PREDICTIA DIABETULUI

-proiect TCRI-

(Tehnici de căutare și regăsire a informației)

ORBISOR BIANCA-ALEXANDRA

SLIDE 2

Obiectivul proiectului a fost crearea unei aplicații de predicție a diabetului pe baza a 5 parametrii, glucoza, tensiunea arterială diastolică, insulina, indicele de masă corporală și vârsta unei persoane, utilizând o tehnică de învățare supervizată.

SLIDE 3

Am ales această temă deoarece, în întreaga lume, conform Organizației Mondiale a Sănătății, aproximativ 422 de milioane de persoane suferă de diabet. Diabetul este o boală care aduce după sine și ale afecțiuni, precum afecțiuni cardiovasculare, oculare, boli renale și lista poate continua. În plus, În situația epidemiologică din zilele noastre s-a ajuns la concluzia că persoanele care au diabet au un risc mai mare de a se infecta cu virusul SARS-CoV-2, fiind predispuși la forme severe ale bolii. Diabetul ar crește riscul de infecție de două până la trei ori, independent de prezența altor afecțiuni.

SLIDE 4

Urmează O scurtă prezentare a metabolismului glucidelor

SLIDE 5

Corpul are nevoie zilnic de energie, pe care o obține prin alimentație. Mai întâi, mâncarea este transformată în glucoză, prin procesul de digestie. Apoi, glucoza ajunge în sânge și este sintetizată, cu ajutorul insulinei produse de pancreas, în energie. Problemele apar atunci când pancreasul nu produce insulina necesară sau când organismul nu folosește eficient acest hormon. În aceste situații, glucoza nu mai este transformată în energie și se acumulează în organism într-o cantitate mai mare decât ar fi normal, favorizând declanșarea diabetului zaharat.

SLIDE 6

Ca set de date utilizat pentru implementarea modelului am ales un set de date de pe Kaggle, în format csv, având 15000 de înregistrări, cu 7 coloane: glucoza, insulina, vârsta,

SLIDE 7

tensiunea arterială diastolică. Aceasta indică câtă presiune exercită sângele asupra pereților arterelor între bătăile inimii, indicele de masa corporală, calculat pe baza greutății și înălțimii persoanei și Diabetic, prin intermediul căruia se exprimă dacă analizele sunt ale unei persoane care are sau nu diabet. 0 -înseamnă că nu are diabet și 1 – că are diabet

SLIDE 8

Preprocesarea datelor a presupus vizualizarea tipului de date, căutarea de valori nule și completarea acestora cu media pe coloană. În urma analizei setului de date nu am găsit valori nule. În ceea ce privește distribuția datelor, din cele 15000 de înregistrări, 10000 de subiecți nu au diabet și 5000 au diabet.

SLIDE 9

Din graficele ilustrate se poate observa distribuția datelor, și limitele în care variază fiecare parametru studiat.

SLIDE 10

Distribuția datelor pentru fiecare parametru în funcție de parametru Diabet (0 – nu are diabet, 1 – are diabet). Din grafice se observă faptul că independent cei cinci parametri nu pot prezice dacă o persoană are sau nu diabet

SLIDE 11

În urma realizării corelațiilor dintre variabile se poate observa că parametrul vârstă este cel mai puternic corelat cu variabila dependentă Diabet 0,34, fiind urmat, de insulină cu 0,25, indicele de masă corporală cu 0,21, glucoza 0,13, și cel mai slab corelat este tensiunea arterială diastolică cu 0,091. Din matricea de corelare ilustrată în dreapta se observă că parametrii sunt slab corelați între ei.

SLIDE 12

Pentru implementare predicției am ales să utilizez un arbore de decizie.

SLIDE 13

Arborele de decizie este o tehnică de învățare supervizată, datele fiind etichetate. Este un clasificator cu structură de arbore, nodurile interne reprezintă caracteristicile setului de date, ramurile regulile de decizie, iar nodurile frunză reprezintă rezultatele.

SLIDE 14

Descrierea algoritmului Pasul 1: Se găsește cel mai bun atribut din set, utilizând ASM (eiesem)(Attribute Selection Measure)prin intermediul căruia se calculează câștigul de informație. ASM-UL exprimă cât de multă informație furnizează o caracteristică despre o clasă. Pe baza acestei informații construim arborele. Astfel, nodul cu cel mai mare câștig de informație este ramnificat primul. Pasul 2: Se împarte nodul rădăcină în subseturi ale valorilor posibile ale celui mai bun atribut ales la pasul 1. Pasul 3: Se generează nodul care conține cel mai bun atribut, iar procesul continua până când nu se mai poate continua clasificarea.

SLIDE 15

Etapele implementării predicției, în plus față de acuratețe, precizie, recall și matricea de confuzie am mai creat patru indicatori pentru a putea observa evoluția predicției, și anume. Indicatorul 1 ilustrează cât la sută reprezintă predicțiile corecte, indicatorul 2 cât la sută reprezintă predicțiile incorecte, iar indicatorii 3 și 4 sunt cei mai importanți în cadrul evoluării predicției. Indicatorul 3 reprezintă câte persoane au fost diagnosticate pozitiv, dar în realitate erau negative, iar indicatorul 4 reprezintă câte persoane care au diabet, însă predicția a spune că nu suferă de această boală, indicatori 3 si 4 fiind raportati la numărul total de predicții greșite.

SLIDE 16

Până la a ajunge la predicția care să îndeplinească toate aspectele de interes, și anume FN mai mic decât FP și indicator 4 mai mic decât indicator 3 s-au implementat mai multe variante. Pe scurt, ce ne interesează este ca numărul predicțiilor incorecte să fie cât mai mic, iar numărul persoanelor care au diabet, dar predicția a fost că nu au să fie mai mic decât cazurile de fals pozitiv. De ce? Pentru că o persoană a cărui rezultat este pozitiv, în cele mai multe cazuri o să efectueze investigații suplimentare, pe când o persoană

cu rezultat FN, persoana are diabet, dar rezultatul indică că nu are, puțin probabil o să ceară analize suplimentare.

Prima variantă a presupus utilizarea a 4 parametri, glucoza, tensiunea arterial diastolică, indicele de masă corporală și vârsta, pe baza cărora s-a efectuat predicția. repartizarea datelor s-a faăcut astfel– 25% date de testare și 75% date de antrenare. Dacă analizăm indicatorii observăm faptul că FN este mai mare decât FP, iar indicatorul 4 este mai mare decât indicatorul 3, 0,50 cu 0,49. Rezultatul nu se încadrează cerințelor.

SLIDE 17

A doua variantă a adus, față de prima, adăugarea unui nou parametru, insulina. Se observă îmbunătățiri pentru toți parametrii studiați, însă indicatorul 4 este în continuare mai mare decât indicatorul 3, 0,51 cu 0,48.

SLIDE 18

A treia variantă și ultima a adus ca schimbare, față de cea de-a doua variantă, modificarea distribuției datelor, s-a mărit procentul datelor pentru antrenare, de la 75% la 86%. Această schimbare a condus la atingerea obiectivelor stabilite, FN mai mic decât FP și indicator 4 mai mic decât indicator3. Se observă și o îmbunătățire a celorlalți parametrii.

SLIDE 19

Pentru implementarea predicției am folosit ca limbaj de programare python, pentru crearea modelului bibliotecile pandas,numpy și scikits learn(saichiț), iar pentru interfața cu utilizatorul s-a folosit tkinter(tichintăr).

SLIDE 20

După crearea aplicației am vrut să testez modelul pe un set de analize a unor pacienți diagnosticați cu diabet. Se poate observa ca Din cele 10 înregistrări 8 au fost predicționate corect și două incorect.

SLIDE 21

În cadrul proiectului am reușit să realizez o aplicație de predicția a diabetului, utilizând un algoritm de învățare supervizată. Prin adăugarea unui parametru (și anume insulina) și creșterea numărului de date de învățare am îmbunătățit modelul și am atins cerința stabilită FN < FP.

S-a utilizat ca mediu de dezvoltare Spyder (Python 3.7).