Anton Stefan, 331CC

`Tema2 - IA`

ML aplicat

Durata implementare: ~26 ore

Detalii implementare: Cerinta de MLP pornind de la codul de laborator nu a fost implementata. In rest toate cerintele temei au fost acoperite.

Workflow: Analiza și vizualizarea datelor constituie prima etapă a fluxului de lucru, începând cu încărcarea seturilor de date pentru riscul de credit și diabet din fișiere CSV, utilizând biblioteca pandas. Aceasta facilitarea manipulării și analizei eficiente a datelor.

Următorul pas implică analiza atributelor, unde identific tipurile de atribute (numerice, categoriale și ordinale) din fiecare set de date. Acest proces este fundamental în determinarea metodelor statistice și a tehnicilor de vizualizare adecvate pentru fiecare tip de date.

Continui cu sumarizarea statistică a atributelor numerice, calculând statistici descriptive precum media, mediana, quartilele și numărul de valori lipsă. Aceasta ajută la înțelegerea structurii datelor, a tendinței centrale și a dispersiei, utilizând funcțiile oferite de pandas.

În ceea ce privește vizualizările:

Utilizez boxplot-uri pentru a detecta outlierii în atributele numerice și pentru a observa distribuția datelor. Generez histograme pentru a ilustra distribuția frecvențelor atributele categoriale, facilitând o înțelegere vizuală a distribuției fiecărei categorii. Procesul de prelucrare a datelor include:

Imputarea, unde gestionez valorile lipsă înlocuindu-le cu media pentru atributele numerice și cu modul pentru cele categoriale, asigurându-mă că setul de date este complet pentru modelare. Tratarea outlierilor, ajustând valorile extreme pentru a corespunde cu un interval acceptabil, calculat pe baza intervalului interquartil (IQR). Scalarea caracteristicilor, aplicând standardizare pentru atributele numerice pentru a asigura că modelul va trata echitabil variabilele, independent de scara lor originală. Implementarea și evaluarea modelului include două abordări principale:

Pentru Random Forest, efectuez o codificare one-hot pentru variabilele categoriale și ajustez parametrii modelului, cum ar fi numărul de arbori și adâncimea maximă, pentru a optimiza performanța, inclusiv gestionarea echilibrată a claselor dezechilibrate. Pentru MLP (Perceptron Multi-Strat), configurez modelul cu straturi ascunse și parametri specifici, cum ar fi rata de învățare și oprirea timpurie, pentru a preveni supra-antrenarea și pentru a îmbunătăți convergența.

Cerinta 3.1 grafice Analiza

CREDIT_RISK

Matricea de Correlație Am observat în matricea de corelație că rata dobânzii și suma împrumutată sunt moderat corelate pozitiv, ceea ce indică faptul că împrumuturile mai mari tind să aibă rate ale dobânzii mai mari. Istoricul de credit influențează și el rata dobânzii și suma împrumutată, reflectând utilizarea istoricului de credit ca un indicator cheie de risc pentru bănci.

Distribuția Statutului Rezidențial Histograma statutului rezidențial arată că majoritatea solicitantilor sunt chiriași sau proprietari, cu foarte puține cazuri neidentificate. Aceasta varietate subliniază diversitatea condițiilor de locuit ale aplicanților.

Scopul Împrumutului Împrumuturile sunt cel mai frecvent solicitate pentru studii, sănătate și afaceri. Acest lucru sugerează că solicitările de împrumut sunt predominant motivate de necesități personale și investiții în dezvoltarea personală sau profesională.

Distribuția Ratingului Împrumutului Majoritatea împrumuturilor sunt clasificate ca având un rating 'excelent' sau 'foarte bun'. Acest lucru indică faptul că majoritatea solicitantilor au un istoric de credit solid, ceea ce este un semn bun pentru creditori.

Istoricul Creditului Am constatat că majoritatea solicitantilor nu au avut probleme anterioare cu creditul, conform distribuției stării istoricului de credit. Acesta este un indicator pozitiv al sănătății financiare a solicitantilor.

Distribuția Stabilității Distribuția ratingului de stabilitate arată că majoritatea solicitantilor sunt considerați stabili financiar, cu o minoritate semnificativă având ratinguri scăzute de stabilitate, aspect ce poate afecta deciziile de acordare a împrumuturilor.

Aprobarea Împrumuturilor Observăm că un număr considerabil de împrumuturi sunt aprobate față de cele respinse, reflectând o tendință a băncii de a aproba împrumuturi pentru solicitanti cu profiluri de risc scăzut, asa cum reiese din ratingurile de credit predominant pozitive.

Boxplot-uri Boxplot-urile pentru variabilele numerice, cum ar fi vârsta solicitantului, venitul, istoricul de credit și suma împrumutată, arată o dispersie variabilă și prezența unor valori extreme. Spre exemplu, venitul solicitantului și suma împrumutată prezintă o varietate largă și multe valori extreme, aspecte care necesită investigații suplimentare pentru a asigura acuratețea datelor și evaluarea adecvată a riscurilor asociate cu aceste împrumuturi.

DIABET

Matricea de Corelație Matricea de corelație arată legături între diferite variabile de sănătate, cum ar fi indicele de masă corporală și obiceiurile alimentare, precum și între exercițiile fizice și parametrii cardiovasculari.

Distribuția Claselor Majoritatea subiecților sunt non-diabetici, cu un număr semnificativ mai mic de cazuri de diabet tip 1 și tip 2.

Histograme pentru Diverse Variabile Interesul pentru îngrijirea sănătății: Majoritatea subiecților arată un interes mare pentru sănătate. Nivelul de educație: Predomină cei cu educație universitară. Jogging: Majoritatea participanților practică jogging rar. Colesterol crescut: O proporție moderată de persoane au colesterol crescut. Genul: Distribuția este echilibrată între bărbați și femei. Fumători: Mai puțin de jumătate dintre subiecți sunt fumători.

Boxplot-uri pentru Variabile Numerice Boxplot-urile arată variații semnificative în metricile clinice, cum ar fi tensiunea arterială și nivelurile de colesterol, evidențiind prezența outlierilor care necesită atenție suplimentară.

Cerinta 3.3 - PREPROCESARE | ALGORITMI Random Forest, MLP

CREDIT RISK - PREPROCESAT

Matricea de Corelație:

Corelațiile între variabile sunt mai clare și mai bine definite, indicând o normalizare și scalare eficientă a datelor. Distribuția Statutului Rezidențial:

Datele sunt complete, fără valori lipsă, și arată o distribuție clară a statusului rezidențial. Scopul Împrumutului:

Distribuțiile sunt complete și precise, fără valori lipsă, reflectând motivele clare ale împrumuturilor. Distribuția Ratingului Împrumutului:

Toate categoriile sunt bine definite și fără valori lipsă, facilitând o evaluare corectă a ratingurilor. Istoricul Creditului:

Datele sunt complete și reflectă corect istoricul de credit al solicitanților. Distribuția Stabilității:

Distribuția este clară și fără date lipsă, oferind o imagine precisă a stabilității financiare. Aprobarea Împrumuturilor:

Distribuția datelor este completă, permițând o analiză exactă a aprobărilor și respingerilor. Boxplot-uri pentru Variabilele Numerice:

Valorile extreme sunt gestionate mai bine, iar variabilele sunt standardizate, arătând o variație și dispersie controlate.

DIABET - PREPROCESAT

Matricea de Corelație După preprocesare: Matricea de corelație arată mai puține corelații între variabile, deoarece variabilele care aveau multe valori lipsă sau valori extreme au fost ajustate sau eliminate.

Modificări observate:

Corelațiile între variabilele de sănătate au fost ajustate, rezultând o imagine mai clară a relațiilor reale dintre variabile. Distribuția Claselor După preprocesare:

Distribuția claselor este acum echilibrată. Toți subiecții sunt clasificați într-o singură clasă, ceea ce sugerează că valorile lipsă sau incorecte din etichetele de clasă au fost corectate.

Modificări observate:

Clasele 1 și 2 au fost eliminate sau reclasificate, rezultând o singură clasă dominantă (0). Histograme pentru Diverse Variabile După preprocesare: Distribuția variabilelor categorice a fost ajustată pentru a reflecta datele curățate. Valorile lipsă au fost completate, iar variabilele au fost standardizate.

Modificări observate:

Valorile lipsă și valorile extreme au fost gestionate, rezultând o distribuție mai uniformă și mai realistă a variabilelor. Boxplot-uri pentru Variabile Numerice După preprocesare: Boxplot-urile arată o reducere a outlierilor și o distribuție mai uniformă a variabilelor numerice, indicând că valorile extreme au fost ajustate sau eliminate.

Modificări observate:

Variabilitatea extremă și outlierii au fost minimizați, ceea ce indică o curățare eficientă a datelor.

Evaluarea Algoritmilor de Clasificare

Random Forest

Credit Risk: Antrenament: Acuratețe 100%, performanță excelentă pentru toate clasele. Test: Acuratețe 100%, model robust. Impact Dezechilibru: Fără impact negativ datorită ponderării claselor.

Diabet: Antrenament: Acuratețe 56%, confuzie între clase. Test: Acuratețe 42%, performanță slabă pentru clasele minoritare. Impact Dezechilibru: Dezechilibrul a redus semnificativ performanța modelului.

MLP (Multi-Layer Perceptron)

Credit Risk: Antrenament: Acuratețe 99.93%, performanță bună. Test: Acuratețe 78.2%, dificultăți cu clasele negative. Impact Dezechilibru: Performanța afectată pentru clasele minoritare.

Diabet: Antrenament: Acuratețe 100%, doar clasa majoritară prezisă. Test: Acuratețe 72.3%, precizie slabă pentru clasele 1 și 2. Impact Dezechilibru: Performanța modelului afectată semnificativ.

HIPERPARAMETRII

Modelul MLPClassifier pentru credit risk folosește o rețea neurală cu două straturi ascunse (50 și 30 de neuroni), antrenată pentru maximum 1000 de iterații, cu o rată de învățare inițială de 0.1. Parametrul random_state este setat pentru reproducibilitate. 10% din date sunt folosite pentru validare. Antrenamentul se

oprește anticipat dacă nu există îmbunătățiri timp de 10 epoci consecutive, prevenind supraînvățarea. Acești hiperparametri echilibrează complexitatea modelului, timpul de antrenament și performanța pe setul de validare.

Modelul RandomForestClassifier pentru credit risk folosește 100 de arbori de decizie, cu o adâncime maximă de 10 pentru a limita complexitatea. Split-urile se fac doar dacă există cel puțin 10 exemple, iar fiecare frunză trebuie să conțină cel puțin 5 exemple. Se folosește criteriul Gini pentru a măsura calitatea split-urilor. Parametrul class_weight este setat pe 'balanced' pentru a aborda dezechilibrul de clase. Numărul de caracteristici luate în considerare la fiecare split este limitat la rădăcina pătrată a numărului total de caracteristici. Parametrul random_state este setat pentru reproducibilitate. Acești hiperparametri sunt aleși pentru a preveni supraînvățarea și pentru a aborda dezechilibrul de clase.

Credit Risk Model Comparison

Algorithm	Accuracy	Precision_False	Recall_False	F1-Score_False	Precision_True	Recall_True	F1-Score_True
MLPClassifier	0.782	0.0	0.0	0.0	0.78	1.0	0.88
RandomForestClassifier	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Diabetes Model Comparison

Algorithm	Accuracy	Precision_Class_0	Recall_Class_0	F1-Score_Class_0
MLPClassifier	0.723	0.72	1.0	0.84
RandomForestClassifier	0.417	0.73	0.56	0.63

Algorithm	Precision_Class_1	Recall_Class_1	F1-Score_Class_1	Precision_Class_2	Recall_Class_2	F1-Score_Class_2
MLPClassifier	0.0	0.0	0.0	0.0	0.0	0.0
RandomForestClassifier	0.33	0.43	0.05	0.0	0.0	0.0

Pentru setul de date Credit Risk, RandomForestClassifier a avut performanțe perfecte în evaluare, ceea ce poate sugera un posibil overfitting. Pentru setul de date Diabetes, MLPClassifier a avut performanțe mai bune comparativ cu RandomForestClassifier, dar ambele modele au avut dificultăți în a clasifica corect clasele 1 și 2.