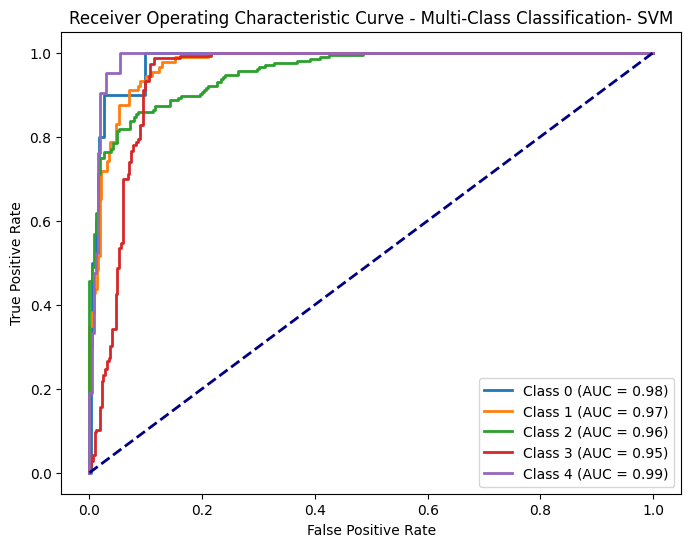
Modelul SVM Classifier este o tehnică de clasificare care se bazează pe găsirea unui hyperplan optim pentru separarea datelor. În implementare, datele au fost standardizate folosind StandardScaler.

În implementarea dată, se utilizează SVM cu probabilitate activată, adăugând un aspect probabilistic la clasificare. Acest lucru permite modelului să ofere nu doar o predicție binară, ci și o măsură a încrederii în această predicție.

Pentru a evalua performanța modelului SVM în clasificarea binară, s-au utilizat diverse metrici, precum acuratețea și matricea de confuzie. Aceste metrici furnizează o înțelegere detaliată a capacității modelului de a face predicții corecte și de a gestiona clasele pozitive și negative.

Pentru a extinde SVM la problemele de clasificare multiclase, s-a utilizat abordarea *One-vs-Rest* (OvR), unde fiecare clasă este tratată ca o problemă binară separată. Acest lucru înseamnă că sunt antrenate multiple clasificatoare binare, iar clasa cu cea mai mare scor este aleasă ca predicție finală.

Evaluarea performanței SVM în clasificarea multiclase implică adesea utilizarea AUC (area de sub curba ROC) pentru fiecare clasă în parte. Aceasta oferă o măsură a abilității modelului de a face distincții clare între clasele diferite.

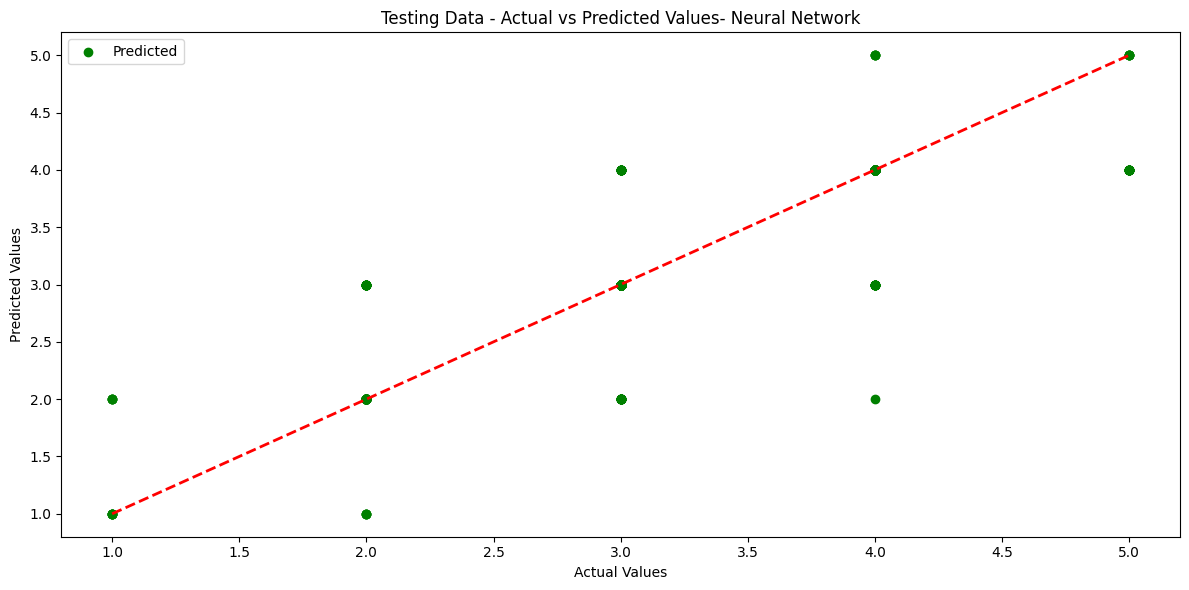
 

## Rețeaua Neuronală

Rețelele neuronale reprezintă o paradigmă de învățare automată inspirată de modul în care funcționează creierul uman, încercând să simuleze procesele neuronale. Rețeaua neurală implementată este un model de regresie, utilizat pentru a prezice valorile continue asociate calității îngrijirii pacienților. Procesul de antrenare al rețelei neurale implică optimizarea parametrilor modelului pentru a minimiza eroarea dintre predicțiile modelului și valorile reale. Acest proces este realizat prin metoda de propagare înapoi (backpropagation), unde gradientul funcției de pierdere este calculat și folosit pentru a actualiza ponderile rețelei.

Pentru a evalua performanța rețelei neurale, diverse metrici sunt utilizate, cum ar fi eroarea medie absolută (MAE), care măsoară diferența absolută între predicțiile modelului și valorile reale. Aceasta oferă o evaluare a preciziei predicțiilor modelului în ceea ce privește calitatea îngrijirii pacienților. De asemenea, în timpul antrenării, se monitorizează acuratețea modelului pe setul de date de antrenare și testare pentru a evalua capacitatea acestuia de a generaliza la date noi.

Un avantaj major al utilizării rețelelor neuronale în contextul calității îngrijirii pacienților constă în capacitatea lor de a învăța reprezentări complexe din datele brute, fără a necesita o preprocesare extensivă. Acestea pot detecta și exploata modele subtile și intricate în datele medicale, contribuind astfel la o înțelegere mai profundă a factorilor care influențează calitatea îngrijirii pacienților.

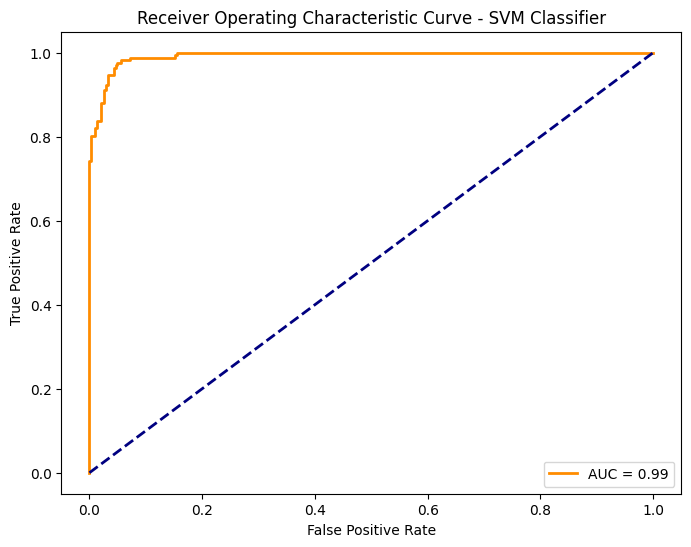
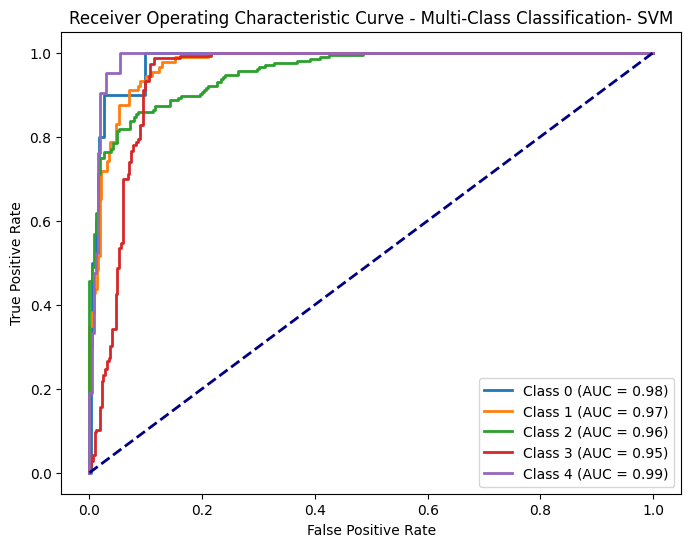


## Decision Tree Regressor

Arborii de decizie sunt algoritmi puternici, folosiți atât în probleme de regresie, cât și în cele de clasificare. Pentru *Decision Tree Regressor*, un arbore de decizie este antrenat pentru a realiza o predicție numerică (regresie) a calității îngrijirii pacienților (atributul "*Quality of patient care star rating*"). Modelul se antrenează pe setul de date și apoi își face predicțiile pe setul de testare. Este evaluată performanța modelului prin calculul Mean Squared Error (MSE), care măsoară dispersia dintre valorile prezise și cele reale.

## Decision Tree Classifier & Multiclass Classifier

În funcția *train\_decision\_tree\_classifier*, un arbore de decizie este antrenat pentru a realiza o clasificare binară a calității îngrijirii pacienților (în funcție de pragul 3.5). Modelul este evaluat în funcție de acuratețe (95%) și este afișată o matrice de confuzie, oferind o perspectivă asupra calității clasificării. Apoi, în *train\_decision\_tree\_classifier\_multiclass*, este implementat un arbore de decizie pentru a realiza o clasificare multi-clasă a calității îngrijirii pacienților, pentru fiecare atribut important. AUC (Area Under the Curve) este calculat pentru fiecare clasă în parte, iar performanța este afișată sub forma unei curbe ROC (Receiver Operating Characteristic). Acuratețea scade ușor, la 93%.

# Concluzii

Analizând performanța modelelor de clasificare (*Random Forest*, *SVM*, *Decision* *Tree* și *Neural* *Network*), putem observa că *Random* *Forest* și *SVM* au obținut rezultate remarcabile, cu o acuratețe de aproximativ 95% și AUC de aproximativ 99%. Aceste modele au fost foarte bune în clasificarea în două clase, obținând rezultate echilibrate în ceea ce privește precizia (precision), amplitudinea acoperirii (recall) și măsura F1-score, în special pentru clasa 0 și clasa 1. Scorul F1 este o măsură a performanței unui model de clasificare și este calculat pe baza valorilor de precizie și amplitudine.

*Random Forest Classifier* și *SVM Classifier* au prezentat o acuratețe comparabilă, cu performanțe ridicate în identificarea ambelor clase. Pentru ambele modele, precizia și amplitudinea acoperirii au fost ridicate, ceea ce indică capacitatea lor de a face predicții corecte și de a recupera majoritatea exemplelor pozitive.

*Decision Tree Classifier* a avut o performanță ușor inferioară în comparație cu Random Forest și SVM, cu o acuratețe de aproximativ 93% și o AUC de 92%. Acest model a obținut rezultate bune în identificarea clasei 0, dar s-a confruntat cu o scădere a preciziei și amplitudinii acoperirii pentru clasa 1. Performanțele relativ echilibrate ale acestui model indică capacitatea sa de a face predicții bune pentru ambele clase*. Neural Network Classifier* a prezentat performanțe inferioare comparativ cu celelalte modele, cu o acuratețe de aproximativ 89,8% și AUC de 97%.

În concluzie, analiza modelelor de învățare automată pentru evaluarea calității îngrijirii pacienților în domeniul sănătății a relevat performanțe diferite pentru diversele abordări utilizate. Setul de date a fost supus unui proces riguros de preprocesare, care a inclus eliminarea duplicatelor, gestionarea datelor lipsă, transformarea atributelor și eliminarea outlier-ilor, pregătind astfel datele pentru analiza ulterioară. Putem conclude că alegerea unui model depinde de contextul specific și de necesitățile problemei în cauză. Cu toate că *Random Forest* și *SVM* au prezentat rezultate excelente, este important să se evalueze atent avantajele și dezavantajele fiecărui model, luând în considerare specificul domeniului medical și importanța unor caracteristici precum interpretabilitatea, înainte de a face o alegere finală pentru implementare.

