Բովանդակություն

Ներածություն

Գլուխ 1 Մեքենայական ուսուցման և տվյալների վերլուծության փոխգործակցությունը

1.1 Տվյալների վերլուծություն

1.2 Մեքենայական ուսուցում

1.2.1 Մեքենայական ուսուցման տեսակները

Գլուխ 2 Դասակարգման մեթոդներ

2.1 Լոգիստիկ ռեգրեսիա (Logistic Regression)

2.2 Նաիվ Բայեսի դասակարգիչ (Naive Bayes Classifier)

2.3 K-ամենամոտ հարևաններ (K-Nearest Neighbors, KNN)

2.4 Գծային դիսկրիմինանտ վերլուծություն (Linear Discriminant Analysis, LDA)

2.5 Քառակուսային դիսկրիմինանտ վերլուծություն (Quadratic Discriminant Analysis, QDA)

Գլուխ 3 Ծրագրային մաս

3.1 Տվյալների հավաքածուի նկարագրություն և խնդրի ձևակերպում

3.2 Տվյալների նախնական վիճակագրական վերլուծություն

3.2.1 Վիճակագրական հաշվարկների արդյունքներ

3.3 Լոգիստիկ ռեգրեսիայի մոդելի կառուցում և կիրառություն

3.3.1 Տվյալների նախապատրաստում

3.3.2 Տվյալների բաժանում ուսուցման և փորձարկման խմբերի

3.3.3 Մոդելի գնահատման չափանիշներ

3.3.4 Լոգիստիկ ռեգրեսիայի մոդելի արդյունքների գնահատում

3.6 K-Nearest Neighbors մոդելի կառուցում և կիրառություն

3.6.1 Տվյալների նախապատրաստում

3.6.2 Տվյալների բաժանում ուսուցման և փորձարկման խմբերի

3.6.3 *k*պարամետրի ընտրություն

3.6.4 KNN մոդելի կիրառություն

3.6.5 Մոդելի գնահատման չափանիշներ

3.6.6 KNN մոդելի արդյունքների ներկայացում

3.7 Naive Bayes մոդելի կառուցում և կիրառություն

3.7.1 Տվյալների նախապատրաստում

3.7.2 Տվյալների բաժանում ուսուցման և փորձարկման խմբերի

3.7.3 Naive Bayes մոդելի կիրառություն

3.7.4 Մոդելի գնահատման չափանիշներ

3.7.5 Naive Bayes մոդելի արդյունքների ներկայացում

3.8 Մոդելների համեմատական վերլուծություն

Եզրակացություն

Օգտագործված գրականության ցանկ

Ներածություն

Վերջին տարիներին տեղեկատվական տեխնոլոգիաների զարգացումը և թվայնացման գործընթացների ակտիվացումը հանգեցրել են տվյալների կտրուկ աճի տարբեր ոլորտներում: Այս պայմաններում տվյալների արդյունավետ վերլուծությունը և դրանց հիման վրա կանխատեսումների իրականացումը դարձել են ժամանակակից գիտության և կիրառական խնդիրների կարևորագույն ուղղություններից մեկը: Առողջապահության ոլորտում այս մոտեցումները առանձնահատուկ նշանակություն ունեն, քանի որ հիվանդների վիճակի գնահատումը և ռիսկերի ճիշտ կանխատեսումը կարող են էական ազդեցություն ունենալ բուժման ընթացքի և արդյունքների վրա:

COVID-19 համավարակը ցույց տվեց, որ հիվանդության ընթացքը տարբեր հիվանդների մոտ կարող է էապես տարբերվել, և առաջանում է անհրաժեշտություն գնահատել յուրաքանչյուր հիվանդի ռիսկի մակարդակը: Այս խնդիրը հնարավոր է լուծել մեքենայական ուսուցման մեթոդների միջոցով, որոնք թույլ են տալիս մեծածավալ տվյալների հիման վրա հայտնաբերել օրինաչափություններ և կառուցել կանխատեսման մոդելներ:

Սույն աշխատանքը նվիրված է COVID-19 հիվանդների տվյալների հիման վրա ռիսկի մակարդակի կանխատեսման խնդրի ուսումնասիրմանը: Աշխատանքում կիրառվում են դասակարգման երեք մեթոդներ՝ Logistic Regression, K-Nearest Neighbors և Naive Bayes, որոնք օգտագործվում են հիվանդների բարձր և ցածր ռիսկային խմբերին պատկանելությունը որոշելու համար:

Գործնական մասում իրականացվում է տվյալների նախնական մշակում, ներառյալ բացակայող արժեքների մշակումը և նպատակային փոփոխականի կառուցումը, ինչպես նաև մոդելների ուսուցում և գնահատում: Արդյունքների գնահատման համար կիրառվում են accuracy, precision, recall և F1-score չափանիշները, որոնք թույլ են տալիս համեմատել տարբեր մոդելների արդյունավետությունը:

Աշխատանքի նպատակն է ուսումնասիրել մեքենայական ուսուցման մեթոդների կիրառելիությունը իրական բժշկական տվյալների վրա և գնահատել, թե որ մոդելն է

առավել արդյունավետ COVID-19 հիվանդների ռիսկի մակարդակի կանխատեսման համար։

Գլուխ 1 Մեքենայական ուսուցման և տվյալների վերլուծության փոխգործակցությունը

1.1. Տվյալների վերլուծություն

Տվյալների վերլուծությունը գործընթաց է, որը հնարավորություն է տալիս կուտակված կամ փորձարկումների և դիտարկումների միջոցով ստացված չմշակված (հում) տվյալները հետազոտել, մաքրել, փոխակերպել և մոդելավորել՝ նպատակ ունենալով նրանցից ստանալ օգտակար ինֆորմացիա, անել եզրահանգումներ, ընդունել որոշումներ:

Ներկայումս կարելի է ասել տվյալներն ամենուր են. գրեթե բոլոր ոլորտներում առընչվում ենք տվյալների հետ և կազմակերպություններն ու հետազոտողները իրենց առջև դրված զանազան խնդիրների լուծումը, առաջացած հարցերի պատասխաններ փորձում են ստանալ այդ տվյալների վերլուծության ճանապարհով:, Այն թույլ է տալիս կատարել կանխատեսումներ, կառուցել ստրատեգիաներ, որոշումներ ընդունել և իրականացնել այլ գործառույթներ:

Տվյալների վերլուծությունը իրականացվում է մի քանի փուլերով, որոնք ամփոփ նկարագրվում են հետևյալ փուլերով՝

- ✓ Տվյալների հավաքագրում, ձեռքվերում: Տվյալագիտության, մասնավորապես տվյալների վերլուծության կարևորագույն խնդիրներից մեկը ճիշտ տվյալների հավաքածու ընտրելն է: Տվյալների հավաքագրման տարբեր տեսակներ կան՝ կախված թե որ ոլորտից, ինչպիսի աղբյուրներից են դրանք կուտակվում: Դրանք կարող են կուտակվել տարբեր աղբյուրներից. ընկերությունների, հետազոտողների, կառավարական, վեբ կայքերից և այլ գործունեություններից առաջաց օրինակստացվող:
- ✓ Տվյալների մաքրում: Սխալ, անհամապատասխան կամ անհասկանալի տվյալների հայտնաբերում, ուղղում կամ հեռացում՝ վերլուծության ճշտությունն ու հուսալիությունը ապահովելու համար:
- ✓ Տվյալների ձևափոխում: Մաքրված տվյալների ձևափոխում՝ վերլուծության և մոդելավորման համար հարմար տեսքի բերելու համար: Սա իր հերթին բազմաճյուղ

գործընթաց է, կախված ձեռք բերված տվյալների բնույթից, վերլուծության պահանջներից և նպատակներից։

✓ Տվյալների վիզուալիզացիա և հիմական բնութագրիչների հաշվարկներ։ Սա այն փուլն է, որը թույլ է տալիս նախնական պատկերացում կազմել տվյալների վերաբերյալ, տեսնել տվյալների կառուցվածքը, գնահատել հնարավոր խնդիրներ ու բացթողումները։

✓ Տվյալների մոդելավորում։ Տվյալների մոդելավորումը տվյալների վերլուծության հիմական փուլն է, որը ներկայացնում է տվյալ տվյալներին և առաջադրվող հարցերին համապատասխան վերլուծական ալգորիթմի կիրառումը՝ որոշումների ընդունման, կանխատեսման, դասակարգման և այլ նպատակով։

1.2. Մեքենայական ուսուցում

Մեքենայական ուսուցումը, որը հաճախ կրճատվում է որպես ML, արհեստական ինտելեկտի (AI) ենթաճյուղ է, որը թույլ է տալիս համակարգիչներին մուտք արված տվյալներից սովորել և ըստ այդմ որոշումներ կամ կանխատեսումներ կայացնել։ Այլ կերպ ասած տվյալագետը համակարգչին տրամադրում է մեծ քանակությամբ տվյալները և ML ալգորթմների կիրառումը հնարավորություն է տալիս ինքնուրույն բացահայտել օրինաչափություններ, իրականացնել կանխատեսումներ կամ կայացնել որոշումներ։

Իր հիմքում մեքենայական ուսուցումը որոշումների կայացման և կանխատեսման նպատակով ստեղծում է մոդելներ և դրանք իրականացնում է մեծածավալ տվյալների վրա։

Տվյալների վերլուծության գործընթացը հիմք է ստեղծում ML մոդելների կառուցման համար։ Մեքենայական ուսուցման ցանկացած նախագիծ սկսվում է տվյալների վերլուծության քայլերից։ Մեքենայական ուսուցման ցանկացած նախագիծ սկսվում է տվյալների վերլուծության քայլերից. Վերլուծաբանը հավաքում, մաքրում և նախապատրաստում է հում տվյալները։ Սա շատ կարևոր փուլ է, քանի որ հակառակ դեպքում ML մոդելը կսովորի տվյալների մեջ առկա ցանկացած սխալ կամ կողմնակալություն, որը կարող է հետագայում, օրինակ, կանխատեսումների ժամանակ բերել սխալ արդյունքի։ Տվյալների կառուցվածքը հասկանալու, անոմալների առկայությունը բացահայտելու առումով դարձյալ կարևորվում է տվյալների վերլուծության վիզուալիզացիայի մոտեցումները։ ML մոդելների ճշգրտությունը

ապահովելու առումով տվյալների վերլուծաբանը կարիք է ունենում առկա տվյալներից ստեղծել ավելի իմաստալից հատկանիշներ։

Երբ տվյալները մաքուր են, կառուցվածքային և պատրաստ, ML մասնագետը կախված լուծվելիք խնդրից իրականացնում է մոդելի ընտրություն համապատասխան ալգորիթմի կիրառմամբ։ Այնուհետև իրականացվում է մոդելի մարզումը, այն է՝ տվյալներից սովորելը, որից հետո իրականացվում է մոդելի գնահատումը՝ մոդելի աշխատանքի ճշգրտությունը և արդյունավետությունը։

Վերջնական արդյունքը կրկին վերադառնում է վերլուծության ոլորտ, այսինքն ML մոդելը տրամադրում է կանխատեսումներ, որոնք էլ տվյալների վերլուծաբանը ներկայացնում և մեկնաբանում է որոշումներ կայացնելու առումով, տվյալ ոլորտի ներկայացուցիչներին հասկանալի լեզվով և տերմինաբանությամբ։

1.2.1 Մեքենայական ուսուցման տեսակները

Մեքենայական ուսուցումը ալգորիթմները կարելի է բաժանել երեք տեսակի՝

--*Վերահսկվող ուսուցում։* Մոդելը մարզվում է պիտակավորված տվյալների վրա։ Մուտքային տվյալներից յուրաքանչյուրը ունի իր ելքային արժեքը, որը կոչվում է պիտակ։ Այստեղ նպատակն է կանխատեսել ելքային պիտակը՝ հինվելով մուտքային տվյալների և պիտակների միջև հարաբերությունների վրա։ Այս մոդելներից են ռեգրեսիայի և դասակարգման մոդելները։ ML-ում դասակարգման համար օգտագործվում են ինչպես հիմական վիճակագրական մեթոդներ (օրինակ՝ Լոգիստիկ ռեգրեսիա, LDA, QDA, Naive Bayes), այնպես էլ ժամամակակից (օրինակ՝ KNN, Decision Trees, Random Forest, Support Vector Machines, Neural Networks)։

---Չվերահսկվող ուսուցում։ Մոդելը մարզվում է չպիտակավորված տվյալների վրա։ Այստեղ մոդելը պետք է ինքնուրույն բացահայտի թաքնված օրինաչափությունները, հայտնավերի տվյալների ներքին կառուցվածքը, դրանք կլաստերացնի կամ նվազեցնի չափողականությունը՝ չկորցնելով կարևոր տեղեկություններ։

---Ուժեղացված ուսուցում։ Ուժեղացված ուսուցումը տարբերվում է վերահսկվող ուսուցումից նրանով, որ կարիք չկա պիտակավորված մուտքային կամ ելքային զույգերի ներկայացման և ոչ օպտիմալ գործողությունների հստակ ուղղման կարիք։ Մոդելը սովորում է փորձի և սխալի միջոցով։ Յուրաքանչյուր ճիշտ գործողությունը տալիս է պարգև (reward), իսկ սխալը՝ պատիժ (penalty)։ Մոդլի նպատակն է սովորել

գործողությունների հերթականությունը, որը կմեծացնի կուտակային պարգևը: Այս մոդելները կիրառվում են ռոբոտաշինությունում, ավտոմատացված խաղերում, ինքնավար մեքենաներում:

Գլուխ 2 Դասակարգման մեթոդներ

Ամենատարածված դասակարգման մեթոդներից են լոգիստիկ ռեգրեսիան(logistic regression), գծային դիսկրիմինանտ վերլուծություն (linear discriminant analysis), քառակուսային դիսկրիմինանտ վերլուծություն (quadratic dis criminant analysis), Նաիվ Բայեսի դասակարգիչ (Naive Bayes Classifier) և K-ամենամոտ հարևաններ(K-Nearest Neighbors, KNN):

2.1 Լոգիստիկ ռեգրեսիա (Logistic Regression)

Լոգիստիկ ռեգրեսիան դասակարգման ամենահիմնական և ամենատարածված մեթոդներից է, որը կիրառվում է այն դեպքերում, երբ կախյալ փոփոխականը երկակի (բինար) է՝ օրինակ՝ 0/1, այո/ոչ, դրական/բացասական: Մեթոդի նպատակը տվյալ դիտարկման համար որոշելն է, թե ինչ հավանականությամբ է այն պատկանում դրական (1) դասին՝ հիմնվելով մուտքային (անկախ) փոփոխականների արժեքների վրա:

Ի տարբերություն գծային ռեգրեսիայի, որը վերադարձնում է անընդհատ արժեքներ և կարող է դուրս գալ $(-\infty, +\infty)$ միջակայքից, լոգիստիկ ռեգրեսիան ապահովում է, որ կանխատեսված արժեքը լինի 0–1 միջակայքում, ինչը թույլ է տալիս այն մեկնաբանել որպես հավանականություն:

Լոգիստիկական ռեգրեսիան արտահայտում է վիճակագրական կապը՝ հետևյալ կախվածության տեսքով՝

$$P\{\boldsymbol{Y} = 1|\boldsymbol{X} = (X_1, X_2, \dots, X_K) = F(\boldsymbol{X}):$$

Այսինքն՝ մոդելն արտահայտում է կախվածությունը Y=1 արժեքի տեղի ունենալու հավանականության և անկախ փոփոխականների արժեքների միջև. կախվածության օրենքը անկախ փոփոխականների կոնկրետ ազդեցության դեպքում արդյունքային փոփոխականի 1 արժեքով ելքի հավանականության միջև է:

Լոգիստիկ մոդելը սակայն ընդունված է ներկայացնել հետազոտվող երևույթի տեղի ունենալու և տեղի չունենալու հավանականությունների հարաբերության միջոցով՝

$$\frac{P\{Y_i = 1\}}{P\{Y_i = 0\}} = \frac{P\{Y_i = 1\}}{1 - P\{Y_i = 1\}} \qquad (1)$$

Այս հարաբերությունը կոչվում է շանսերի (նախընտրությունների) հարաբերակցություն:

Մոդելը նկարագրվում է շանսերի հարաբերակցության լոգարիթմի միջոցով արտահայտվող հետևյալ գծային կապի տեսքով՝

$$ln\left(\frac{P\{y_i = 1\}}{P\{y_i = 0\}}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i, \qquad i = 1,2, \dots n \qquad (2)$$

Այստեղ ε_i – երը բաշխված են լոգիստիկ օրենքով: Այս հավասարումը ցույց է տալիս, թե մուտքային փոփոխականների ինչ կոմբինացիա է մեծացնում կամ նվազեցնում դիտարկման դրական դասին պատկանելու հավանականությունը:

Պարամետրերի գնահատում: Լոգիստիկ ռեգրեսիայի պարամետրերը (β-ները) սովորաբար գնահատվում են Առավելագույն ճշմարտանմանության մեթոդով:

2.2 Նաիվ Բայեսի դասակարգիչ (Naive Bayes Classifier)

Նաիվ Բայեսի դասակարգիչը հիմված է Բայեսի տեսության վրա և ենթադրում է, որ բոլոր մուտքային փոփոխականները անկախ են իրարից՝ անկախ իրականությունից: Այս ենթադրությունը շատ հաճախ աշխատում է լավ, նույնիսկ, եթե իրականում փոփոխականները մասնակի կախված են: Յուրաքանչյուր դիտարկման համար հաշվում ենք տարբեր դասերի հավանականությունները ըստ Բայեսի տեսության, և դասակարգում այն դասին, որի հավանականությունը առավելագույնն է: Մոդելը շատ արագ է աշխատում և լավ է աշխատում մեծ տվյալների և բարձր չափաբաշխումների դեպքում:

Մաթեմատիկական մոդելն է՝

$$P(C_k \mid x_1, x_2, \ldots, x_n) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(x_1, x_2, \ldots, x_n)}$$

որտեղ՝

- ✓ C_k— k-րդ դասը,
- ✓ x_i— մուտքային փոփոխականները,
- ✓ $P(C_k)$— priori հավանականությունը,
- ✓ $P(x_i \mid C_k)$— դասի համար համապատասխան փոփոխականի պայմանական հավանականությունը։

Օրինակ. Էլ. փոստերի spam-ֆիլտրացիայի մեջ՝ յուրաքանչյուր բառը դիտարկվում է որպես փոփոխական։ Նաիվ Բայեսը հաշվում է, թե տվյալ բառերի հավաքածուն որքանով է հավանական spam կամ ոչ spam էլ. փոստի համար, և դասակարգում է համապատասխան կերպ։

2.3 K-ամենամոտ հարևաններ (K-Nearest Neighbors, KNN)

KNN-ը ոչ-պարամետրական և instance-based (նմուշի վրա հիմված) մեթոդ է, որը գտնում է մուտքային տվյալի K ամենամոտ "հարևաններին" և դասակարգում այն ըստ դրանց։ KNN մեթոդը չունի մոդելի կառուցում կամ պարամետրերի գնահատում։ Այստեղ K-ն որոշվում է, թե քանի հարևան հաշվելուց հետո է որոշում կայացվում։, իհարկե հաշվի է առնվում նաև հարևանների միջև հեռավորությունը (օրինակ, Էվկլիդյան)։

Օրինակ. Եթե մենք ունենք տվյալներ ուսանողների գնահատականների և մասնակցության մասին, նոր ուսանողի դասը որոշելու համար KNN-ը կտեսնի, որ նրա ամենամոտ K ուսանողները հիմնականում բարձր գնահատականներով են, և, հետեւաբար, նոր ուսանողը դասակարգվում է բարձր գնահատական ունեցող դասում։

2.4 Գծային դիսկրիմինանտ վերլուծություն (Linear Discriminant Analysis, LDA)

Գծային դիսկրիմինանտ վերլուծությունը դասակարգման մեթոդ է, որը օգտագործվում է դիտարկումները երկու կամ ավելի դասերի բաժանելու համար՝ հիմված տվյալների հատկությունների (մուտքային փոփոխականների) վրա։ Մեթոդի հիմնական նպատակն է գտնել այն գծային կոմբինացիան, որը առավելագույն

տարբերություն է ստեղծում դասերի միջև և միաժամանակ նվազեցնում դասերի ներսում գոյացող փոփոխականությունը: LDA գնահատում է դասերի միջինների և կովարիացիոնների հիման վրա ստեղծվող դիսկրիմինանտ ֆունկցիաներ, որոնք թույլ են տալիս տարբերակել դասերը: Յուրաքանչյուր դիտարկում դասակարգվում է այն դասին, որի համար հաշված հավանականությունը (posterior probability) առավելագույնն է:

Մաթեմատիկական մոդելն է՝

$$f_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln(\pi_k)$$

որտեղ՝

- ✓ x— դիտարկման մուտքային վեկտորը,
- ✓ μ_k— k-րդ դասի միջին վեկտոր,
- ✓ Σ— ընդհանուր կովարիացիոն մատրիցա,
- ✓ π_k— k-րդ դասի priori հավանականությունը,
- ✓ $f_k(x)$— k-րդ դասի համար հաշվարկված գնահատականը (score):

Դիտարկումը դասակարգվում է այն դասին, որի $f_k(x)$ արժեքը առավելագույնն է:

Օրինակ. Ենթադրենք, որ ունենք բժշկական տվյալներ՝ հիվանդների հասակի, քաշի ու այլ չափումների վերաբերյալ, և ցանկանում ենք կանխատեսել, թե հիվանդը վտանգավոր խախտման (High Risk) կամ նվազ ռիսկի (Low Risk) դասին է պատկանում:LDA-ն կստեղծի դիսկրիմինանտ ֆունկցիաները՝ հիմված այդ չափումների միջինների և կովարիացիոնների վրա, և յուրաքանչյուր հիվանդ կդասակարգվի այն դասին, որի ֆունկցիան առավելագույն է:

2.5 Քառակուսային դիսկրիմինանտ վերլուծություն (Quadratic Discriminant Analysis, QDA)

Քառակուսային դիսկրիմինանտ վերլուծությունը (QDA) շատ նման է LDA-ին, սակայն թույլ է տալիս, որ յուրաքանչյուր դաս ունենա իր առանձին կովարիացիոն մատրիցան: Սա նշանակում է, որ դասերի ներսում փոփոխականությունները կարող են տարբեր լինել, և դիսկրիմինանտ ֆունկցիաները չեն սահմանափակվում գծային տեսքով: QDA-ն

հատկապես օգտակար է, երբ տվյալների դասերը չեն կարող լավ բաժանվել մեկ գծային սահմանով, և դասերի տարբերությունը բարդ (քառակուսային) է: Յուրաքանչյուր դիտարկում դասակարգվում է այն դասին, որի posterior probability-ն առավելագույն է:

Մաթեմատիկական մոդելն է՝

$$f_k(x) = -\frac{1}{2}\ln \mid \Sigma_k \mid -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \ln(\pi_k)$$

որտեղ՝

- ✓ x— դիտարկման մուտքային վեկտորը,
- ✓ μ_k— k-րդ դասի միջին վեկտոր,
- ✓ Σ_k— k-րդ դասի կովարիացիոն մատրիցա,
- ✓ π_k— k-րդ դասի priori հավանականություն,
- ✓ $f_k(x)$— k-րդ դասի հաշվարկված գնահատականը:

Օրինակ. Եթե բժշկական տվյալների հիման վրա ցանկանում ենք գնահատել հիվանդների ռիսկը՝ որոշ չափումների հիման վրա, QDA-ն կկարողանա ավելի ճկուն կերպով տարբերակել բարձր և ցածր ռիսկ ունեցող հիվանդներին՝ հաշվի առնելով, որ յուրաքանչյուր դասի փոփոխականությունների տարածքը տարբեր է:

2.7 Տվյալների նկարագրական վիճակագրական բնութագրիչներ

Այս ենթագլխում ներկայացվում են տվյալների նկարագրական վիճակագրական հիմնական բնութագրիչները, որոնք կիրառվում են տվյալների նախնական վերլուծության ընթացքում: Դրանք հնարավորություն են տալիս գնահատել տվյալների կենտրոնական միտումը և տարածվածությունը, ինչպես նաև ընդհանուր պատկերացում կազմել տվյալների բաշխման վերաբերյալ: Մասնավորապես, դիտարկվում են միջին արժեքը, մեդիանը և ստանդարտ շեղումը, որոնք լայնորեն կիրառվում են տվյալների վերլուծության մեջ և հիմք են հանդիսանում հետագա վիճակագրական և մեքենայական ուսուցման մեթոդների կիրառման համար:

Միջին արժեք. ցույց է տալիս տվյալների կենտրոնական միտումը և հաշվարկվում է հետևյալ բանաձևով.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

որտեղ՝

- ✓ x_i - տվյալների արժեքներն են,
- ✓ n - դիտարկումների ընդհանուր քանակը։

Մեդիան. այն արժեքն է, որը բաժանում է տվյալների շարքը երկու հավասար մասերի։

Տվյալների տարածվածություն. սա գնահատելու համար օգտագործվում է **ստանդարտ շեղումը**, որը ցույց է տալիս, թե որքանով են տվյալները շեղված միջին արժեքից։ Հաշվարկվում է հետևյալ բանաձևով.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

որտեղ՝

- ✓ $\bar{x}$ - միջին արժեքն է։

Հաշվարկվում են նաև տվյալների **նվազագույն** և **առավելագույն արժեքները**, որոնք թույլ են տալիս գնահատել տվյալների ամբողջ միջակայքը։

Բացի քանակական հատկանիշներից, տվյալների հավաքածուում առկա են նաև **բինար** և **կատեգորիական** հատկանիշներ։ Այսպիսի հատկանիշները ներկայացնում են որոշակի վիճակի կամ հատկության առկայությունը և սովորաբար ունեն սահմանափակ թվով հնարավոր արժեքներ։

Բինար հատկանիշների դեպքում տվյալների արժեքները կարող են ընդունել երկու հիմնական արժեք.

- ✓ 1 - այո (առկա է)
- ✓ 2 - ոչ (առկա չէ)

Տվյալների հավաքածուում բինար հատկանիշներ են.

- ✓ pneumonia
- ✓ pregnancy
- ✓ diabetes
- ✓ copd
- ✓ asthma
- ✓ inmsupr
- ✓ hypertension

- ✓ cardiovascular
- ✓ renal_chronic
- ✓ other_disease
- ✓ obesity
- ✓ tobacco
- ✓ intubed
- ✓ icu
- ✓ usmer
- ✓ sex

Այս հատկանիշների դեպքում հաշվարկվում է յուրաքանչյուր արժեքի **հաճախականությունը** (frequency), որը ցույց է տալիս, թե տվյալների հավաքածուում քանի դիտարկում ունի տվյալ հատկանիշի կոնկրետ արժեքը։

Հաճախականությունը հաշվարկվում է հետևյալ ձևով.

$$f_i = \frac{n_i}{n}$$

որտեղ՝

- ✓ n_i - տվյալ արժեք ունեցող դիտարկումների քանակն է,
- ✓ n - դիտարկումների ընդհանուր քանակը։

Կատեգորիական հատկանիշների դեպքում հաշվարկվում է յուրաքանչյուր կատեգորիայի դիտարկումների քանակը և դրանց հարաբերական հաճախականությունը։

Այս տվյալների հավաքածուում կատեգորիական հատկանիշներն են.

- ✓ medical_unit
- ✓ patient_type
- ✓ classification_final

Հատուկ փոփոխական

date_died

Այս փոփոխականը ներկայացնում է հիվանդի մահվան ամսաթիվը։

- ✓ եթե արժեքը 9999-99-99 է, նշանակում է հիվանդը չի մահացել
- ✓ այլ դեպքում նշվում է մահվան ամսաթիվը

Այն օգտագործվում է նպատակային փոփոխական (target) կառուցելու համար։

3.7 Լոգիստիկ ռեգրեսիայի վիճակագրական մոդել

Լոգիստիկ ռեգրեսիան վիճակագրական մեթոդ է, որը կիրառվում է այն դեպքերում, երբ նպատակային փոփոխականը ունի երկու հնարավոր արժեք։ Այսպիսի խնդիրները կոչվում են բինար դասակարգման խնդիրներ։ Լոգիստիկ ռեգրեսիայի նպատակը գնահատելն է որոշակի իրադարձության տեղի ունենալու հավանականությունը՝ հիմնվելով մի քանի բացատրող կամ կանխատեսող փոփոխականների վրա։

Այս մեթոդը լայնորեն կիրառվում է տարբեր ոլորտներում, հատկապես բժշկական տվյալների վերլուծության մեջ, որտեղ հաճախ անհրաժեշտ է կանխատեսել հիվանդության առաջացման կամ ծանր ընթացքի հավանականությունը՝ հիմնվելով հիվանդի առողջական տվյալների վրա։

Լոգիստիկ ռեգրեսիան թույլ է տալիս գնահատել հավանականությունը, որ դիտարկումը պատկանում է որոշակի դասի։ Եթե այդ հավանականությունը գերազանցում է որոշակի շեմային արժեք (օրինակ՝ 0.5), ապա դիտարկումը դասակարգվում է որպես դրական դեպք, հակառակ դեպքում՝ բացասական։

Լոգիստիկ ֆունկցիա

Լոգիստիկ ռեգրեսիայի հիմքում ընկած է լոգիստիկ ֆունկցիան, որը սահմանվում է հետևյալ կերպ.

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-z}}$$

որտեղ

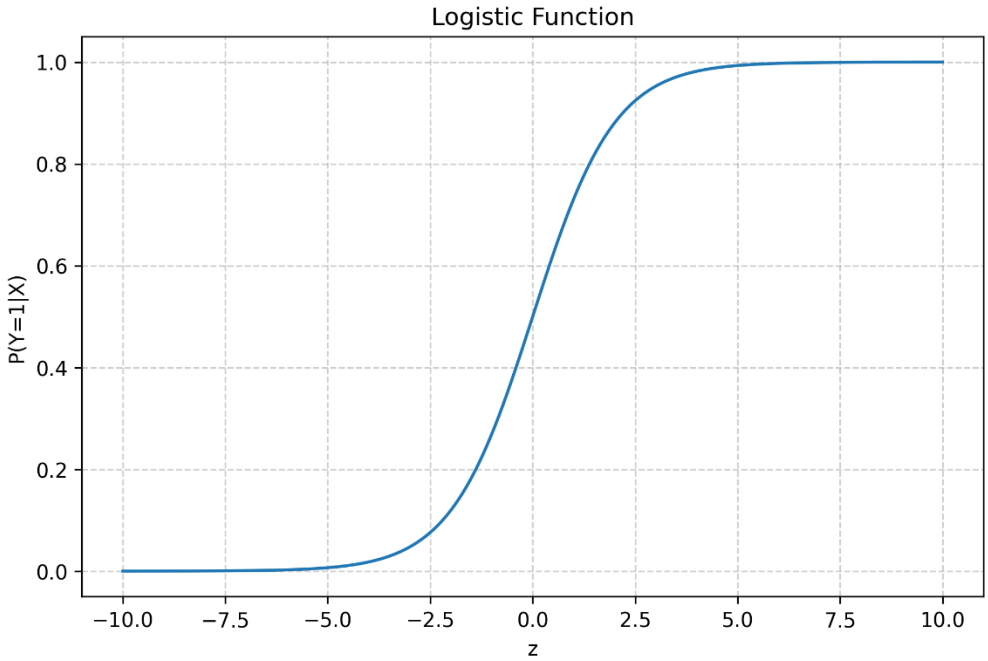
$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Այստեղ՝

- ✓ $x_1, x_2, \ldots, x_k$ - կանխատեսող փոփոխականներն են,
- ✓ β_0 - ազատ անդամ է,
- ✓ $\beta_1, \beta_2, \ldots, \beta_k$ - մոդելի գործակիցներն են։

Այս ֆունկցիայի կարևոր հատկությունն այն է, որ այն միշտ վերադարձնում է 0-ից 1 միջակայքում գտնվող արժեքներ, որոնք կարելի է մեկնաբանել որպես հավանականություններ։

Լոգիստիկ ֆունկցիայի գրաֆիկը ունի S-աձև տեսք և ցույց է տալիս, թե ինչպես է փոփոխվում իրադարձության հավանականությունը կախված բացատրող փոփոխականներից։



Logit փոխակերպում

Լոգիստիկ ռեգրեսիան իրականում մոդելավորում է ոչ թե հավանականությունը, այլ այդ հավանականության լոգարիթմական հարաբերությունը, որը կոչվում է logit։

Logit ֆունկցիան սահմանվում է հետևյալ կերպ.

$$\log\left(\frac{p}{1-p}\right)$$

որտեղ p-ը իրադարձության հավանականությունն է։

Այսպիսով լոգիստիկ ռեգրեսիայի մոդելը կարելի է ներկայացնել հետևյալ ձևով.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Այս փոխակերպումը թույլ է տալիս հավանականությունների ոչ գծային կախվածությունը ներկայացնել գծային մոդելի տեսքով։

Պարամետրերի գնահատում

Լոգիստիկ ռեգրեսիայի մոդելի գործակիցները գնահատվում են առավելագույն հավանականության մեթոդի (Maximum Likelihood Estimation) միջոցով։ Այս մեթոդի նպատակն է գտնել այնպիսի պարամետրերի արժեքներ, որոնց դեպքում դիտարկված տվյալների ստացման հավանականությունը առավելագույնն է։

Հավանականության ֆունկցիան կարելի է ներկայացնել հետևյալ ձևով.

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

որտեղ՝

y_i— նպատակային փոփոխականի արժեքն է,
p_i— մոդելի կողմից հաշվարկված հավանականությունը։

Այս մեթոդը թույլ է տալիս գնահատել մոդելի պարամետրերը այնպես, որ ստացված մոդելը լավագույնս նկարագրի տվյալները։

Գործակիցների մեկնաբանություն

Լոգիստիկ ռեգրեսիայի գործակիցները հաճախ մեկնաբանվում են **Odds ratio**-ի միջոցով։ Odds ratio-ն ցույց է տալիս, թե որքանով է փոխվում իրադարձության հավանականությունը տվյալ փոփոխականի փոփոխության դեպքում։

Odds ratio-ն հաշվարկվում է հետևյալ կերպ.

$$OR = e^{\beta}$$

Եթե odds ratio-ի արժեքը մեծ է 1-ից, ապա տվյալ փոփոխականը մեծացնում է իրադարձության տեղի ունենալու հավանականությունը։ Եթե այն փոքր է 1-ից, ապա տվյալ փոփոխականը նվազեցնում է այդ հավանականությունը։

Այսպիսի մեկնաբանությունը լայնորեն կիրառվում է բժշկական և սոցիալական հետազոտություններում։

Գլուխ 3 Ծրագրային մաս

3.1 Տվյալների հավաքածուի նկարագրություն և խնդրի ձևակերպում

Այս աշխատանքում դասակարգման մեթոդների կիրառությունը ուսումնասիրելու նպատակով օգտագործվել է COVID-19 հիվանդների վերաբերյալ տվյալների հավաքածու։ Տվյալների հավաքածուն ներբեռնվել է բաց տվյալների հարթակից՝ Kaggle կայքից։

Հատկանիշների նկարագրություն

Տվյալների հավաքածուում ներկայացված հատկանիշների մեծ մասը հանդիսանում են բինար (boolean) փոփոխականներ, որոնց դեպքում արժեքը 1 նշանակում է «այո», իսկ 2 նշանակում է «ոչ»: Որոշ դաշտերում հանդիպող 97, 98 և 99 արժեքները ներկայացնում են բացակայող կամ անհայտ տվյալներ:

Տվյալների հավաքածուում ներառված հիմնական հատկանիշները հետևյալն են.

sex - հիվանդի սեռը:

1 - իգական, 2 - արական:

age - հիվանդի տարիքը:

classification_final – covid-19 թեստի վերջնական արդյունքը:

1-3 արժեքները նշանակում են, որ հիվանդը վարակված է covid-19-ով տարբեր աստիճաններով, իսկ 4 և ավելի արժեքները նշանակում են, որ հիվանդը վարակված չէ կամ թեստի արդյունքը անորոշ է:

patient_type - բուժման տեսակը:

1 - հիվանդը ստացել է ամբուլատոր բուժում և վերադարձել է տուն,

2 - հիվանդը հոսպիտալացվել է:

pneumonia – ցույց է տալիս, արդյոք հիվանդի մոտ առկա է թոքաբորբ:

pregnancy – ցույց է տալիս, արդյոք հիվանդը հղի է:

diabetes – ցույց է տալիս շաքարային դիաբետի առկայությունը:

copd – ցույց է տալիս քրոնիկական օբստրուկտիվ թոքային հիվանդության առկայությունը:

asthma – ցույց է տալիս ասթմայի առկայությունը հիվանդի մոտ:

inmsupr – ցույց է տալիս, արդյոք հիվանդը ունի թուլացած իմունային համակարգ:

hypertension – ցույց է տալիս բարձր արյան ճնշման առկայությունը:

cardiovascular – ցույց է տալիս սրտանոթային հիվանդությունների առկայությունը:

renal_chronic – ցույց է տալիս քրոնիկական երիկամային հիվանդության առկայությունը:

other_disease – ցույց է տալիս այլ ուղեկցող հիվանդությունների առկայությունը:

obesity – ցույց է տալիս գիրության առկայությունը հիվանդի մոտ:

tobacco – ցույց է տալիս, արդյոք հիվանդը ծխող է:

usmer – ցույց է տալիս, արդյոք հիվանդը բուժվել է առաջին, երկրորդ կամ երրորդ մակարդակի բժշկական հաստատությունում:

medical_unit – բժշկական հաստատության տեսակը, որը տրամադրել է բուժումը:

intubed – ցույց է տալիս, արդյոք հիվանդը միացված է եղել շնչառական ապարատին:

icu – ցույց է տալիս, արդյոք հիվանդը գտնվել է ինտենսիվ թերապիայի բաժանմունքում:

date_died – հիվանդի մահվան ամսաթիվը: Եթե արժեքը 9999-99-99 է, նշանակում է հիվանդը չի մահացել:

Խնդրի ձևակերպում

Այս աշխատանքի հիմնական նպատակն է կանխատեսել COVID-19 հիվանդի առողջական վիճակի ռիսկի մակարդակը (date_died հիման վրա կառուցված նպատակային փոփոխական)՝ հիմնվելով հիվանդի առողջական տվյալների (age, sex), ախտանիշների և կլինիկական վիճակի (pneumonia, intubed, icu, patient_type) և բժշկական պատմության (diabetes, hypertension, cardiovascular, renal_chronic, copd, asthma, obesity, other_disease, inmsupr) վրա:Տվյալների հավաքածուում ներկայացված տարբեր բժշկական հատկանիշները կարող են ազդել հիվանդության ընթացքի և դրա ծանրության վրա: Այդ պատճառով այս աշխատանքում դիտարկվում է դասակարգման խնդիր, որի նպատակն է որոշել, թե արդյոք տվյալ հիվանդը գտնվում է բարձր ռիսկային խմբում, թե ոչ:

Նպատակային փոփոխականը ներկայացվում է երկու դասերով.

- ✓ 1 - բարձր ռիսկ (high risk)

- ✓ 0 - ցածր ռիսկ (low risk)

Խնդրի լուծման համար կիրառվում են մեքենայական ուսուցման երկու դասակարգման մեթոդներ՝

- ✓ Logistic Regression [1]
- ✓ K-Nearest Neighbors (KNN) [2]
- ✓ Naïve Bayes [3]

Այս մեթոդների միջոցով կառուցված մոդելները հետագայում կգնահատվեն և կհամեմատվեն՝ օգտագործելով համապատասխան գնահատման չափանիշներ, ինչպիսիք են accuracy, precision, recall և F1-score, որպեսզի գնահատվի դրանց արդյունավետությունը տվյալ տվյալների հավաքածուի դեպքում:

3.2 Տվյալների նախնական վիճակագրական վերլուծություն

Նախնական վիճակագրական վերլուծությունը թույլ է տալիս հասկանալ տվյալների բաշխումը, հայտնաբերել հնարավոր արտասովոր արժեքներ և գնահատել տվյալների համապատասխանությունը հետագա մեքենայական ուսուցման մոդելների կիրառման համար:

Քանակական հատկանիշների համար հաշվարկվում են հետևյալ հիմնական ցուցանիշները.

- ✓ միջին արժեք (mean)
- ✓ մեդիան (median)
- ✓ ստանդարտ շեղում (standard deviation)
- ✓ նվազագույն արժեք (minimum)
- ✓ առավելագույն արժեք (maximum)

3.2.1 Վիճակագրական հաշվարկների արդյունքներ

Աղյուսակ 1. Հիվանդների տարիքի /age/ վիճակագրական բնութագիրչները

դիտարկումների (n)	1048575
միջին արժեք (mean)	41.79
մեդիան (median)	40
ստանդարտ շեղում	16.91
նվազագույն արժեք	0
առավելագույն արժեք	121

Բինար հատկանիշները հանդիսանում են կատեգորիական հատկանիշների հատուկ դեպք, որոնք ընդունում են միայն երկու արժեք՝ «այո» կամ «ոչ», մինչդեռ կատեգորիական հատկանիշները կարող են ունենալ երկուից ավելի տարբեր կատեգորիաներ։

Աղյուսակ 2. Բինար հատկանիշների բաշխում

pneumonia	դիտարկումների քանակ	հաճախականություն
1	140038	0.1336
2	892534	0.8512
99	16003	0.0153
diabetes	դիտարկումների քանակ	հաճախականություն
1	124989	0.1192
2	920248	0.8776
98	3338	0.0032
hypertension	դիտարկումների քանակ	հաճախականություն
1	162729	0.1552
2	882742	0.8418
98	3104	0.0030
obesity	դիտարկումների քանակ	հաճախականություն
1	159816	0.1524

2	885727	0.8447
98	3032	0.0029

Կատեգորիական հատկանիշների բաշխման ուսումնասիրություն

Կատեգորիական հատկանիշների դեպքում հաշվարկվել է յուրաքանչյուր կատեգորիայի դիտարկումների քանակը և դրանց հարաբերական հաճախականությունը: Որոշ կատեգորիաների միջև առկա է անհավասար բաշխում:

3.3 Լոգիստիկ ռեգրեսիայի մոդելի կառուցում և կիրառություն

Դիտարկվումէ բինար դասակարգման խնդիր, որի նպատակն է կանխատեսել COVID-19 հիվանդի ռիսկի մակարդակը՝ հիմվելով նրա բժշկական տվյալների վրա:

Որպես կախյալ (նպատակային) փոփոխական օգտագործվել է **target** բինար փոփոխականը, որը կառուցվել է **date_died** հատկանիշի հիման վրա, որտեղ

- ✓ 1 - բարձր ռիսկ (հիվանդը մահացել է),
- ✓ 0 - ցածր ռիսկ (հիվանդը չի մահացել):

Որպես անկախ փոփոխականներ օգտագործվել են հիվանդի առողջական վիճակը և բժշկական պատմությունը նկարագրող հատկանիշները, մասնավորապես՝ age, sex, pneumonia, diabetes, hypertension, cardiovascular, renal_chronic, copd, asthma, obesity, tobacco, inmsupr, other_disease, patient_type, intubed, icu, medical_unit, usmer:

Այսպիսով, մոդելի նպատակը գնահատելն է այն հավանականությունը, որ տվյալ հիվանդը պատկանում է բարձր ռիսկային խմբին՝ կախված նշված անկախ փոփոխականներից:

3.3.1 Տվյալների նախապատրաստում

Մեքենայական ուսուցման մոդելի կառուցումից առաջ անհրաժեշտ է նախապատրաստել տվյալները: Նախապատրաստման ընթացքում իրականացվել են հետևյալ քայլերը.

Առաջին հերթին մշակվել են բացակայող կամ անհայտ տվյալները։ Տվյալների հավաքածուում որոշ հատկանիշների դեպքում հանդիպում են 97, 98 և 99 արժեքներ, որոնք ներկայացնում են բացակայող կամ անհայտ տվյալներ։ Այսպիսի արժեքները փոխարինվել են բացակայող արժեքներով կամ հեռացվել են տվյալների մշակման ընթացքում։

Այնուհետև կառուցվել է նպատակային փոփոխականը։ Տվյալների հավաքածուում առկա **date_died** փոփոխականի հիման վրա ստեղծվել է նոր բինար փոփոխական, որը ցույց է տալիս հիվանդության ռիսկի մակարդակը։

- ✓ 1 - բարձր ռիսկ (հիվանդը մահացել է)
- ✓ 0 - ցածր ռիսկ (հիվանդը չի մահացել)

Այս փոփոխականը օգտագործվել է որպես նպատակային փոփոխական մոդելի ուսուցման համար։

3.3.2 Տվյալների բաժանում ուսուցման և փորձարկման խմբերի

Մոդելի արդյունավետությունը գնահատելու համար տվյալների հավաքածուն բաժանվել է երկու մասի.

- ✓ ուսուցման տվյալներ (training set)
- ✓ փորձարկման տվյալներ (test set)

Սովորաբար տվյալների մոտավորապես 80%-ը օգտագործվում է ուսուցման համար, իսկ 20%-ը՝ մոդելի փորձարկման համար։

3.3.3 Մոդելի գնահատման չափանիշներ

Տվյալների նախապատրաստումից հետո կիրառվել է Logistic Regression դասակարգման մեթոդը։

Լոգիստիկ ռեգրեսիայի մոդելի արդյունավետությունը գնահատվել է մի քանի հիմնական չափանիշների միջոցով.

- ✓ accuracy
- ✓ precision
- ✓ recall
- ✓ F1-score

Լոգիստիկ ռեգրեսիայի մոդելի աշխատանքի արդյունավետությունը գնահատելու համար օգտագործվում են մի քանի հիմնական չափանիշներ: Այս չափանիշները թույլ են տալիս գնահատել, թե որքան ճիշտ է մոդելը դասակարգում դիտարկումները:

Accuracy-ն ցույց է տալիս, թե մոդելը ընդհանուր դիտարկումների քանի տոկոսն է ճիշտ դասակարգել: Այն հաշվարկվում է հետևյալ բանաձևով.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

որտեղ

- ✓ TP (True Positive) - ճիշտ կանխատեսված դրական դեպքեր
- ✓ TN (True Negative) - ճիշտ կանխատեսված բացասական դեպքեր
- ✓ FP (False Positive) - սխալ դրական կանխատեսումներ
- ✓ FN (False Negative) - սխալ բացասական կանխատեսումներ

Այս չափանիշը ցույց է տալիս մոդելի ընդհանուր ճշգրտությունը:

Precision-ը ցույց է տալիս, թե մոդելի կողմից դրական կանխատեսված դեպքերից քանիսն են իրականում դրական:

$$Precision = \frac{TP}{TP + FP}$$

Այս չափանիշը կարևոր է այն դեպքերում, երբ սխալ դրական կանխատեսումները ցանկալի չեն:

Recall-ը (կամ զգայունություն) ցույց է տալիս, թե իրական դրական դեպքերից քանիսն է մոդելը ճիշտ հայտնաբերել:

$$Recall = \frac{TP}{TP + FN}$$

Այս չափանիշը կարևոր է հատկապես բժշկական խնդիրներում, քանի որ այն ցույց է տալիս, թե որքան լավ է մոդելը հայտնաբերում իրական դրական դեպքերը:

F1-score-ը հանդիսանում է precision և recall չափանիշների համակցված գնահատականը: Այն հաշվարկվում է հետևյալ բանաձևով.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Այս չափանիշը օգտագործվում է այն դեպքերում, երբ անհրաժեշտ է հավասարակշռված գնահատել մոդելի ճշգրտությունն ու զգայունությունը:

3.3.4 Լոգիստիկ ռեգրեսիայի մոդելի արդյունքների գնահատում

Լոգիստիկ ռեգրեսիայի մոդելի արդյունավետությունը գնահատելու համար տվյալների հավաքածուն բաժանվել է ուսուցման և փորձարկման մասերի: Ուսուցման տվյալների քանակը կազմել է 33,035 դիտարկում, իսկ փորձարկման տվյալների քանակը՝ 8,259 դիտարկում:

Accuracy

Մոդելի ընդհանուր ճշգրտությունը (accuracy) կազմել է.

$$Accuracy = 0.749$$

Սա նշանակում է, որ մոդելը ճիշտ է դասակարգել փորձարկման (test) տվյալների մոտավորապես 74.9%-ը:

Լոգիստիկ ռեգրեսիայի մոդելի գործակիցների վերլուծությունը ցույց է տալիս, թե յուրաքանչյուր անկախ փոփոխական ինչ ազդեցություն ունի հիվանդի բարձր ռիսկային խմբում գտնվելու հավանականության վրա:

Feature	Coefficient (β)	Odds Ratio
Patient_type	1.634586	5.127333

Sex	0.817293	2.264362
Pregnant	0.711148	2.036328
Icu	0.324760	1.383698
Tobacco	0.197698	1.218594
Asthma	0.167716	1.182601
Age	0.042488	1.043403
Cardiovascular	-0.006767	0.993256
Medical_unit	-0.026507	0.973841
Obesity	-0.092973	0.911218
Inmsupr	-0.110261	0.895601
Copd	-0.136571	0.872344
Hipertension	-0.162189	0.850281
Other_disease	-0.171213	0.842642
Diabetes	-0.252257	0.777045
Usmer	-0.271978	0.761871
Classification_final	-0.444844	0.640924
Renal_chronic	-0.655563	0.519150
Pneumonia	-0.860629	0.422896
Intubed	-2.538580	0.078978

Ամենամեծ դրական ազդեցություն ունեցող հատկանիշներից է **patient_type** փոփոխականը ($\beta = 1.63$, OR ≈ 5.13), ինչը ցույց է տալիս, որ հոսպիտալացված հիվանդների մոտ բարձր ռիսկի հավանականությունը զգալիորեն մեծ է: Նմանապես, **sex** ($\beta \approx 0.82$, OR ≈ 2.26) և **pregnant** ($\beta \approx 0.71$, OR ≈ 2.04) փոփոխականները նույնպես մեծացնում են բարձր ռիսկի հավանականությունը:

Բացի այդ, հիվանդության ծանր ընթացքը բնութագրող որոշ հատկանիշներ, ինչպիսիք են **icu** ($\beta \approx 0.32$) և **tobacco** ($\beta \approx 0.20$), ունեն դրական ազդեցություն, ինչը տրամաբանական է, քանի որ այս գործոնները հաճախ կապված են ավելի ծանր կլինիկական վիճակի հետ: Նույնպես նկատվում է, որ **age** փոփոխականը ունի դրական

գործակից (β ≈ 0.04), ինչը նշանակում է, որ տարիքի աճի հետ մեծանում է բարձր ռիսկի հավանականությունը:

Միևնույն ժամանակ, որոշ հատկանիշներ ունեն բացասական գործակիցներ: Օրինակ՝ **intubed** (β ≈ -2.54, OR ≈ 0.08) և **pneumonia** (β ≈ -0.86, OR ≈ 0.42) փոփոխականները ցույց են տալիս հակառակ ազդեցություն: Նման արդյունքները կարող են պայմանավորված լինել տվյալների բաշխմամբ, հատկանիշների փոխկապակցվածությամբ կամ տվյալների նախնական մշակման առանձնահատկություններով:

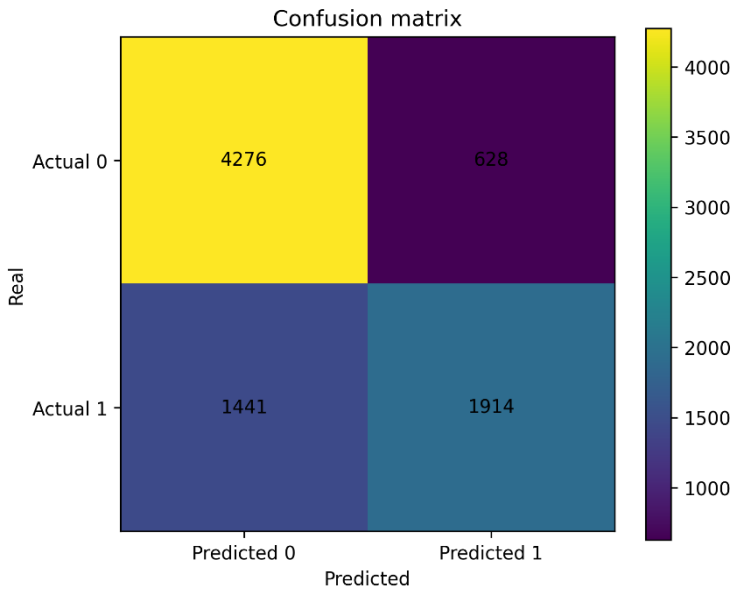
Ընդհանուր առմամբ, մոդելի արդյունքները ցույց են տալիս, որ հիվանդի վիճակը բնութագրող որոշ գործոններ (օրինակ՝ հոսպիտալացում, ինտենսիվ թերապիա, տարիք) ունեն առավել նշանակալի ազդեցություն ռիսկի մակարդակի վրա, մինչդեռ մյուս փոփոխականների ազդեցությունը կարող է լինել ավելի թույլ կամ կախված տվյալների կառուցվածքից:

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	4276	628
Actual 1	1441	1914

Այս աղյուսակից կարելի է տեսնել, որ

- 4276 դեպքերում մոդելը ճիշտ է կանխատեսել ցածր ռիսկը
- 1914 դեպքերում մոդելը ճիշտ է կանխատեսել բարձր ռիսկը
- 628 դեպքերում ցածր ռիսկը սխալ դասակարգվել է որպես բարձր ռիսկ
- 1441 դեպքերում բարձր ռիսկը սխալ դասակարգվել է որպես ցածր ռիսկ



Precision

Բարձր ռիսկի համար precision-ը կազմել է.

$$Precision = 0.75$$

Սա նշանակում է, որ մոդելի կողմից բարձր ռիսկ դասակարգված դեպքերի մոտ 75%-ը իրականում բարձր ռիսկային են:

Recall

$$Recall = 0.57$$

Սա նշանակում է, որ մոդելը հայտնաբերել է բարձր ռիսկային հիվանդների մոտավորապես 57%-ը:

F1-score

Բարձր ռիսկի համար

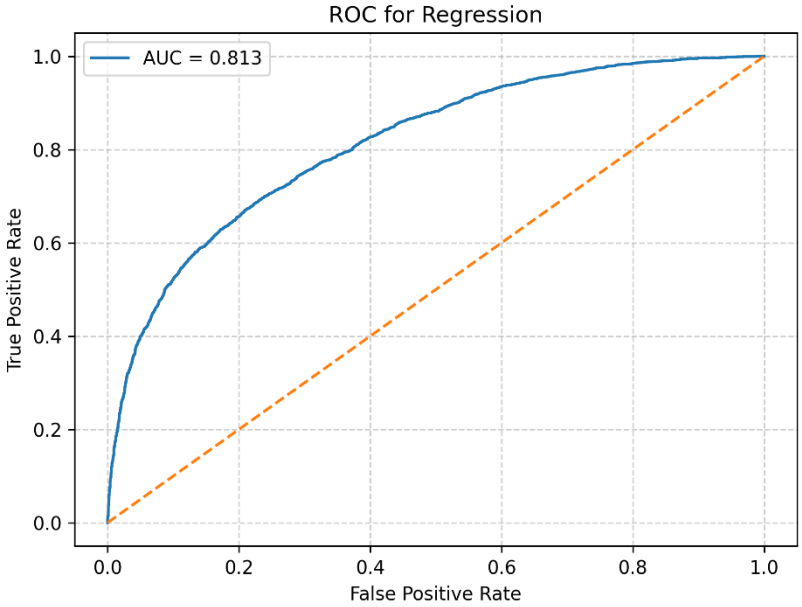
$$F1 = 0.65$$

Սա ցույց է տալիս, որ մոդելը միջին մակարդակով է կատարում բարձր ռիսկային հիվանդների դասակարգումը:

ROC Curve

ROC կորը ցույց է տալիս լոգիստիկ ռեգրեսիայի մոդելի տարբերակիչ կարողությունը:

$$AUC = 0.813$$



Ստացված AUC = 0.813 արժեքը ցույց է տալիս, որ մոդելը բավականին լավ է տարբերակում բարձր և ցածր ռիսկային դեպքերը ROC կորի համար ստացվել է

Արդյունքները ցույց են տալիս, որ լոգիստիկ ռեգրեսիայի մոդելը բավականին լավ է աշխատում ընդհանուր ճշգրտության տեսանկյունից: Սակայն recall արժեքը ցույց է տալիս, որ մոդելը չի հայտնաբերում բարձր ռիսկային բոլոր հիվանդներին: Սա կարևոր է հատկապես բժշկական խնդիրներում, քանի որ բարձր ռիսկային դեպքերի բաց թողնելը կարող է ունենալ լուրջ հետևանքներ:

Այս խնդիրը կարող է պայմանավորված լինել մի քանի գործոններով: Նախ, տվյալների հավաքածուում կարող է առկա լինել դասերի անհավասարակշռություն (class imbalance), երբ բարձր ռիսկային դեպքերը ներկայացված են ավելի փոքր քանակով, ինչի արդյունքում մոդելը ավելի լավ է սովորում ցածր ռիսկային դեպքերը: Բացի այդ, որոշ կարևոր հատկանիշներ կարող են բավարար չափով չարտացոլել հիվանդության ծանրության աստիճանը կամ ունենալ փոխկապակցվածություն միմյանց հետ, ինչը դժվարացնում է մոդելի կողմից ճիշտ սահմանազատումը:

Խնդրի լուծման համար կարելի է փոխել դասակարգման շեմը (decision threshold)՝ բարձրացնելով recall-ը՝ նույնիսկ precision-ի որոշ նվազման հաշվին: Կարելի է նաև կիրառել տվյալների վերակշռման մեթոդներ (օրինակ՝ oversampling կամ undersampling), ինչպես նաև օգտագործել կշռված մոդելներ (class weights), որոնք ավելի մեծ ուշադրություն են դարձնում բարձր ռիսկային դասին: Բացի այդ, հնարավոր է փորձարկել այլ մոդելներ կամ իրականացնել հատկանիշների ընտրություն և ինժեներիա՝ մոդելի որակը բարելավելու համար:

3.4 K-Nearest Neighbors մոդելի կառուցում և կիրառություն

3.4.1 Տվյալների նախապատրաստում

KNN մոդելի կառուցումից առաջ անհրաժեշտ է իրականացնել տվյալների նախապատրաստում: Այս փուլում օգտագործվել է նույն տվյալների հավաքածուն, որն օգտագործվել էր լոգիստիկ ռեգրեսիայի դեպքում, որպեսզի հնարավոր լինի հետագայում համեմատել երկու մեթոդների արդյունքները:

Նախ տվյալների հավաքածուում մշակվել են բացակայող կամ անհայտ տվյալները: Որոշ հատկանիշների դեպքում հանդիպող 97, 98 և 99 արժեքները ներկայացնում են բացակայող կամ անհայտ տվյալներ, ուստի դրանք փոխարինվել են բացակայող արժեքներով կամ հեռացվել հետագա վերլուծությունից:

Այնուհետև կառուցվել է նպատակային փոփոխականը՝ հիմնվելով date_died հատկանիշի վրա:

- ✓ 1 - բարձր ռիսկ, եթե հիվանդը մահացել է,
- ✓ 0 - ցածր ռիսկ, եթե հիվանդը չի մահացել:

KNN մեթոդում առանձնահատուկ կարևորություն ունի նաև հատկանիշների մասշտաբը, քանի որ մեթոդը հիմված է հեռավորությունների հաշվարկի վրա: Եթե տարբեր հատկանիշներ ունեն տարբեր միջակայքներ, ապա մեծ միջակայք ունեցող հատկանիշները կարող են անհամաչափ ազդեցություն ունենալ հեռավորության վրա: Այդ պատճառով տվյալները նախքան մոդելի կիրառումը նորմալացվել կամ ստանդարտացվել են:

3.4.2 Տվյալների բաժանում ուսուցման և փորձարկման խմբերի

Ինչպես լոգիստիկ ռեգրեսիայի դեպքում, KNN մոդելի համար ևս տվյալների հավաքածուն բաժանվել է երկու մասի՝

- ուսուցման տվյալներ (training set),
- փորձարկման տվյալներ (test set)։

Տվյալների մոտավորապես 80%-ը օգտագործվել է ուսուցման համար, իսկ 20%-ը՝ մոդելի փորձարկման համար։ Այս մոտեցումը հնարավորություն է տալիս գնահատել մոդելի աշխատանքը ոչ միայն ուսուցման տվյալների, այլ նաև նոր, նախկինում չօգտագործված դիտարկումների վրա։

3.4.3 *k*պարամետրի ընտրություն

KNN մեթոդում կարևոր դեր ունի *k*պարամետրը, որը ցույց է տալիս, թե քանի հարևան պետք է հաշվի առնել նոր դիտարկման դասը որոշելու համար։

Այս աշխատանքում ընտրվել է

$$k = 5$$

արժեքը։

3.4.4 KNN մոդելի կիրառություն

Տվյալների նախապատրաստումից հետո կիրառվել է K-Nearest Neighbors դասակարգման մեթոդը։ Մոդելի ուսուցման ընթացքում ուսուցման տվյալների յուրաքանչյուր դիտարկում պահպանվում է, և նոր դիտարկման դեպքում հաշվարկվում են նրա հեռավորությունները ուսուցման տվյալների բոլոր դիտարկումներից։

Այնուհետև ընտրվում են ամենամոտ 5 դիտարկումները, և նոր դիտարկման դասը որոշվում է դրանց մեծամասնության սկզբունքով։

Այսպիսով KNN մոդելը տվյալ հիվանդին դասակարգում է որպես բարձր կամ ցածր ռիսկային՝ հիմնվելով նրան առավել նման հիվանդների տվյալների վրա։

3.4.5 Մոդելի գնահատման չափանիշներ

KNN մոդելի աշխատանքի արդյունավետությունը գնահատելու համար օգտագործվել են նույն չափանիշները, որոնք կիրառվել են լոգիստիկ ռեգրեսիայի դեպքում, որպեսզի հնարավոր լինի իրականացնել ուղիղ համեմատություն:

Օգտագործվել են հետևյալ չափանիշները.

- ✓ accuracy
- ✓ precision
- ✓ recall
- ✓ F1-score

3.4.6 KNN մոդելի արդյունքների ներկայացում

K-Nearest Neighbors մոդելի արդյունավետությունը գնահատելու համար տվյալների հավաքածուն, ինչպես լոգիստիկ ռեգրեսիայի դեպքում, բաժանվել է ուսուցման և փորձարկման մասերի: Ուսուցման տվյալների քանակը կազմել է 33,035 դիտարկում, իսկ փորձարկման տվյալների քանակը՝ 8,259 դիտարկում:

Accuracy

KNN մոդելի ընդհանուր ճշգրտությունը կազմել է.

$$Accuracy = 0.724$$

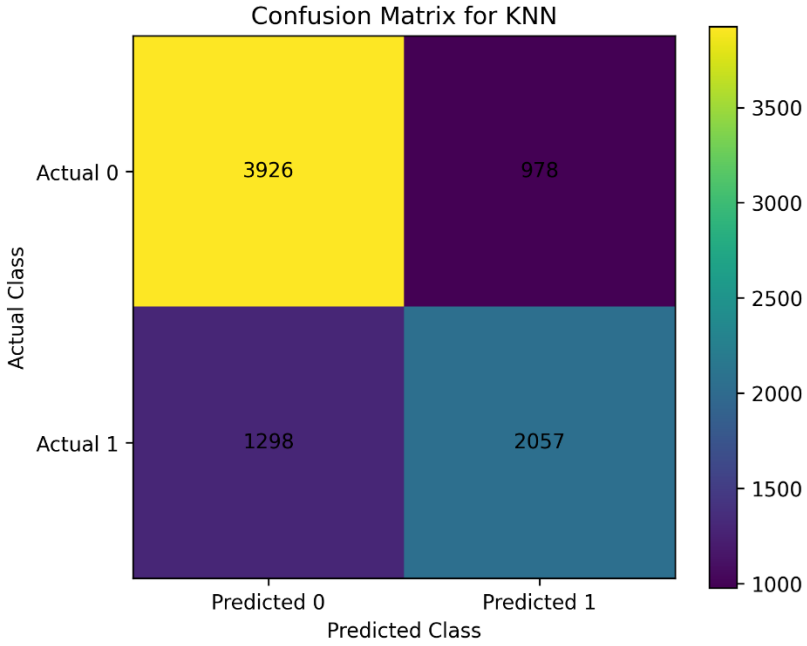
Սա նշանակում է, որ մոդելը ճիշտ է դասակարգել փորձարկման տվյալների մոտավորապես 72.4%-ը:

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	3926	978
Actual 1	1298	2057

Այս աղյուսակից երևում է, որ

- 3926 դեպքերում մոդելը ճիշտ է կանխատեսել ցածր ռիսկը,
- 2057 դեպքերում մոդելը ճիշտ է կանխատեսել բարձր ռիսկը,
- 978 դեպքերում ցածր ռիսկը սխալ դասակարգվել է որպես բարձր ռիսկ,
- 1298 դեպքերում բարձր ռիսկը սխալ դասակարգվել է որպես ցածր ռիսկ։



Precision

Բարձր ռիսկի համար precision-ը կազմել է.

$$Precision = 0.68$$

Սա նշանակում է, որ մոդելի կողմից բարձր ռիսկ դասակարգված դեպքերի մոտ 68%-ը իրականում բարձր ռիսկային են։

Recall

Բարձր ռիսկի համար recall-ը կազմել է.

$$Recall = 0.61$$

Սա նշանակում է, որ մոդելը ճիշտ է հայտնաբերել բարձր ռիսկային հիվանդների մոտավորապես **61%**-ը:

F1-score

Բարձր ռիսկի համար F1-score-ը կազմել է.

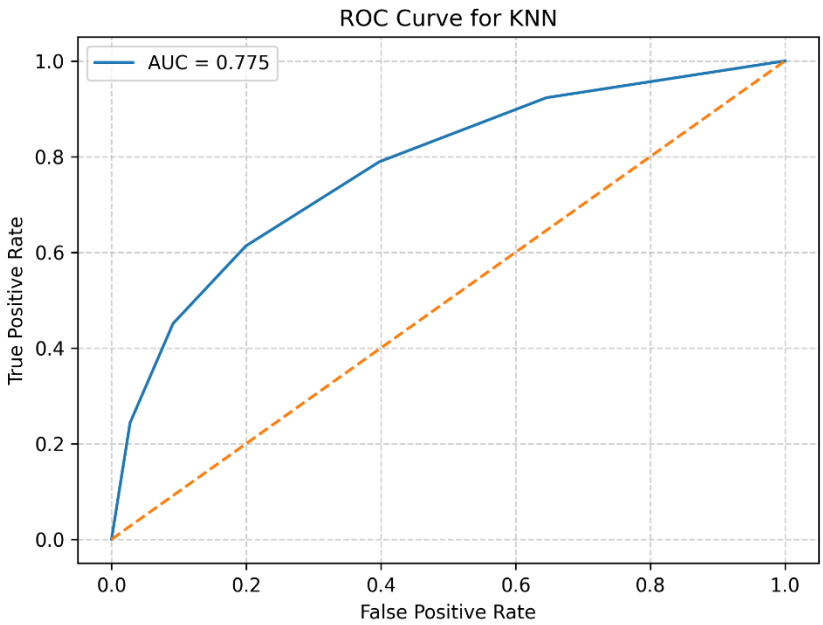
$$F1 = 0.64$$

Սա ցույց է տալիս, որ մոդելը միջին մակարդակով է կատարում բարձր ռիսկային հիվանդների դասակարգումը:

ROC Curve

KNN մոդելի ROC կորի համար ստացվել է

$$AUC = 0.775$$



Այս արժեքը ցույց է տալիս, որ մոդելը ունի բավարար տարբերակիչ կարողություն, սակայն այն մի փոքր ցածր է լոգիստիկ ռեգրեսիայի մոդելի համապատասխան ցուցանիշից:

3.5 Naive Bayes մոդելի կառուցում և կիրառություն

3.5.1 Տվյալների նախապատրաստում

Naive Bayes դասակարգման մեթոդի կիրառման համար օգտագործվել է նույն տվյալների հավաքածուն, ինչ լոգիստիկ ռեգրեսիայի և K-Nearest Neighbors մոդելների դեպքում, որպեսզի հնարավոր լինի իրականացնել արդյունքների ուղիղ համեմատություն։
Նախ տվյալների հավաքածուում մշակվել են բացակայող կամ անհայտ տվյալները։ Ինչպես նախորդ մոդելների դեպքում, 97, 98 և 99 արժեքները դիտարկվել են որպես բացակայող կամ անհայտ տվյալներ և փոխարինվել են բացակայող արժեքներով կամ հեռացվել տվյալների մշակման ընթացքում։

Այնուհետև կառուցվել է նպատակային փոփոխականը՝ հիմվելով date_died հատկանիշի վրա.

- ✓ 1 – բարձր ռիսկ, եթե հիվանդը մահացել է,
- ✓ 0 – ցածր ռիսկ, եթե հիվանդը չի մահացել։

Naive Bayes մեթոդը հիմված է հավանականությունների տեսության վրա և ենթադրում է, որ հատկանիշները պայմանականորեն անկախ են միմյանցից՝ տվյալ դասի պայմաններում։ Թեև իրական տվյալներում այս ենթադրությունը միշտ չէ, որ ամբողջությամբ կատարվում է, մեթոդը հաճախ ապահովում է բավականին լավ արդյունքներ և հատկապես արդյունավետ է որպես պարզ ու արագ բազային դասակարգիչ։

3.5.2 Տվյալների բաժանում ուսուցման և փորձարկման խմբերի

- ✓ Ինչպես նախորդ մոդելների դեպքում, տվյալների հավաքածուն բաժանվել է երկու մասի՝
- ✓ ուսուցման տվյալներ (training set),
- ✓ փորձարկման տվյալներ (test set)։ `

Տվյալների մոտավորապես 80%-ը օգտագործվել է ուսուցման համար, իսկ 20%-ը՝ փորձարկման համար։ Այս մոտեցումը թույլ է տալիս գնահատել մոդելի աշխատանքը նոր դիտարկումների վրա և ստանալ համեմատելի արդյունքներ մյուս մեթոդների հետ։

3.6.3 Naive Bayes մոդելի կիրառություն

Տվյալների նախապատրաստումից հետո կիրառվել է Naive Bayes դասակարգման մեթոդը։ Այս մեթոդը յուրաքանչյուր նոր դիտարկման համար հաշվարկում է, թե այն ինչ հավանականությամբ է պատկանում յուրաքանչյուր դասին՝ հիմնվելով հատկանիշների արժեքների վրա։
Այնուհետև դիտարկումը վերագրվում է այն դասին, որի հավանականությունը ավելի մեծ է։

Այս աշխատանքում Naive Bayes մոդելի նպատակն է եղել կանխատեսել, թե արդյոք տվյալ COVID-19 հիվանդը գտնվում է բարձր ռիսկային խմբում, թե ոչ՝ հիմնվելով նրա բժշկական և կլինիկական հատկանիշների վրա։

3.6.4 Մոդելի գնահատման չափանիշներ

Naive Bayes մոդելի աշխատանքի արդյունավետությունը գնահատելու համար օգտագործվել են նույն չափանիշները, որոնք կիրառվել են լոգիստիկ ռեգրեսիայի և KNN մոդելների դեպքում.

- ✓ accuracy
- ✓ precision
- ✓ recall
- ✓ F1-score
- ✓ ROC AUC

Այս չափանիշները հնարավորություն են տալիս գնահատել, թե որքան ճիշտ է մոդելը դասակարգում բարձր և ցածր ռիսկային դեպքերը, ինչպես նաև համեմատել այն մյուս մոդելների արդյունքների հետ։

3.6.5 Naive Bayes մոդելի արդյունքների ներկայացում

Naive Bayes մոդելի արդյունավետությունը գնահատելու համար այն ուսուցանվել է ուսուցման տվյալների վրա և փորձարկվել փորձարկման տվյալների վրա։
Ստացվել են accuracy, precision, recall, F1-score և ROC AUC ցուցանիշները, ինչպես նաև confusion matrix-ը։

Accuracy

Naive Bayes մոդելի ընդհանուր ճշգրտությունը կազմել է.

$$Accuracy = 0.899$$

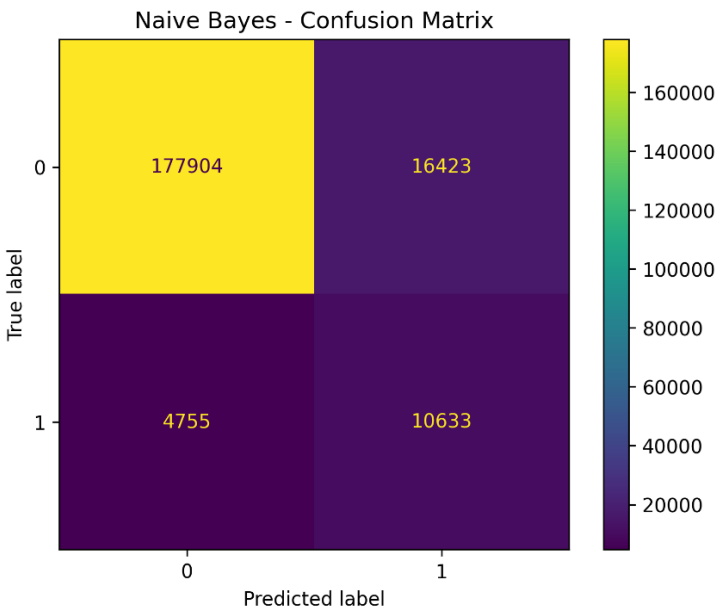
Սա նշանակում է, որ մոդելը ճիշտ է դասակարգել փորձարկման տվյալների մոտավորապես 89.9%-ը։

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	177904	16423
Actual 1	4755	10633

Այս աղյուսակից երևում է, որ

- 177904 դեպքերում մոդելը ճիշտ է կանխատեսել ցածր ռիսկը,
- 10633 դեպքերում մոդելը ճիշտ է կանխատեսել բարձր ռիսկը,
- 16423 դեպքերում ցածր ռիսկը սխալ դասակարգվել է որպես բարձր ռիսկ,
- 4755 դեպքերում բարձր ռիսկը սխալ դասակարգվել է որպես ցածր ռիսկ։



Precision

Բարձր ռիսկի համար precision-ը կազմել է.

$$Precision = 0.393$$

Սա նշանակում է, որ մոդելի կողմից բարձր ռիսկ դասակարգված դեպքերի մոտ 39.3%-ը իրականում բարձր ռիսկային են:

Recall

Բարձր ռիսկի համար recall-ը կազմել է.

$$Recall = 0.691$$

Սա նշանակում է, որ մոդելը ճիշտ է հայտնաբերել բարձր ռիսկային հիվանդների մոտավորապես 69.1%-ը:

F1-score

Բարձր ռիսկի համար F1-score-ը կազմել է.

$$F1 = 0.501$$

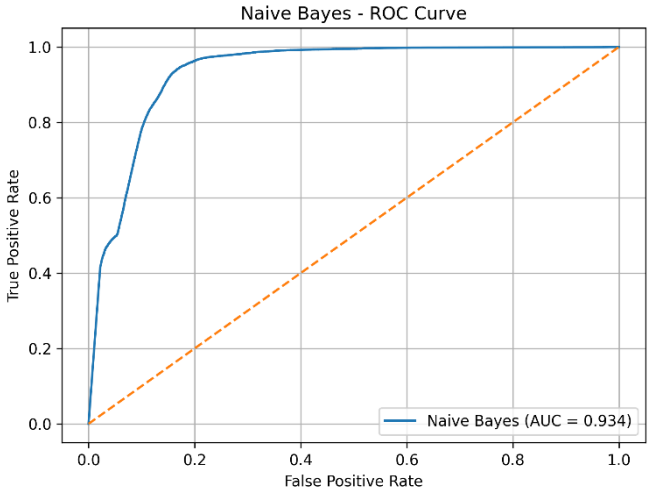
Սա ցույց է տալիս, որ մոդելը միջին մակարդակով է կատարում բարձր ռիսկային հիվանդների դասակարգումը:

ROC Curve

Naïve Bayes մոդելի ROC կորի համար ստացվել է

$$AUC = 0.934$$

Այս արժեքը ցույց է տալիս, որ մոդելը ունի բավականին բարձր տարբերակիչ կարողություն:



3.7 Մոդելների համեմատական վերլուծություն

Այս բաժնում համեմատվում են լոգիստիկ ռեգրեսիայի և K-Nearest Neighbors մեթոդների արդյունքները՝ գնահատելու համար, թե որ մոդելն է ավելի արդյունավետ տվյալ խնդրի լուծման համար:

Մոդել	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.749	0.75	0.57	0.65	0.813
KNN	0.724	0.68	0.61	0.64	0.775
Naïve Bayes	0.899	0.393	0.691	0.501	0.934

Աղյուսակից երևում է, որ Naive Bayes մոդելը ապահովել է ամենաբարձր accuracy (0.899) և AUC (0.934) արժեքները, ինչը ցույց է տալիս նրա բարձր ընդհանուր ճշգրտությունն ու տարբերակիչ կարողությունը: Սակայն precision-ի ցածր արժեքը (0.393) ցույց է տալիս, որ մոդելը հաճախ սխալ կերպով դասակարգում է ցածր ռիսկային հիվանդներին որպես բարձր ռիսկային:

Լոգիստիկ ռեգրեսիայի մոդելը ցուցաբերել է ավելի հավասարակշռված արդյունքներ՝ համեմատաբար բարձր precision և F1-score արժեքներով, ինչը նշանակում է, որ այն ավելի կայուն է բարձր ռիսկի կանխատեսման տեսանկյունից։

KNN մոդելը ցուցաբերել է միջին արդյունքներ՝ մի փոքր ավելի բարձր recall-ով, ինչը նշանակում է, որ այն որոշ չափով ավելի լավ է հայտնաբերում բարձր ռիսկային հիվանդներին, սակայն ընդհանուր ճշգրտությամբ զիջում է մյուս մոդելներին։

Եզրակացություն

Այս աշխատանքում ուսումնասիրվել է մեքենայական ուսուցման դասակարգման մեթոդների կիրառությունը COVID-19 հիվանդների ռիսկի մակարդակի կանխատեսման համար: Օգտագործվել են Logistic Regression, K-Nearest Neighbors և Naive Bayes մոդելները, որոնց միջոցով իրականացվել է տվյալների վերլուծություն և դասակարգում:

Ստացված արդյունքները ցույց են տվել, որ տարբեր մոդելներ ունեն տարբեր առավելություններ: Naive Bayes մոդելը ապահովել է բարձր accuracy և AUC ցուցանիշներ, սակայն ունեցել է ցածր precision, ինչը նշանակում է, որ այն հաճախ սխալ կերպով դասակարգում է ցածր ռիսկային հիվանդներին որպես բարձր ռիսկային:

Լոգիստիկ ռեգրեսիայի մոդելը ցուցաբերել է ավելի հավասարակշռված արդյունքներ՝ համեմատաբար բարձր precision և F1-score արժեքներով, ինչը այն դարձնում է ավելի հուսալի բարձր ռիսկային հիվանդների հայտնաբերման տեսանկյունից:

KNN մոդելը ցուցաբերել է միջին արդյունավետություն և կարող է դիտարկվել որպես լրացուցիչ մեթոդ:

Ընդհանուր առմամբ, կարելի է եզրակացնել, որ տվյալ խնդրի համար առավել կիրառելի է Logistic Regression մոդելը, քանի որ այն ապահովում է արդյունքների լավ հավասարակշռություն:

Գրականության ցանկ

- Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly, 2019
- Scikit-learn Documentation, https://scikit-learn.org
- Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press
- Kaggle Dataset – COVID-19 Data
- Tom Mitchell, Machine Learning, McGraw-Hill