Tema 2 IA

Primul set de date pentru Diabet:

Informații Generale

Informațiile de bază despre dataset-uri:

Test Dataset: 2000 de înregistrări și 25 de coloane.

Train Dataset: 8000 de înregistrări și 25 de coloane.

Full Dataset: 10000 de înregistrări și 25 de coloane.

Histogram for Diabetes

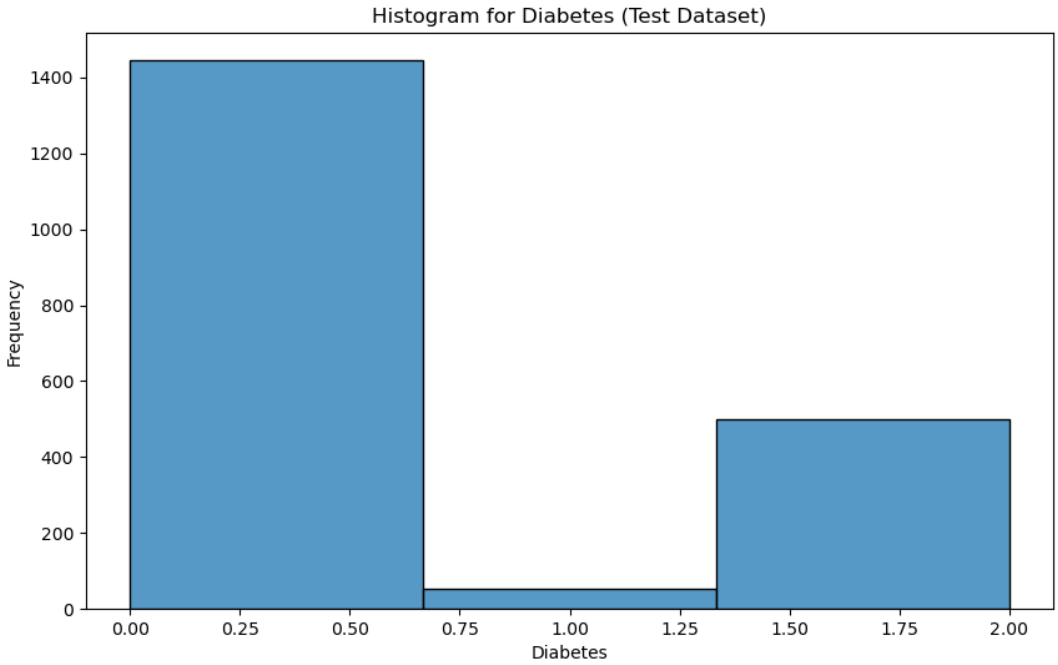
Un histogramă arată distribuția frecvenței valorilor unei variabile. În cazul nostru, graficul histogramă pentru coloana 'Diabetes' din fiecare dataset (test, train, full) arată numărul de cazuri pentru fiecare valoare a diabetului (0 sau 1).

Diabetes Histogram:

Valorile 0 reprezintă indivizi fără diabet.

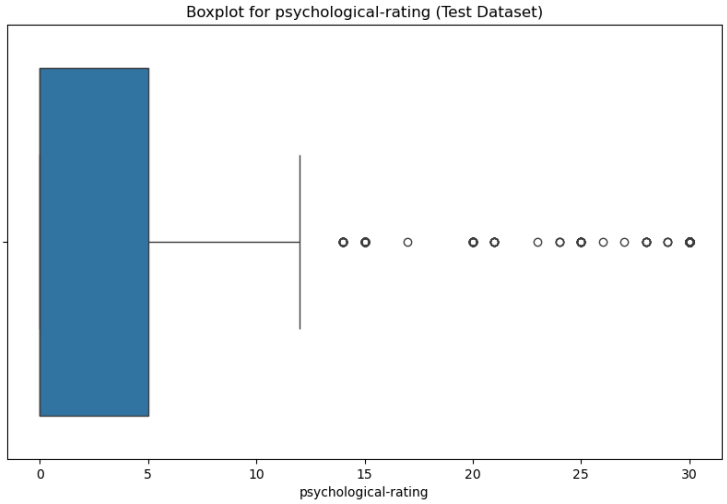
Valorile 1 reprezintă indivizi cu diabet.

Dacă există o valoare mare de 0 comparativ cu 1, aceasta indică un dezechilibru al clasei în datele noastre.



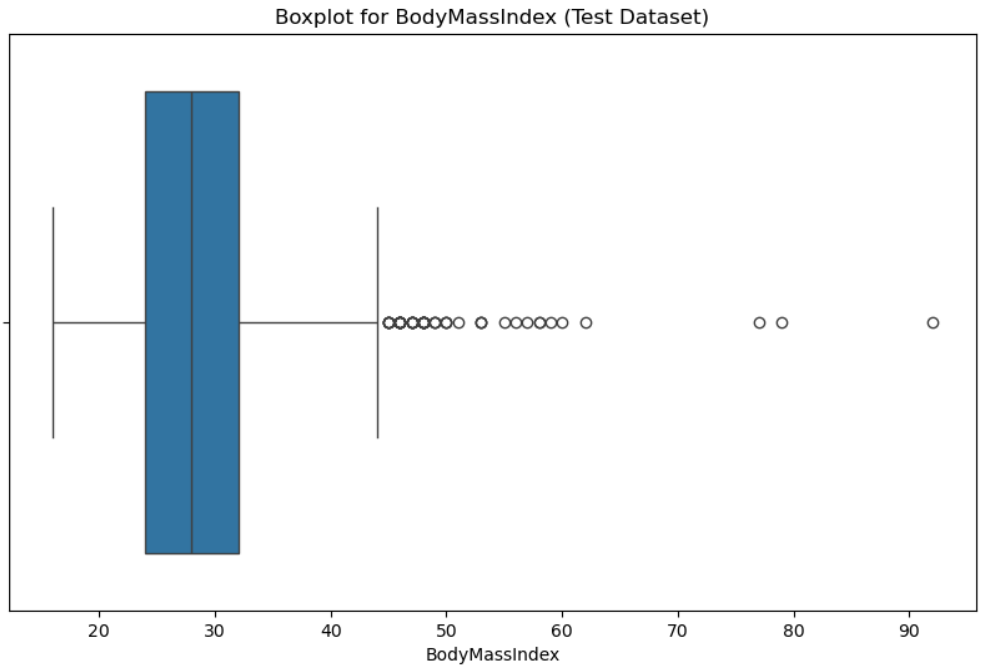
1. Histogram for Diabetes (Test Dataset) - diabetes_histogram_Test.png:

Aceasta histogramă arată distribuția frecvenței variabilei 'Diabetes' în setul de date de testare. Ajută la înțelegerea echilibrului clasei variabilei țintă, ceea ce este crucial pentru construirea de modele robuste de învățare automată. Un set de date echilibrat asigură că modelul nu favorizează o clasă în detrimentul alteia.



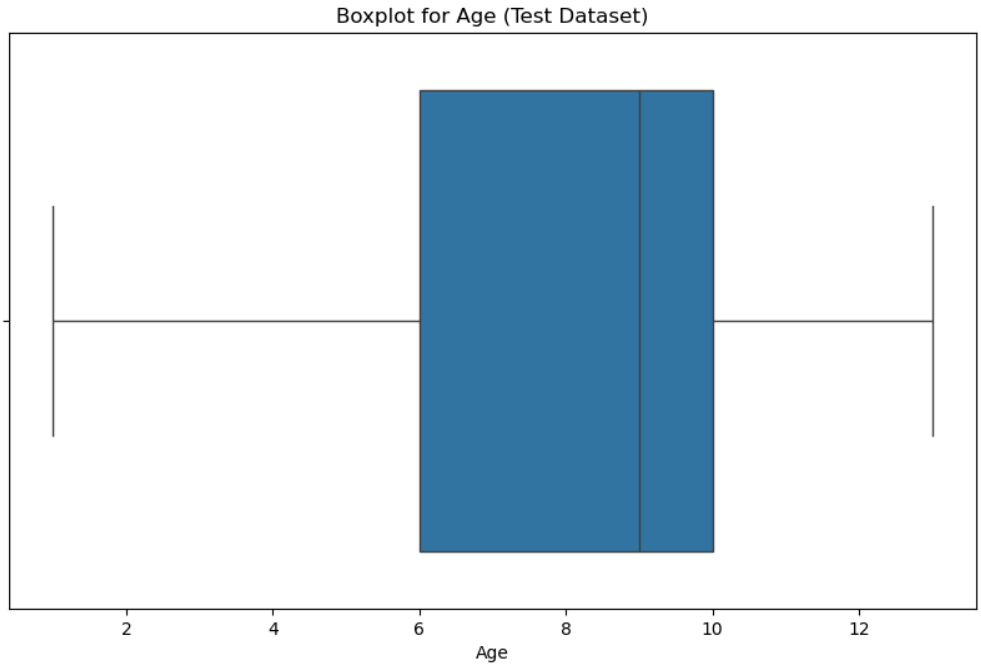
2. **Boxplot for psychological-rating (Test Dataset) - boxplot_psychological-rating_Test.png:**

Descriere: Acest boxplot vizualizează distribuția valorilor pentru atributul numeric continuu 'psychological-rating' în setul de date de testare. Boxplot-ul arată o sumă de cinci valori: minim, primul quartil, mediană, al treilea quartil și maxim. Este util pentru detectarea valorilor anormale și înțelegerea răspândirii și asimetriei datelor.



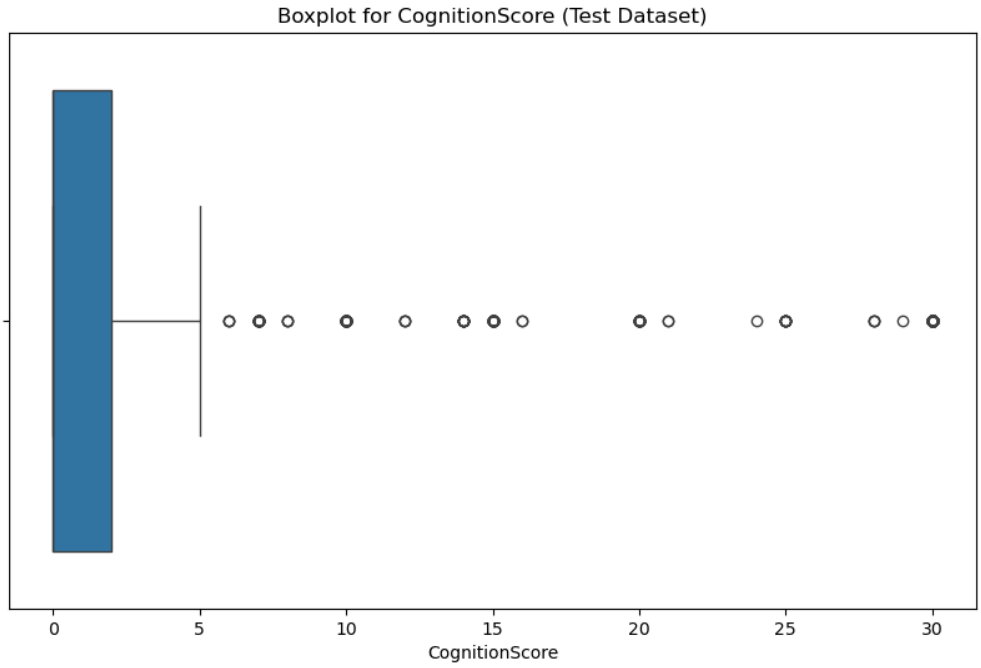
3. **Boxplot for BodyMassIndex (Test Dataset) - boxplot_BodyMassIndex_Test.png:**

Descriere: Acest boxplot arată distribuția valorilor pentru atributul numeric continuu 'BodyMassIndex' în setul de date de testare. Este utilizat pentru a identifica valori anormale și pentru a înțelege răspândirea și distribuția datelor.



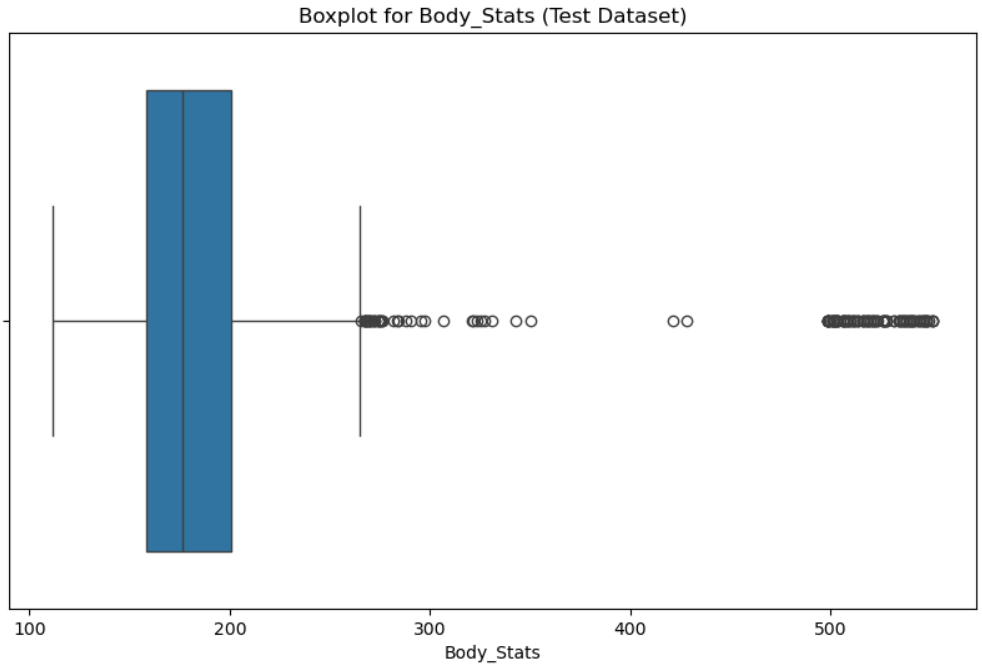
4. Boxplot for Age (Test Dataset) - boxplot_Age_Test.png:

Descriere: Acest boxplot arată distribuția valorilor pentru atributul numeric continuu 'Age' în setul de date de testare. Ajută la detectarea valorilor anormale și la înțelegerea răspândirii și distribuției vârstelor în setul de date



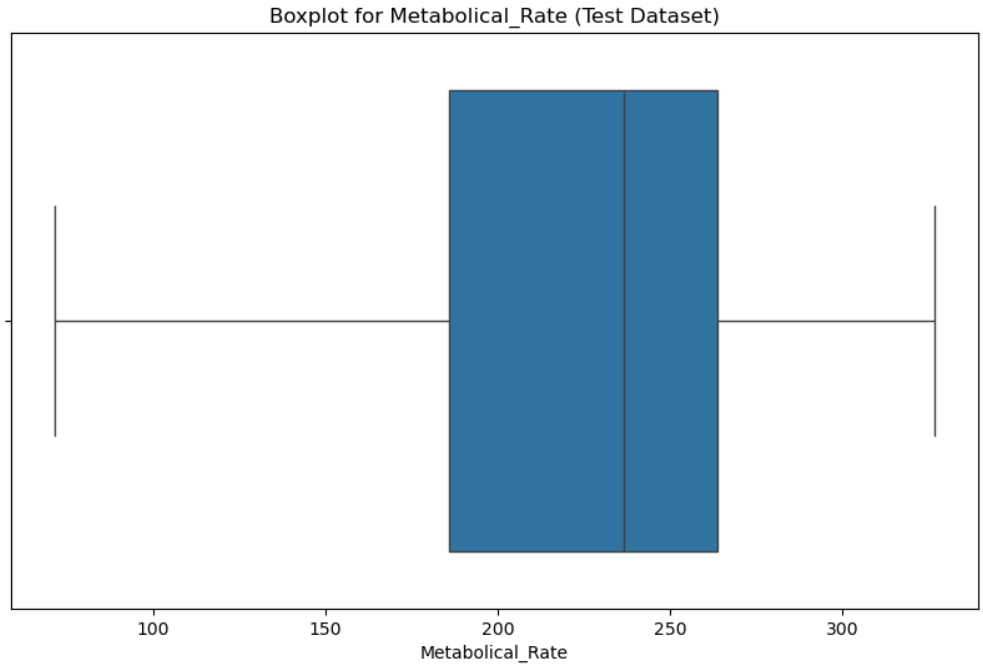
5. Boxplot for CognitionScore (Test Dataset) - boxplot_CognitionScore_Test.png:

Descriere: Acest boxplot arată distribuția valorilor pentru atributul numeric continuu 'CognitionScore' în setul de date de testare. Este util pentru a înțelege răspândirea și variația scorurilor de cogniție și pentru a detecta valorile anormale.



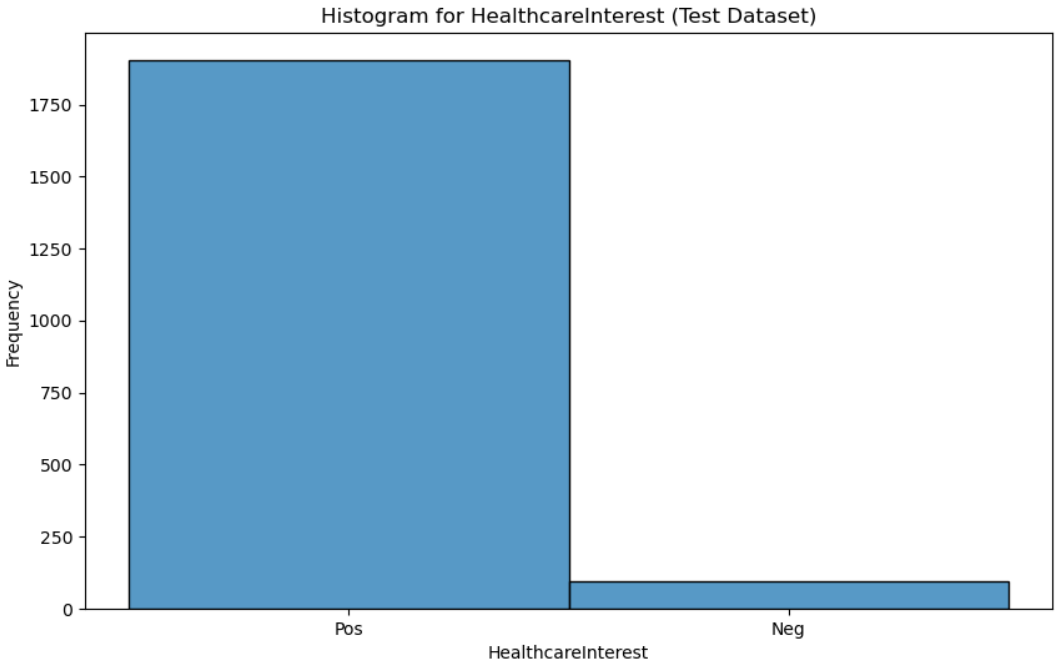
1.
6. Boxplot for Body_Stats (Test Dataset) - boxplot_Body_Stats_Test.png:

Descriere: Acest boxplot arată distribuția valorilor pentru atributul numeric continuu 'Body_Stats' în setul de date de testare. Ajută la identificarea valorilor anormale și la înțelegerea răspândirii datelor corporale.



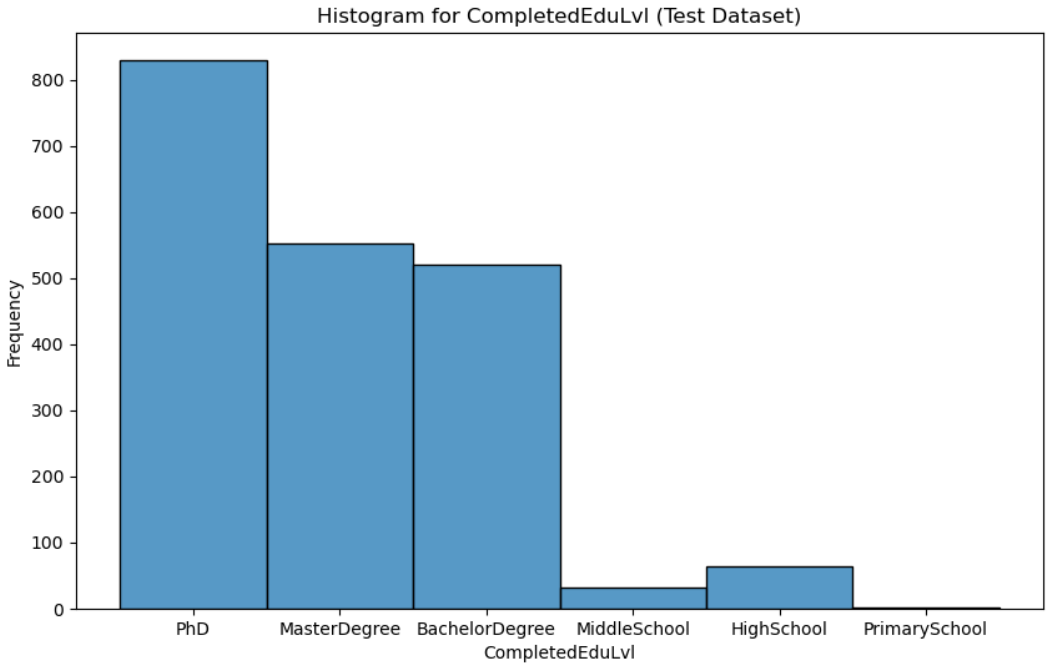
7. Boxplot for Metabolical_Rate (Test Dataset) - boxplot_Metabolical_Rate_Test.png:

Descriere: Acest boxplot arată distribuția valorilor pentru atributul numeric continuu 'Metabolical_Rate' în setul de date de testare. Este util pentru a înțelege variația ratei metabolice și pentru a detecta valorile anormale.



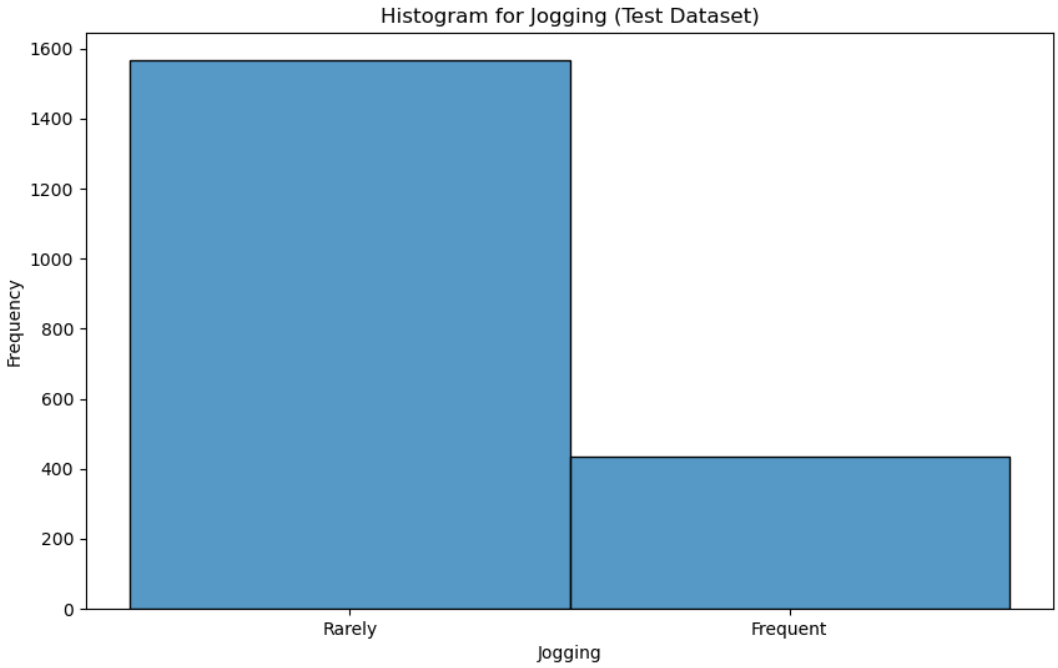
1.
8. Histogram for HealthcareInterest (Test Dataset) - histogram_HealthcareInterest_Test.png:

Descriere: Această histogramă arată distribuția frecvenței valorilor pentru atributul categoric 'HealthcareInterest' în setul de date de testare. Ajută la înțelegerea distribuției și frecvenței fiecărei categorii.



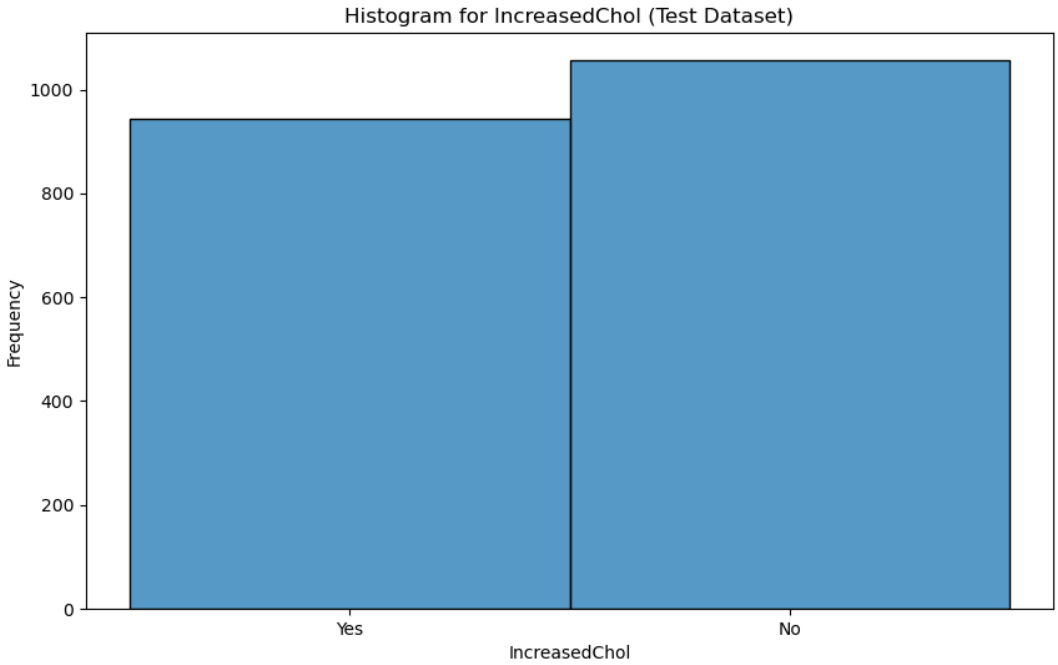
9. Histogram for CompletedEduLvl (Test Dataset) - histogram_CompletedEduLvl_Test.png:

Descriere: Această histogramă arată distribuția frecvenței valorilor pentru atributul ordinal 'CompletedEduLvl' în setul de date de testare. Este utilă pentru a înțelege distribuția nivelurilor de educație completate.



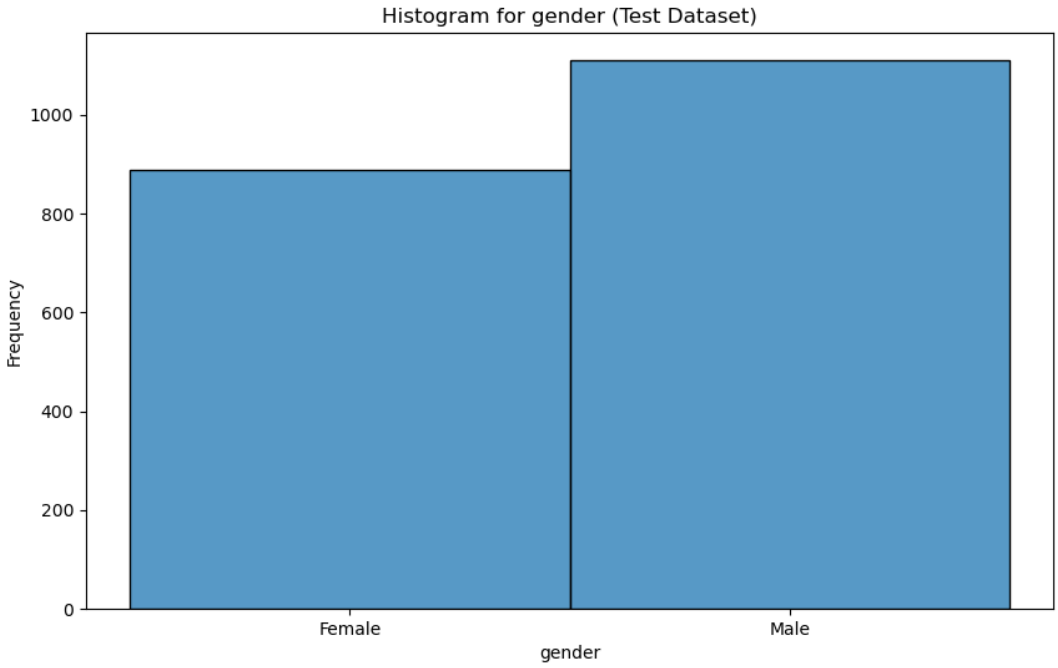
10. Histogram for Jogging (Test Dataset) - histogram_Jogging_Test.png:

Descriere: Această histogramă arată distribuția frecvenței valorilor pentru atributul categoric 'Jogging' în setul de date de testare. Ajută la înțelegerea distribuției și frecvenței fiecărei categorii.



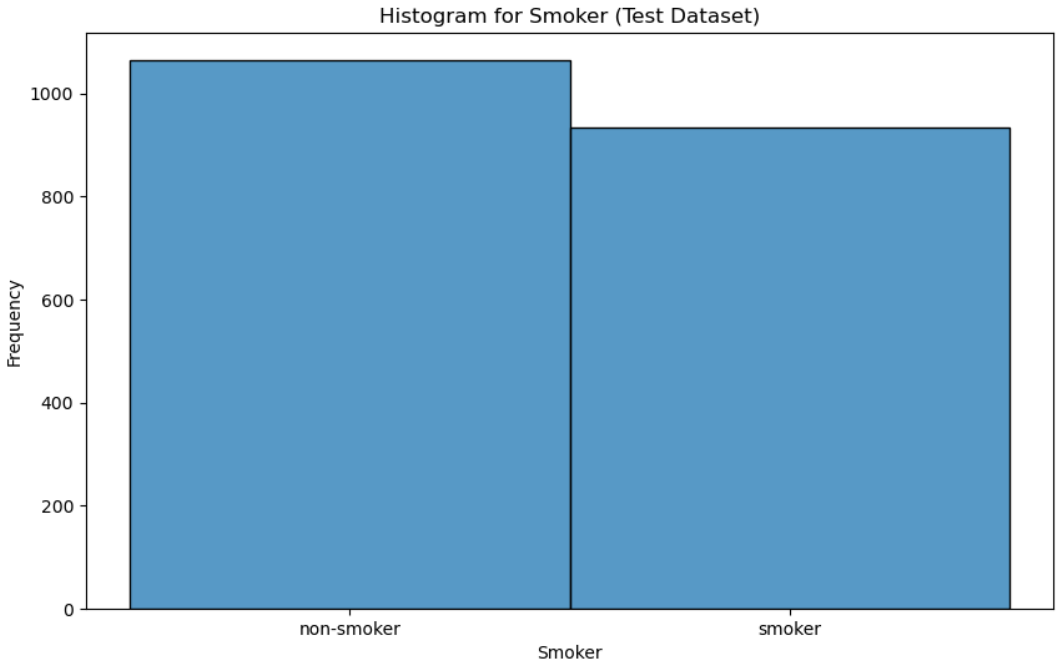
11. Histogram for IncreasedChol (Test Dataset) - histogram_IncreasedChol_Test.png:

Descriere: Această histogramă arată distribuția frecvenței valorilor pentru atributul categoric 'IncreasedChol' în setul de date de testare. Este utilă pentru a înțelege distribuția și frecvența fiecărei categorii.



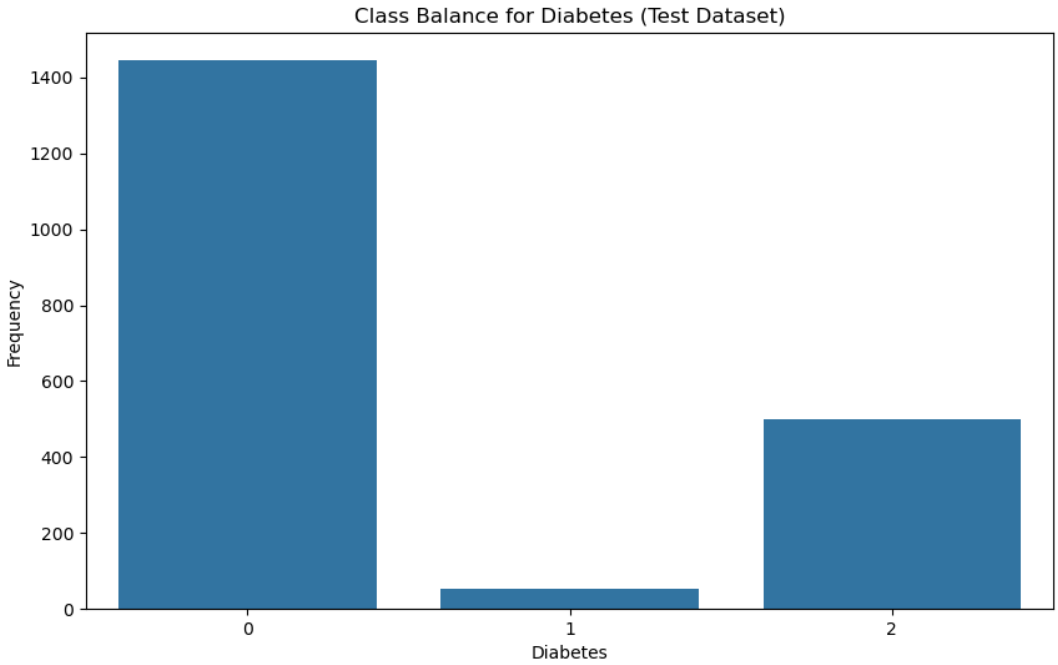
12. Histogram for gender (Test Dataset) - histogram_gender_Test.png:

Descriere: Această histogramă arată distribuția frecvenței valorilor pentru atributul categoric 'gender' în setul de date de testare. Ajută la înțelegerea distribuției și frecvenței fiecărei categorii.



13. Histogram for Smoker (Test Dataset) - histogram_Smoker_Test.png:

Descriere: Această histogramă arată distribuția frecvenței valorilor pentru atributul categoric 'Smoker' în setul de date de testare. Este utilă pentru a înțelege distribuția și frecvența fiecărei categorii.



14. Class Balance for Diabetes (Test Dataset) - class_balance_Diabetes_Test.png:

Descriere: Acest countplot arată frecvența fiecărei clase în variabila țintă 'Diabetes' în setul de date de testare. Ajută la înțelegerea dacă setul de date este echilibrat sau dezechilibrat. Seturile de date dezechilibrate pot necesita tehnici suplimentare de reechilibrare pentru a asigura performanța corectă a modelului.

CREDIT RISK

Histogram for Loan Approval Status:

Descriere: Această histogramă arată distribuția frecvenței variabilei 'loan_approval_status' în setul de date. Ajută la înțelegerea echilibrului clasei variabilei țintă. Un set de date echilibrat este crucial pentru construirea de modele robuste de învățare automată.

Descriptive Statistics for Continuous Numeric Attributes:

Descriere: Acest tabel oferă un rezumat al tendinței centrale, dispersiei și formei distribuției setului de date pentru atributele numerice continue. Statisticile cheie includ media, deviația standard, minimul, maximul și valorile quartilelor. Aceste statistici oferă o imagine de ansamblu asupra caracteristicilor datelor, cum ar fi valorile medii și variabilitatea.

Boxplot for Continuous Numeric Attributes:

Descriere: Boxplot-urile vizualizează distribuția datelor pe baza unui rezumat de cinci numere: minim, primul quartil, mediană, al treilea quartil și maxim. Sunt utile pentru detectarea valorilor anormale și înțelegerea răspândirii și asimetriei datelor. Fiecare boxplot reprezintă distribuția unui singur atribut numeric continuu.

Histograms for Categorical and Ordinal Attributes:

Descriere: Aceste histogramă arată distribuția frecvenței variabilelor categorice/ordinale în setul de date. Ajută la înțelegerea distribuției și frecvenței fiecărei categorii, esențială pentru ingineria caracteristicilor și construirea modelului.

Class Balance for Loan Approval Status:

Descriere: Acest countplot arată frecvența fiecărei clase în variabila țintă 'loan_approval_status'. Ajută la înțelegerea dacă setul de date este echilibrat sau dezechilibrat. Seturile de date dezechilibrate pot necesita tehnici suplimentare de reechilibrare pentru a asigura performanța corectă a modelului.

Correlation Matrix for Numeric Attributes:

Descriere: Această matrice de corelație arată coeficienții de corelație între perechile de atribute numerice. Ajută la înțelegerea forței și direcției relației dintre variabile. Corelațiile ridicate între atribute pot indica redundanță, ceea ce înseamnă că una dintre variabile poate fi eliminată pentru a reduce complexitatea modelului.

Categorical Correlation Analysis:

Descriere: Rezultatele testului Chi-Square indică dacă există o corelație semnificativă între perechile de atribute categorice. O valoare p scăzută (de obicei < 0,05) sugerează o relație semnificativă între variabile. Această analiză ajută la identificarea perechilor de variabile categorice care