УНИВЕРЗИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“ – БИТОЛА
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ
И КОМУНИКАЦИСКИ ТЕХНОЛОГИИ -БИТОЛА

Студиска програма: Информатички науки и компјутерско инженерство

ТЕХНИКИ ЗА МАШИНСКО УЧЕЊЕ И РЕСЕМПЛИРАЊЕ ЗА ПРЕДИКЦИЈА ОД НЕБАЛАНСИРАНИ МЕДИЦИНСКИ ПОДАТОЦИ

докторски проект

Кандидат
м-р Нора Пиреци Сејдиу
бр. Индекс:11/20/III

Ментор
ред. проф. д-р Благој Ристевски

Битола
септември, 2024 година

СОДРЖИНА

ТЕХНИКИ ЗА МАШИНСКО УЧЕЊЕ И РЕСЕМПЛИРАЊЕ ЗА ПРЕДИКЦИЈА ОД НЕБАЛАНСИРАНИ МЕДИЦИНСКИ ПОДАТОЦИ

Нора Пиреци Сејдиу
Универзитет „Св. Климент Охридски“ – Битола
0000-0002-7195-2266
pireci.nora@uklo.edu.mk

Благој Ристевски
Универзитет „Св. Климент Охридски“ – Битола
0000-0002-8356-1203
blagoj.ristevski@uklo.edu.mk

Апстракт

Доминацијата на дијабетесот бара точни модели на предвидување, посебно кога се работи со множества податоци кои покажуваат небалансирана распределба на класите. Оваа студија ги истражува перформансите на два класификатори за машинско учење, XGBoost и Balanced Random Forest, комбинирано со две напредни техники за ресемплирање SMOTEtomek и ADASYN, за предвидување на дијабетес со користење на високо небалансирано множество податоци за здравствени индикатори за дијабетес. Множеството податоци, со сооднос на неизбалансираност на класите од приближно 6:1, постави предизвици во прецизното идентификување на малцинската класа (случаи со дијабетес). Целта на истражувањето беше да се евалуира како различните техники на ресемплирање и класификатори влијаат на перформансите на моделот во смисла на точност, прецизност, recall и F1-резултатот. Студијата користи четири различни конфигурации на модели: SMOTEtomek со XGBoost, SMOTEtomek со Balanced Random Forest, ADASYN со XGBoost и ADASYN со Balanced Random Forest. Резултатите покажаа дека моделот SMOTEtomek со Balanced Random Forest ги надмина сите други модели, постигнувајќи точност од 89,47% и F1-резултат од 0,91 за малцинската класа, ефикасно балансирајќи ја прецизноста и recall. Додека, XGBoost, и покрај неговата севкупна ефикасност, покажа слаба прецизност во предвидувањето на инстанците од малцинската класа, што доведе до повисока стапка на лажни позитивни инстанци (false positive). Студијата покажа дека комбинирањето на SMOTEtomek со Balanced Random Forest го нуди најробусното решение за предвидување дијабетес од небалансирани множества податоци . Идната работа ќе ја истражува примената на длабокото учење и невронските мрежи за понатамошно подобрување на перформансите на моделот, особено во препознавање на комплексни облици (patterns) во податоците поврзани со здравството.

Клучни зборови: XGBoost класификатор, Balanced Random Forest класификатор, SMOTETomek, ADASYN, небалансирани множества податоци.

1. Вовед

Брзиот напредок во машинското учење (ML) ја револуционизираше предиктивната анализа во различни области, посебно во здравствената заштита. Сепак, значаен предизвик кој опстојува е прашањето на небалансираните множества податоци, каде што едната класа е недоволно помалку застапена во споредба со другата. Оваа небалансираност влијае на перформансите на алгоритмите за машинско учење, честопати доведувајќи до пристрасни предикции во корист на мнозинската класа, со што се занемаруваат критичните малцински случаи (инстанци). Во медицината како што е предвидувањето на дијабетес, каде што точната идентификација на пациентите изложени на ризик е најважна, оваа небалансираност предизвикува сериозна загриженост за веродостојноста и точноста на предвидливите модели (Abushahla & Pala, 2024; Kumar et al., 2022; Mohammed et al., 2020).

Решавањето на неизбалансираноста на класите кај машинското учење беше во фокусот на бројни студии. Abushahla и Pala (2024) применила различни ML алгоритми како што се XGBoost, Random Forest и CatBoost во комбинација со техники за ресемплирање (повторно земање на примероци) како SMOTE за да се подобри предвидувањето (предикцијата) на дијабетес. Други студии, како што се тие на Alkharabsheh et al. (2022), истражуваше како различни алгоритми за машинско учење се однесуваат кај небалансираните множества податоци во контекст на дизајн на софтвер за детекција на миризби. Овие студии ја нагласуваат важноста на решавање на неизбалансираност ана класите за да се подобри способноста за предвидување на моделите во различни домени, вклучително и во медицината, здравствената заштита и софтверското инженерство.

Во делот посветен на методологијата, се опишува експерименталниот пристап, почнувајќи со претходна обработка на податоците, каде што беа применети техники за ресемплирање (SMOTEtomek и ADASYN) за да се избалансира множеството податоци. Перформансите на секоја техника на ресемплирање се евалуираат со помош на класификатори на XGBoost и Balanced Random Forest. За мерење на ефективноста на моделите се користат стандардни метрики за евалуација како што се точност, прецизност, recall и f1-score, со посебен акцент на детекција на малцинската класа. Множеството податоци, кое вклучува над 200.000 записи со значителна неизбалансираност меѓу дијабетес позитивните и негативните инстанци, претставува релевантен предизвик за стандардните алгоритми. Целта е да се процени колку добро техниките за ресемплирање го подобруваат предвидувањето на малцинската класа (дијабетес-позитивни) додека се одржува целокупната точност на моделот.

Конечно, во делот за анализата, резултатите и дискусијата, презентираме компаративна анализа на четирите различни комбинации на модели и техники за ресемплирање. Резултатите го демонстрираат влијанието на техниките за ресемплирање врз перформансите на моделот, посебно во подобрување на детекција на малцинските класи.

Овој докторски проект обезбедува сеопфатно разбирање за улогата на техниките за ресемплирање во решавањето на дисбалансот на класите, нагласувајќи ја нивната важност за подобрување на предиктивната точност во критичните примени во здравството како што е предвидувањето на дијабетес. Со споредување на перформансите на различни модели за ресемплирање и класификација, целта е да се обезбеди увид во најефективните методи за справување на небалансираните множества податоци.

2. Досегашна работа

Во последниве години, решавањето на предизвикот за небалансирани податоци стана критично во примената на алгоритмите за машинско учење, посебно во областите како што се здравството и софтверското инженерство. Небалансираните множества податоци, каде што едната класа значително ја надминува другата, често доведува до пристрасни предвидувања и намалена ефективност на моделот. Ова прашање е распространето во медицинската дијагностика, каде што идентификувањето на ретки состојби како што е дијабетесот или откривањето аномалии во многу небалансирани множества е од клучно значење за веродостојноста на моделот. Истражувачите се фокусираа на развивање и подобрување на техниките за ефикасно справување со овие дисабалнси, користејќи различни методи на ресемплирање и напредни алгоритми.

Трудот „Optimizing Diabetes Prediction: Addressing Data Imbalance with Machine Learning Algorithms" од Abushahla и Pala (2024) се фокусира на подобрување на моделите за предикција на дијабетес преку решавање на дисбалансот на класите во податоците. Авторите имплементираат различни алгоритми за машинско учење, како што се логистичка регресија, стебла на одлучување, случајна шума, Gradient Boosting, SVM, KNN, наивен Бајес, XGBoost, LightGBM и CatBoost во комбинација со техники за ресемплирање како SMOTE за да се избалансира множеството податоци. Нивниот пристап има за цел да ја подобри точноста и веродостојноста на предвидувањата за дијабетес преку ефикасно справување со дисбалансот меѓу позитивните и негативните случаи (инстанци).

Трудот "A Comparison of Machine Learning Algorithms on Design Smell Detection Using Balanced and Imbalanced Dataset: A Study of God Class" од Alkharabsheh et al. (2022) оценува различни алгоритми за машинско учење за детекција на дизајна на кластата за миризби во софтверот. Студијата ги споредува перформансите на алгоритмот на избалансирани наспроти небалансирани множества податоци, откривајќи како балансот на податоците влијае на точноста на откривањето.

Трудот „A New Concordant Partial AUC and Partial C Statistics for Dimbalanced Data in the Evaluation of Machine Learning Algorithms“ од Carrington et al. (2020) воведува нови

метрики за евалуација на моделите за машинско учење на небалансирани множества податоци. Предложената усогласена парцијална AUC и делумна C статистика имаат за цел да обезбедат попрецизна проценка на перформансите на моделот во ситуации каде што преовладува дисбалансот на класите.

Трудот "A Novel Riboswitch Classification Based on Imbalanced Sequences Achieved by Machine Learning" од Beyen et al. (2020) воведува напредни техники за класификација на Riboswitches во небалансирана низа податоци. Авторите користат методи на прекумерно и недоволно семплирање (oversampling и undersampling) за да се балансираат множествата податоци, избор на одличја за да се идентификуваат клучните атрибути, техники за учење на ансамблот како случајна шума и методи за длабоко учење како што се конволуциските невронски мрежи (CNN) за да се подобри точноста на класификацијата.

Трудот „A resampling Method for Disbalanced Datasets Considering Noise and Overlap“ од Sasada et al. (2020) претставува нова техника за ресемплирање дизајнирана да ги подобри перформансите на моделот на небалансирани множества податоци преку решавање на проблемите со шумот и преклопувањето на класите. Авторите предлагаат метод кој комбинира филтрирање на шумот и намалување на преклопувањето за да се подобри квалитетот на ресемплираните податоци, со цел да се обезбедат попрецизни и понадежни (поверодостојни) резултати од класификација.

Трудот "A New Hybrid Predictive Model to Predict the Early Mortality Risk in Intensive Care Units on a Highly Imbalanced Dataset" од Горбани et al. (2020) воведува хибриден модел за предвидување на ризикот од рана смртност кај пациентите користејќи високо неизбалансирано множество податоци. Авторите комбинираат неколку техники за да ја подобрат прецизноста на предвидувањето, вклучително и методите за учење на ансамблот, како што се случајна шума и Gradient Boosting, и напредни техники за ресемплирање како што е SMOTE за да се реши дисбалансот на класите. Дополнително, тие го интегрираат изборот на карактеристики за да ги подобрат перформансите на моделот со фокусирање на најрелевантните модели за предикција.

Трудот "Boosting Methods for Multi-Class Imbalanced Data Classification: An Experimental Review" од Tanha et al. (2020) обезбедува сеопфатен преглед на техниките за подобрување на класификација на небалансирани множества податоци со дисбаланс меѓу повеќе класи. Авторите оценуваат различни алгоритми за засилување (boosting), вклучително AdaBoost и Gradient Boosting, за да ја проценат нивната ефикасност во справувањето со дисбаланост на класите во случај на повеќе-класна класификација. Тука се нагласува како овие методи ги подобруваат перформансите на класификацијата со фокусирање на тешко предвидливи примери и балансирање на застапеноста на класите.

Трудот "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques" од Ghorbani и Ghouzi (2020) испитува различни техники на ресемплирање за да ги подобри предвидувањата за перформансите (успехот) на учениците. Авторите ги споредуваат методите како што се SMOTE и random undersampling за да се реши дисбалансот меѓу класите во образовните множества податоци. Тие ги

применуваат овие стратегии за ресемплирање на различни алгоритми за машинско учење, вклучително и дрва на одлучување и машини со носечки вектори (SVMs), за да го оценат нивното влијание врз точноста на предвидувањето.

Трудот Significance of Machine Learning for Detection of Malicious Websites on an Unbalanced Dataset" од Ul Hassan et al. (2022) ја истражува техниката на машинско учење за детекција на малициозни веб страници со користење на множества податоци со значителен дисбаланс меѓу класите. Авторите применуваат техники за ресемплирање како што е SMOTE за справување со дисбаланост на податоците и подобрување на перформансите на моделот.

3. Методи и материјали

Различни методи и алатки беа употребени за да се оцени ефикасноста на класификаторите за машинско учење на небалансираните множества податоци за предикција на дијабетес. Експериментите беа спроведени со помош на Python, со тренирање и тестирање на моделот извршена во Jupyter Notebook од платформата Anaconda.

3.1 Опис на множеството податоци

Множеството податоци што се користи во оваа студија е множество податоци за здравствени индикатори за дијабетес, добиен од репозиториумот за машинско учење на UCI (Data ID: 891). Целната променлива, Дијабетес бинарен податок, ја претставува класата 0 (без дијабетес) и класата 1 (дијабетес). Оригиналното множество податоци покажа значителен дисбаланс на класата, со 174.595 примероци од класа 0 и 28.349 примероци од класа 1, што доведе до сооднос од приближно 6:1. Овој дисбаланс поставува предизвици за алгоритмите за машинско учење, посебно во прецизното идентификување на случаите на малцинската класа.

Множеството податоци содржи 21 карактеристики (одличја), вклучувајќи неколку демографски променливи (возраст, пол и ниво на образование), бихејвиорални променливи (навики за пушење, физичка активност, консумирање алкохол) и индикатори поврзани со здравјето (BMI, физичко здравје, ментално здравје). Конкретно, нема вредности што недостасуваат во множеството податоци, што ги поедноставува чекорите на претпроцесирање. Множеството податоци е добро структурирано, што го прави погоден за модели за машинско учење без потреба од екстензивно чистење на податоците.

3.2 Претпроцесирање на податоците

Имајќи го предвид високиот степен на дисбаланс на класите, беа имплементирани техники за ресемплирање за да се балансира множеството податоци и да се подобрат перформансите на класификацијата. Беа користени два метода за oversampling:

- SMOTEtomek: е хибридна техника ресемплирање што комбинира SMOTE (Synthetic Minority Over-sampling Technique) за да генерира синтетички инстанци од малцинската класа и Tomek Links за отстранување на двосмислени инстанци во

близина на границата на одлучување помеѓу класите. Овој комбиниран пристап го подобрува квалитетот на множеството податоци и ги подобрува перформансите на моделот на небалансирани множества податоци (Wang, Wu, Zheng, Niu, & Wang, 2021).

- ADASYN (Adaptive Synthetic Sampling Method): е напредна техника за ресемплирање дизајнирана да одговори на дисбалансот на класите во машинското учење. За разлика од SMOTE, кој генерира синтетички примероци подеднакво, ADASYN се фокусира на инстанци од малцинската класа потешки за класификација. Доделува тежини врз основа на бројот на соседи од мнозинската класа, генерирајќи повеќе синтетички примероци за примери во близина на границата на одлуката. Овој адаптивен пристап ги подобрува перформансите на класификаторот, посебно во откривањето на инстанци на малцинската класа, со намалување на пристрасноста кон мнозинската класа (He et al., 2008).

3.3 Модели за класификација

Два модели на машинско учење беа користени за да се проценат перформансите на балансираното множество податоци: XGBoost Classifier и Balanced Random Forest Classifier.

- XGBoost (Extreme Gradient Boosting) е високо ефикасен и скалабилен алгоритам за машинско учење дизајниран за надгледувани задачи за учење. Се надоврзува на принципот на зголемување на градиентот (gradient boosting), каде дрвата на одлучување се обучуваат последователно, при што секое ги коригира грешките од претходното. Алгоритамот вклучува неколку подобрувања, вклучувајќи регулација (L1 и L2), што помага да се спречи преоптоварување и ја подобрува генерализацијата на моделот. XGBoost е познат по своите перформанси и брзина, користејќи техники како што се паралелна обработка и кастрење на дрва за да се оптимизација на пресметките (Chen & Guestrin, 2016).
- Balanced Random Forest (BRF, балансирана случајна шума): е подобрена верзија на традиционалниот метод на случајна шума дизајнирана да ги решава проблемите во небалансираните множества податоци. BRF работи со модифицирање на стандардниот пристап Random Forest за да создаде избалансирани примероци за подигање за обука на секое дрво. Ова балансирање се постигнува со прилагодување на класите на тренираачкото множество, обезбедувајќи дека малцинските и мнозинските класи се застапени подеднакво. Се воведуваат модификации на BRF за понатамошно подобрување на неговите перформанси во предвидувањето на малцинските класи со инкорпорирање на стратегии кои се однесуваат на специфични предизвици поврзани со небалансирани податоци (Kobyliński & Przepiórkowski, 2008).

3.4 Метрика за евалуација

За да се споредат перформансите на моделите, користени се повеќе метрики за евалуација:

- Accuracy: Ја мери целокупната точност на моделот со пресметување на односот на правилно класифицирани примероци со вкупниот број инстанци.
- Precision: Го означува соодносот на вистински позитивни предвидувања меѓу сите позитивни предвидувања, фокусирајќи се на квалитетот на позитивните класификации.
- Recall: ја мери способноста на моделот правилно да ги идентификува позитивните примери или соодносот на вистински позитивни страни од сите вистински позитивни.
- F1-Score: Хармониска средина на прецизноста и recall, обезбедувајќи единствена метрика која ги балансира двата аспекта.

Перформансите на секој модел беа оценети со користење на овие метрики за да се обезбеди сеопфатно разбирање и споредба на нивната ефикасност при справување со небалнсираните податоци.

3.5 Ресемплирање и избор на класификатор

За да се процени ефективноста на различните техники за ресемплирање и алгоритми за машинско учење, беа применети четири различни конфигурации на модели на небалнсирани множества податоци:

1. SMOTETomek со XGBoost класификатор
2. ADASYN со XGBoost класификатор
3. SMOTETomek со Balanced Random Forest
4. ADASYN со Balanced Random Forest

Компаративните резултати го истакнаа влијанието на различните техники за ресемплирање врз способноста на класификаторот правилно да ја предвиди класата на малцинствата, обезбедувајќи ја ефективноста на предложените техники за управување со небалансирани множества податоци.

4. Анализа, резултати и дискусија

Во предиктивната аналитика, посебно во медицинските области како што е предвидувањето на дијабетес, справувањето со небалансираните множества податоци е значаен предизвик. Прецизната класификација на малцинските случаи станува клучна, бидејќи тие често ги претставуваат најкритичните резултати. За да се реши ова, беа применети различни модели на машинско учење на небалнсирано множество податоци за дијабетес, со фокус на оптимизирање на севкупните перформанси и на откривање на малцинската класа. Моделите што се имплементирани вклучуваат ADASYN во комбинација со Balanced Random Forest, ADASYN во комбинација со XGBoost, SMOTEtomek во комбинација со XGBoost и SMOTEtomek во комбинација со Balanced

Random Forest. Анализата се фокусира на оценување на перформансите на класификацијата врз основа на метрика како што се прецизност, точност, recall и F1-score, со посебен акцент на откривањето на малцинските класи.

4.1 ADASYN со Balanced Random Forest класификатор

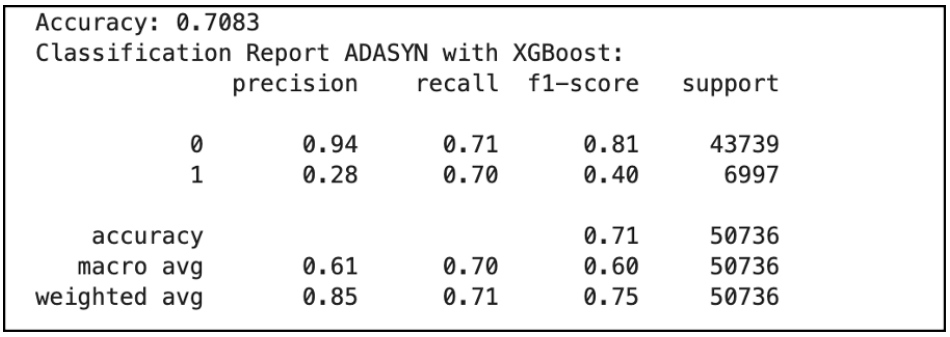
Моделот ADASYN со Balanced Random Forest, како што е прикажано на слика 1, постигна точност од 0,7805, што е умерено високи перформанси, посебно во класифицирањето на мнозинската класа (класа 0). За класа 0, прецизноста (0,91), recall (0,83) и f1-оценката (0,87) укажуваат на силни резултати од класификацијата. Сепак, моделот се соочува со малцинската класа (класа 1), како што е потврдено со ниска прецизност од 0,31 и f1-резултат од 0,38. Повлекувањето за класа 1 е 0,49, што значи дека моделот правилно идентификува речиси половина од случаите на малцинската класа, но погрешно класифицира значителен број лажни позитивнии, што доведува до мала прецизност.

```
Accuracy: 0.7805
Classification Report ADASYN Balanced Random Forest:
              precision    recall  f1-score   support

           0       0.91      0.83      0.87     43739
           1       0.31      0.49      0.38      6997

    accuracy                           0.78     50736
   macro avg       0.61      0.66      0.62     50736
weighted avg       0.83      0.78      0.80     50736
```

Слика бр.1: Извештај за класификација за ADASYN со класификатор Balanced Random Forest на небалансирано множество податоци

4.2 ADASYN со XGBoost класификатор

Моделот ADASYN со XGBoost забележа помала вкупна точност од 0,7083. Изведбата во класа 0 беше реална, со прецизност од 0,94 и f1-резултат од 0,81, но recall-от опадна на 0,71. За малцинската класа, моделот покажа подобрена вредност на recall на 0,70, но прецизноста остана ниска на 0,28, што доведе до мало подобрување на f1-оценката до 0,40 во споредба со ADASYN со Balanced Random Forest. Ова сугерира дека иако моделот е подобар во идентификувањето на малцинската класа, тој сепак произведува значителен број лажни позитивни, што негативно влијае на прецизноста.

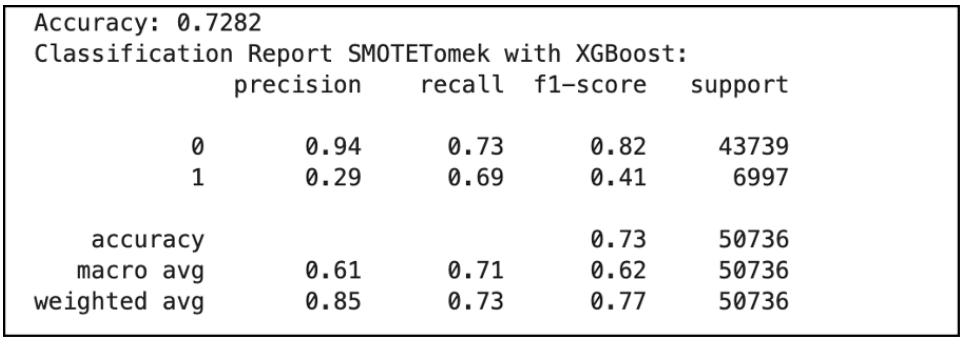


```
Accuracy: 0.7083
Classification Report ADASYN with XGBoost:
              precision    recall  f1-score   support

           0       0.94      0.71      0.81     43739
           1       0.28      0.70      0.40      6997

    accuracy                           0.71     50736
   macro avg       0.61      0.70      0.60     50736
weighted avg       0.85      0.71      0.75     50736
```

Слика бр.2: Извештај за класификација за ADASYN со XGBoost класификатор на неурамнотежен збир на податоци

4.3 SMOTETomek со XGBoost класификатор

Моделот SMOTEtomek со XGBoost покажа слични резултати како ADASYN со XGBoost, со мало подобрување во вкупната точност на 0,7282. За малцинската класа, моделот постигна прецизност од 0,29 и отповикување од 0,69, давајќи f1-оценка од 0,41. Иако ова е мало подобрување во однос на моделот XGBoost базиран на ADASYN, перформансите сепак укажуваат на предизвици во прецизното идентификување на примероци од класа 1. Изведбата на мнозинската класа остана конзистентна, со прецизност од 0,94 и f1-резултат од 0,82. Генерално, моделот се соочува со прецизноста на балансирање и recall за малцинската класа.



```
Accuracy: 0.7282
Classification Report SMOTETomek with XGBoost:
              precision    recall  f1-score   support

           0       0.94      0.73      0.82     43739
           1       0.29      0.69      0.41      6997

    accuracy                           0.73     50736
   macro avg       0.61      0.71      0.62     50736
weighted avg       0.85      0.73      0.77     50736
```

Слика бр.3: Извештај за класификација за SMOTEtomek со XGBoost класификатор на небалансирано множество податоци

4.4 SMOTETomek со Balanced Random Forest класификатор

Моделот SMOTEtomek со Balanced Random Forest ги надмина сите други модели со значителна разлика, постигнувајќи точност од 0,8947. И мнозинската и малцинската класа покажаа избалансирани и висококвалитетни перформанси. Малцинската класа (класа 1) постигна прецизност од 0,91, отповикување од 0,91 и f1-резултат од 0,91. Ова ниво на перформанси покажува дека моделот не само што правилно ги идентификува случаите на малцинската класа, туку тоа го прави и со ниска стапка на лажни позитивни, што резултира

со висок f1-резултат. Мнозинската класа исто така се претстави добро, со прецизност, recall и f1-score, сите блиску до 0,88.

```
Accuracy: 0.8947
Classification Report for SMOTETomek with Balanced Random Forest:
              precision    recall  f1-score   support

           0       0.88      0.88      0.88         8
           1       0.91      0.91      0.91        11

    accuracy                           0.89        19
   macro avg       0.89      0.89      0.89        19
weighted avg       0.89      0.89      0.89        19
```

Слика бр.4: Извештај за класификација за SMOTETomek со класификатор Balanced Random Forest на небалансирано множество податоци

4.5 Дискусија

Резултатите од четирите модели се претставени во однос на вкупната точност и извештаи за класификација кои вклучуваат прецизност, recall и f1-оценки и за мнозинската и за малцинската класа. Во Табела 1 е дадено резиме на клучните индикатори за перформанси.

Model	Accuracy	Precision	Recall	F1-Score
ADASYN со Balanced Random Forest	0.7805	0.31	0.49	0.38
ADASYN со XGBoost	0.7083	0.28	0.70	0.40
SMOTETomek со XGBoost	0.7282	0.29	0.69	0.41
SMOTETomek со Balanced Random Forest	0.8947	0.91	0.91	0.91

Табела 1: Компаративни перформанси на ADASYN и SMOTEtomek со XGBoost и Balanced Random Forest Classifiers за предвидување дијабетес од небалнсирано множество податоци

Споредбата на четирите модели открива неколку клучни сознанија за ефективноста на различните техники за преземање примероци и класификатори за справување со небалнсирани множества податоци:

- Ефективност на SMOTETomek: Моделите што користат SMOTETomek постојано ги надминуваат оние што користат ADASYN, особено во однос на f1-резултатот и вкупната точност. SMOTETomek, кој комбинира прекумерно и недоволно земање примероци, се чини дека обезбедува подобра рамнотежа помеѓу класите, помагајќи поефикасно да се ублажи дисбалансот на класите.
- Супериорни перформанси на балансиран класификатор на случајни шуми: Моделите што користат Balanced Random Forest покажаа супериорни перформанси во споредба со XGBoost, особено во откривањето на малцинската класа. Балансираната случајна шума се чини дека подобро се справува со небалансираните податоци, посебно кога се комбинира со техниката SMOTEtomek, како што е потврдено со високата прецизност, recall и f1-score за малцинската класа.
- Предизвици со XGBoost: Додека XGBoost е познат по своите силни перформанси на многу задачи за класификација, во овој случај се соочи со небалнсирани множества податоци, посебно во однос на прецизноста за малцинската класа. Иако постигна релативно високи вредности за recall, слабата прецизност (под 0,30 и во поставките ADASYN и SMOTEtomek) укажува на значаен проблем со лажни позитивни, што ја намалува неговата севкупна ефикасност.
- Balanced Random Forest со SMOTEtomek: Комбинацијата на SMOTEtomek и Balanced Random Forest ги даде најдобрите вкупни резултати, со речиси совршени перформанси на двете класи. Овој модел обезбедува не само висока точност, туку и многу силна способност за балансирање на прецизноста и потсетувањето и за мнозинските и за малцинските класи. Тоа е јасен избор кога и севкупната точност и откривањето на малцинската класа се критични.
- Наоѓање на вистинскиот баланс меѓу прецизноста и recall: кај сите модели, освен последниот (SMOTETomek со Balanced Random Forest), постоеше конзистентна врска помеѓу прецизноста и recall за малцинската класа. Ова е вообичаен предизвик во небалансираните множества податоци, каде што зголемената вредност на recall често доведува до намалување на прецизноста поради погрешната класификација на примероците од мнозинската класа како примероци од малцинска класа. Конечниот модел успешно ги избалансираше овие две метрики, истакнувајќи ја важноста на правилната комбинација на техники за ресемплирање и класификатори.

5. Заклучок

Овој проект имаше за цел да се справи со предизвиците поврзани со високо небалансираните множества податоци во предвидувањето на дијабетес преку имплементација и евалуација на два класификатори за машинско учење, XGBoost и Balanced Random Forest, во комбинација со два методи за ресемплирање, SMOTEtomek и ADASYN. Главниот фокус беше на имплементација и анализа на четири различни модели како што се:

- SMOTETomek во комбинација со XGBoost
- SMOTETomek во комбинација со Balanced Random Forest
- ADASYN во комбинација со XGBoost, and
- ADASYN во комбинација со Balanced Random Forest.

Методологијата вклучуваше примена на овие техники за ресемплирање за да се балансира високо небалансирано множество податоци и користење на евалуација на перформансите на моделот со клучните метрики како што се точност, прецизност, recall и f1-резултат.

Врз основа на сеопфатна евалуација на перформансите на моделите, анализата откри дека SMOTEtomek со балансирана случајна шума резултираше како модел со најдобри перформанси, постигнувајќи најголема точност (0,89) и f1-резултат од 0,91 за малцинската класа. Ја балансираше високата предиктивна прецизност со способноста за ефикасно справување со нерамнотежата на класата, што го прави особено погоден за примена на множества податоци за небалнсирани медицински и здравствени податоци, како што е предвидување на дијабетес. Од друга страна, моделите кои користат XGBoost, се соочија со предизвици за прецизност, посебно ако се применат со ADASYN. Овие наоди ја нагласуваат важноста од внимателен избор на методи за ресемплирање и класификатори за да се постигнат оптимални перформанси на предвидување на небалансирани множества податоци.

За идно истражување на мојата докторска дисертација, имам намера да ја испитам примената на длабокото учење и невронските мрежи за предикција од небалансирани множества податоци. Додека претходната досегашна работа ги истражуваше традиционалните модели на машинско учење во комбинација со техниките за ресемплирање како SMOTEtomek и ADASYN, мојот фокус ќе се префрли на развивање на нови архитектури на невронски мрежи кои можат подобро да се справат со дисбалансираните податоци. Ова вклучува користење на врвни достигнувања како што се конволуциски невронски мрежи (CNN) за екстракција на карактеристики, мрежи со long short-term меморија (LSTM) за секвенцијално справување со податоци и модели за трансформација за да се подобри препознавањето на облиците во комплексни множества податоци. Дополнително, интегрирањето на техники како што се генеративни соперснички (антагонистички) мрежи (generative adversarial networks) (GANs) за усогласување на податоците за малцинската класа може да помогне во решавањето на дисбалансот на класите. Со инкорпорирање на овие напредни невронски архитектури и неодамнешни трендови како што се само-надгледувано учење, механизми за внимание и длабоко учење на ансамблот, целта е да создадавање поробусен, поточен и скалабилен модел за предвидување дијабетес. Овој пристап ќе ги помести границите на предиктивната аналитика и ќе обезбеди значителен придонес во справувањето со небалансираните медицински множества податоци.

Користена литература

Abushahla, K. H., & Pala, M. A. (2024). Optimizing Diabetes Prediction: Addressing Data Imbalance with Machine Learning Algorithms. ADBA Computer Science, 1(1), 26-35. https://doi.org/10.69882/adba.cs.2024075

Alkharabsheh, K., Alawadi, S., Kebande, V. R., Crespo, Y., Fernández-Delgado, M., & Taboada, J. A. (2022). A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: A study of God class. Information and Software Technology, 143, 106736. https://doi.org/10.1016/j.infsof.2021.106736

Beyene, S. S., Ling, T., Ristevski, B., & Chen, M. (2020). A novel riboswitch classification based on imbalanced sequences achieved by machine learning. PLoS computational biology, 16(7), e1007760. https://doi.org/10.1371/journal.pcbi.1007760

Carrington, A. M., Fieguth, P. W., Qazi, H., Holzinger, A., Chen, H. H., Mayr, F., & Manuel, D. G. (2020). A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC medical informatics and decision making, 20, 1-12. https://doi.org/10.1186/s12911-019-1014-6

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. Journal of Chemical Information and Modeling, 61(6), 2623-2640. https://pubs.acs.org/doi/10.1021/acs.jcim.1c00160

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE access, 8, 67899-67911. 10.1109/ACCESS.2020.2986809

Ghorbani, R., Ghousi, R., Makui, A., & Atashi, A. (2020). A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset. IEEE Access, 8, 141066-141079. 10.1109/ACCESS.2020.3013320

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

Kobyliński, Ł., Przepiórkowski, A. (2008). Definition Extraction with Balanced Random Forests. In: Nordström, B., Ranta, A. (eds) Advances in Natural Language Processing. GoTAL 2008. Lecture Notes in Computer Science(), vol 5221. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-85287-2_23

Kumar, M. S., Khan, M. Z., Rajendran, S., Noor, A., Dass, A. S., & Prabhu, J. (2022). Imbalanced classification in diabetics using ensembled machine learning. Computers, Materials and Continua, 72(3), 4397-4409. https://doi.org/10.32604/cmc.2022.025865

Kumar, V., Lalotra, G. S., Sasikala, P., Rajput, D. S., Kaluri, R., Lakshmanna, K., ... & Uddin, M. (2022, July). Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. In Healthcare (Vol. 10, No. 7, p. 1293). MDPI. https://doi.org/10.3390/healthcare10071293

Mohammed, A. J., Hassan, M. M., & Kadir, D. H. (2020). Improving classification performance for a novel imbalanced medical dataset using SMOTE method. International Journal of Advanced Trends in Computer Science and Engineering, 9(3), 3161-3172. https://doi.org/10.30534/ijatcse/2020/104932020

Sasada, T., Liu, Z., Baba, T., Hatano, K., & Kimura, Y. (2020). A resampling method for imbalanced datasets considering noise and overlap. Procedia Computer Science, 176, 420-429. https://doi.org/10.1016/j.procs.2020.08.043

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big data, 7, 1-47. https://doi.org/10.1186/s40537-020-00349-y

Ul Hassan, I., Ali, R. H., Ul Abideen, Z., Khan, T. A., & Kouatly, R. (2022). Significance of machine learning for detection of malicious websites on an unbalanced dataset. Digital, 2(4), 501-519. https://doi.org/10.3390/digital2040027

Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. Knowledge and Information Systems, 65(1), 31-57. https://doi.org/10.1007/s10115-022-01772-8