УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

ПРИМЕНА АЛГОРИТАМА МАШИНСКОГ УЧЕЊА ЗА ДЕТЕКЦИЈУ ДИЈАБЕТЕСА

Дипломски рад

Ментор:

др. Дражен Драшковић,

доцент

Кандидат:

Андријана Миковић,

0363/2019

Београд, септембар 2023.

САДРЖАЈ

1. Увод

Дијабетес (lat. *diabetes mellitus*) је хронична болест обележена високом количином глукозе у крви. Код људи са дијабетесом типа 1, панкреас не производи довољну количину инсулина, хормона који контролише употребу глукозе у ћелијама ради обезбеђења неопходне енергије. У овом контексту глукоза се задржава у крви, а то изазива низ здравствених проблема. Док дијабетес типа 2 указује на проблем да ћелије више не могу да користе инсулин на прави начин за стварање енергије. Овај облик захтева озбиљнији приступ лечењу и не може да се држи под контролом само изменом лоших животних навика. А може настати погоршањем дијабетеса типа 1.

Дијабетес спада у критична обољења, са учесталим порастом броја људи који пате од њега, у модерном свету. Разни фактори утичу на појаву ове болести, а велики део се односи на стил живота пацијената, као што су килажа, мањак физичке активности, лоша исхрана, године, високи притисак и сл. Код пацијената са дијабетесом, расте ризик од појаве других обољења као што су болести срца, бубрега, проблема са очима, оштећења нерва и сл. [1]. Рано откривање дијабетеса, омогућава брже започињање лечења, и самим тим смањује ризик од развијања других обољења или од добијања дијабетеса типа 2.

Пацијентима се врши дијагностиковање дијабетеса, на основу разних информација о пацијенту и резултата тестова, што производи велику количину медицинских података, који се касније могу анализирати и употребити за проналажење повезаности између других фактора и дијабетеса, за које је можда медицина сматрала да не доприноси давању дате дијагнозе.

У овом раду је коришћења база података која се састоји од информација о 100 000 пацијената, са и без дијабетеса. База укључује атрибуте као што су пол, године, хипертензија (високи крвни притисак), болести срца, историју пушења, индекс телесне масе, ХбА1ц ниво (мери количину глукозе прикачену на хемоглобин у црвеним крвним зрнцима током њиховог животног века, 2-3 месеца), као и ниво глукозе у крви.

У другом поглављу врши се осврт на претходна истраживања. Након тога биће изложена анализа саме базе података, као и утицаја разних фактора на дијагнозу дијабетеса у трећем поглављу. Алгоритми коришћени при класификацији, односно дијагностици

дијабетеса, обрађени су у четвртом поглављу. У поглављу пет анализиран је проблем неуравнотеженог скупа података, и различити приступи решавању проблематике прекомерног узорковања. Затим следе резултати класификације, упоређивањем прецизности и ефикасности алгоритама обрађених у четвртом поглављу.

2. АНАЛИЗА ПРОБЛЕМА И ПРЕГЛЕД ПОСТОЈЕЋИХ РЕШЕЊА

У овом поглављу, биће анализиран глобални проблем дијабетеса који се све више шири у светској популацији. Нагласићемо важност ране дијагнозе овог обољења, као и потребу за аутоматизацијом дијагностичких процеса. Осим тога направиће се осврт на претходна истраживања на ову тему. Током анализе, посебну пажњу ћемо посветити њиховим приступима, изборима алгоритама, базама података као и самим резултатима тих истраживања.

2.1. Опште информације

На основу података интернационалне дијабетес федерације, на свету постоји 537 милиона пацијената који болују од дијабетеса. Предвиђају да до 2045 године, ће 1 у 8 одраслих особа имати ову дијагнозу, тј. оквирно 783 милиона људи, што представља пораст од 46%. Преко 90% пацијената са дијабетесом имају дијабетес типа 2, кључни фактори за раст ове бројке су урбанизација, старење популације, смањење физичке активности као и пораст броја гојазних особа [6].

Могуће је смањити утицај дијабетеса применом превентивних мера, као и раном дијагнозом. Овакве превенције помажу људима да живе са својом дијагнозом и спречавају појаве других обољења и компликација. Аутоматизација дијагностиковања би убрзала процес, и омогућила ранију дијагнозу поготову код пацијената којима медицинска нега није толико доступна.

2.2. Релевантна истраживања

У [3] је одрађена класификација дијабетеса употребом *SVM (*енгл. *support vector machine)* – метода потпорних вектора. Основна идеја овог алгоритма је да пронађе оптималну границу, хипер-раван, између различитих класа података. Захтева пажљиво подешавање хиперпараметара, који су му потребни за класификацију и може бити захтеван за тренирање на великом скупу података. Добијена тачност у овом истраживању је била 78%. Вршено је на подацима из базе података под називом *Prima Indians Diabetes Database (PIMA),* која такође има проблем са неуравнотежености података, као и база коришћена у овом истраживању.

Састоји се од информација о 768 женских пацијената, од којих 500 је негативно на дијабетес. Осим мануелне провере параметара, и уклањања невалидних података, није извршена регулација неуравнотежености базе.

Аутори [5] су користили исту базу података као и [3] са варијабилним аутоенкодером (*VAE Variational Autoencoder*) као тактиком решавања проблема неуравнотежености базе. Након тога, вршено је тестирање неколико алгоритама за детекцију дијабетеса на оригиналној бази података, као и на допуњеној уравнотеженој верзији. Може се приметити да од класичних алгоритама машинског учења најбољу тачност су добили употребом *XGBoost* алгоритма који се базира на стаблу одлучивања, а након тога употребом стабла одлучивања и методом случајне шуме. Ови алгоритми су упоређивани са резултатом класификације дубоким учењем, коришћењем вишеслојне мреже која на допуњеној бази даје нижу тачност од стабла одлучивања и методе случајне шуме. Касније је примењен резређени аутоенкодер (*sparse autoencoder*) и заједно са конволуционом неуронском мрежом враћа тачност од 92,31%. То представља обећавајући резултат, али је ограничен малим скупом података који ова база пружа.

Аутори [1] су вршили класификацију на својој бази података која се састоји од 800 пацијената као и на *PIMA* бази. Резултати овог истраживања показују да на њиховој бази највећу тачност има алгоритам логистичке регресије од 96%, док иста примењена на *PIMA* бази извршава класификацију са тачношћу од 76%. Овај рад нам не даје јасан увид у коришћену базу података, осим информација о називима атрибута и њиховом типу. Не можемо са сигурношћу да знамо да ли њихова база не садржи проблем неуравнотежености као *PIMA* и да ли даје реалну слику о ефикасности класификације.

3.

БАЗА ПОДАТАКА

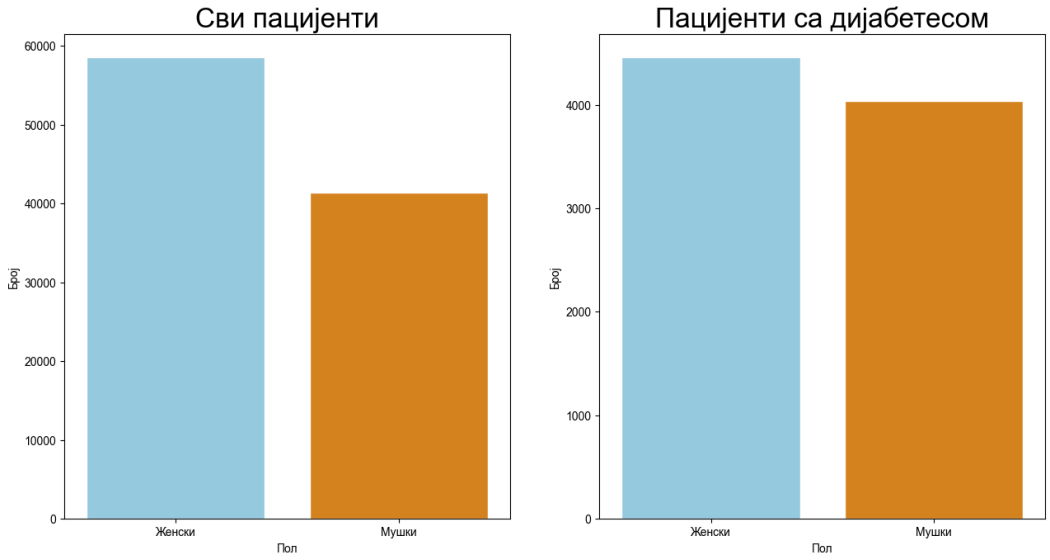
У овом поглављу биће дат опис базе података коришћене у овом истраживању. И анализа утицаја разних атрибута у детекцији дијабетеса код пацијената. Као и избор подскупа улазних атрибута за примену алгоритама машинског учења на проблем дијагностиковања дијабетеса.

3.1. Опште информације

База података коришћена у овом раду састоји се од медицинских и општих подата о пацијентима, заједно са статусом њихове дијагнозе дијабетеса. Она укључује атрибуте као што су пол, године, хипертензија, болести срца, историју пушења, индекс телесне масе, ХбА1ц ниво и ниво глукозе у крви. Садржи информације о 100 000 пацијената, од којих је код 8500 дијагностикиван дијабетес.

3.2. Пол

Дијабетес може подједнако захватити припаднике оба пола. Током трудноће, долази до повећања инсулинске резистенције, односно смањене осетљивошћу ткива на инсулин. Нормалан ниво шећера у крви код трудница се одржава повећањем лучења инсулина, а уколико повећано лучење инсулина не успе да надомести смањену осетљивост ткива, долази до појаве Гестационог дијабетеса, који се не сматра дијагнозом дијабетеса. Жене које имају гестациони дијабетес, имају значајно повећан ризик за развијање дијабетеса типа 2, касније у животу.

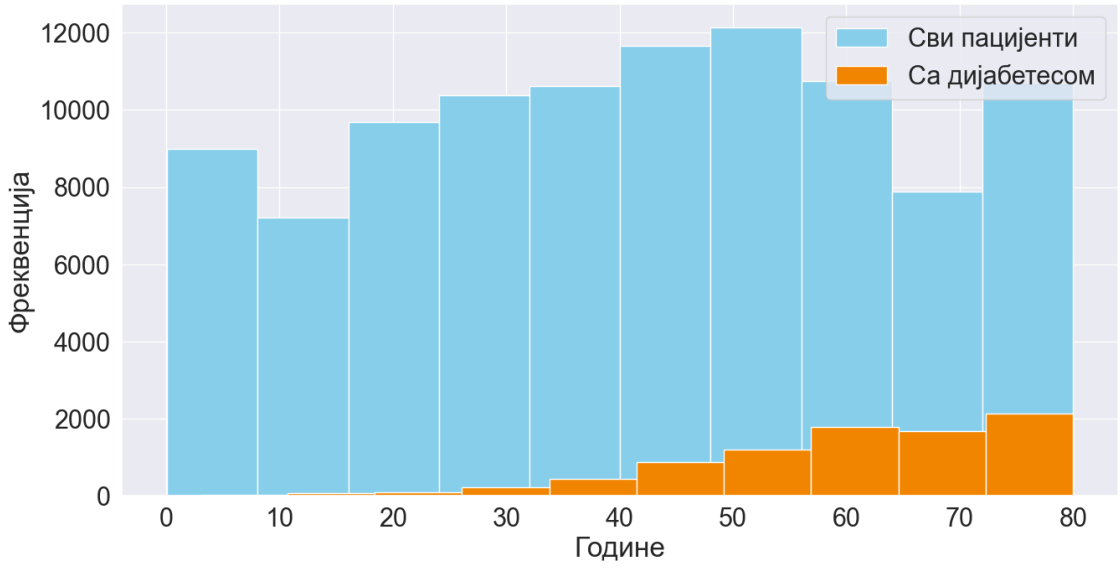


Слика 3.2.1. Графички приказ расподеле родне карактеристике

Као што можемо видети са слике 3.2.1., већи је број жена које имају дијабетес, али је такође већи број жена у самој бази података.

3.3. Године

Године чине битан фактор за дијагностиковање дијабетеса, како особа стари, ризик за развијање ове болести расте, као што можемо видети на слици 3.3.1. Утицај на то има и смањена физичка активност у каснијим стадијумима живота, као и промена нивоа хормона и већи ризик од развијања других болести које могу довести до дијабетеса.



Слика 3.3.1. Дистрибуција година и приказ заступљености дијабетеса у различитим узрастима

3.4. Хипертензија

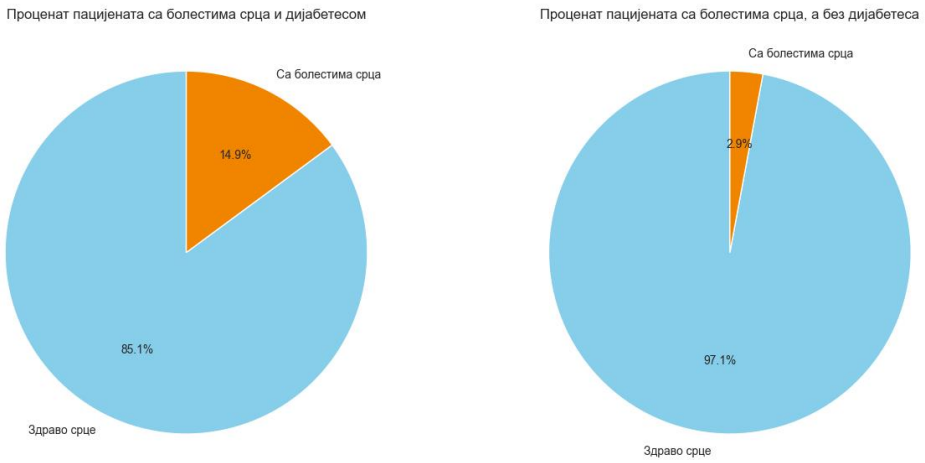
Хипертензија, висок притисак, је обољење које често иде у пару са дијабетесом. Ова два обољења деле ризичне факторе, и ако пацијент болује од једног, повећавају му се шансе развијања другог.

Слика 3.4.1. Заступљеност пацијената са хипертензијом међу пацијентима са дијабетесом и без дијабетеса

На слици 3.4.1. се јасно уочава да је већи број пацијената са дијабетесом такође захваћен хипертензијом, у поређењу са пацијентима који имају само једну дијагнозу.

3.5. Болести срца

Веза између болести срца и дијабетеса је двосмерна, што значи да развијање једног стања повећава ризик од развијања другог. То се дешава зато што ове две болести деле многе заједничке факторе ризика, као што су гојазност, висок крвни притисак итд.

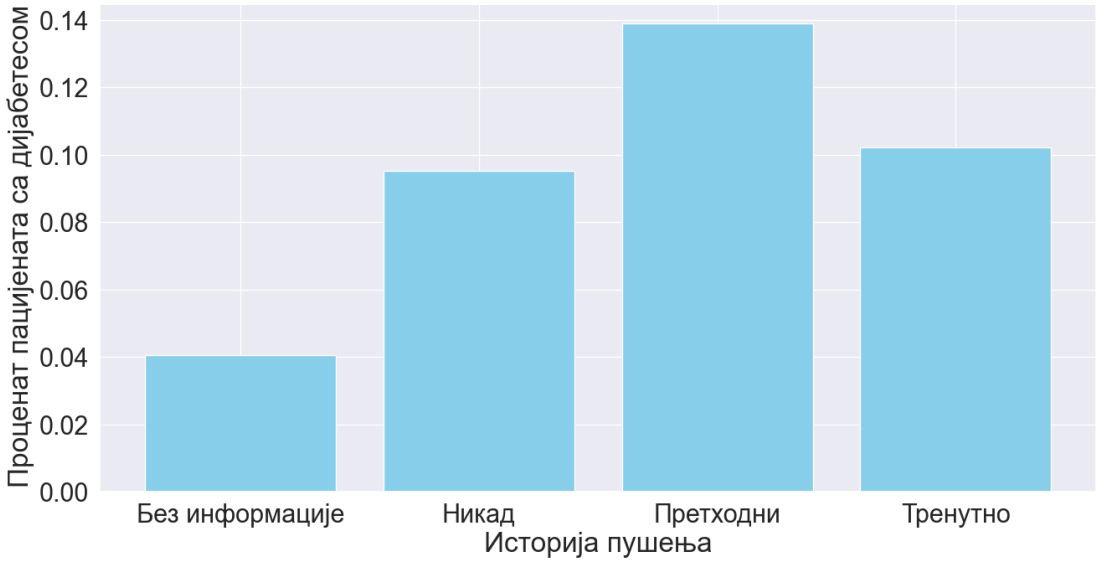


Слика 3.5.1. Заступљеност пацијената са болестима срца међу пацијентима са дијабетесом и без дијабетеса

На слици 3.5.1. се јасно уочава да је већи број пацијената са дијабетесом такође има проблем са срцем, у поређењу са пацијентима који имају само једну дијагнозу.

3.6. Историја пушења

Пушење утиче на више фактора који могу изазвати повећану инсулинску резистенцију, која касније може довести до развијања дијабетеса типа 2. Ова навика такође повећава ризик од развијања обољења срца, која могу довести до дијабетеса [7].

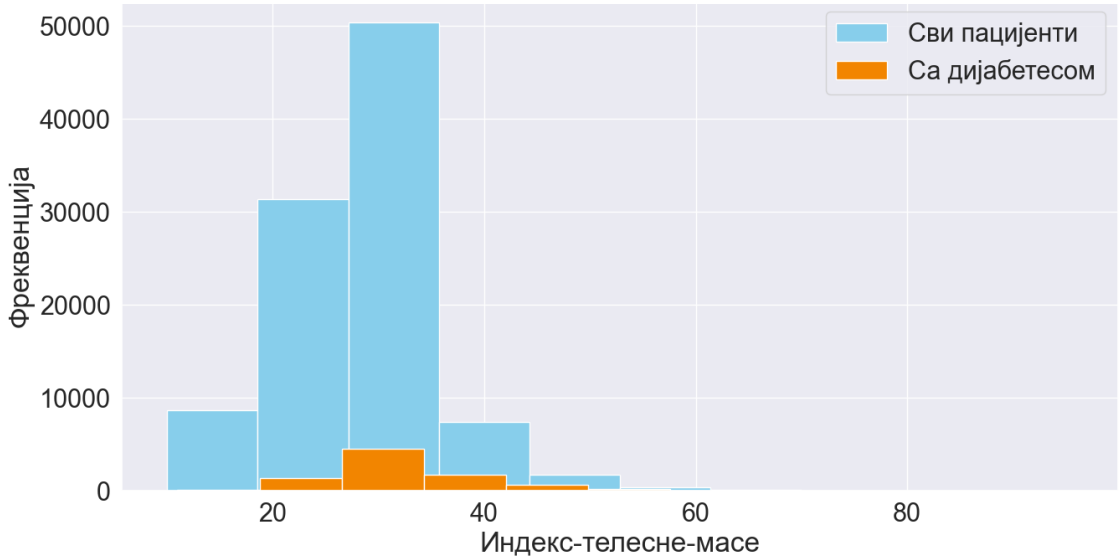


Слика 3.6.1. Проценат пацијената са дијабетесом и одговарајућом информацијом о историји пушења

На слици 3.6.1. уочавамо повећану учестаност дијабетеса међу пацијентима са историјом пушења. Ипак, ова штетна навика није директно повезана са дијабетесом, што значи да се лекари само код половине својих пацијената са дијабетесом одлучују да саветују кориговање ове навике [7].

3.7. Индекс телесне масе

Индекс телесне масе представља однос тежине особе у килограмима и квадрата висине у метрима. Висока вредност индекса телесне масе може указати на висок ниво масних наслага у телу. Он указује на категорије људи који могу имати проблем са претераном количином масних наслага али не представља дијагнозу [8].

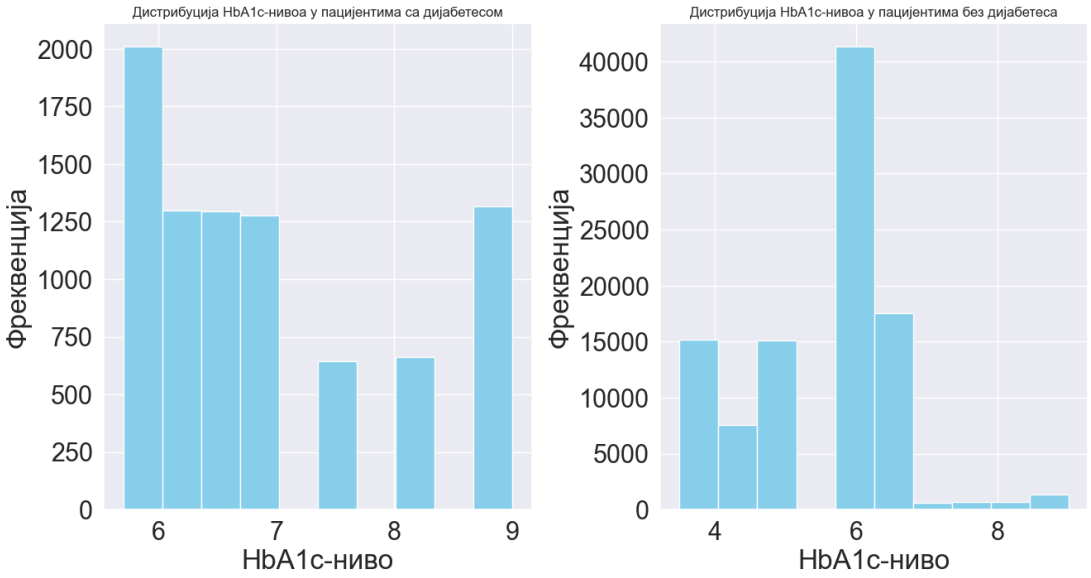


Слика 3.7.1. Дистрибуција индекса телесне масе и приказ заступљености дијабетеса

На слици 3.7.1., примећујемо пораст броја дијагноза дијабетеса код особа са индексом телесне масе 30-40, што представља гојазне особе. Гојазност особа, поготову у ранијим годинама живота, значајно повећава ризик од развијања дијабетеса [9].

3.8. ХбА1ц-ниво

ХбА1ц ниво представља вредност просечне количине нивоа глукозе у црвеним крвним зрнцима током њиховог животног века, 2-3 месеца. Представља информацију о дугорочном нивоу шећера у организму.

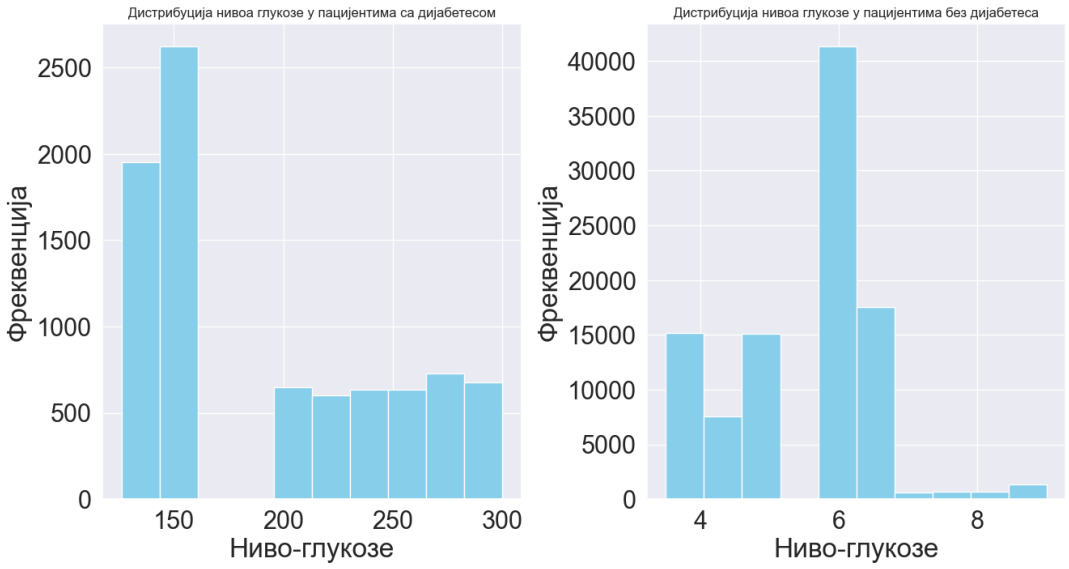


Слика 3.8.1. Дистрибуција ХбА1ц нивоа код пацијената са и без дијабетеса

Као што можемо видети и на слици 3.8.1., више вредности ХбА1ц нивоа могу указивати на дијабетес или његове компликације. Организам особа које имају виши ниво ХбА1ц није у стању да претвори сву потребну глукозу у ћелијску енергију, и она остаје закачена за црвена крвна зрнца.

3.9. Ниво глукозе

Ниво глукозе представља количину шећера у крви, у датом тренутку.

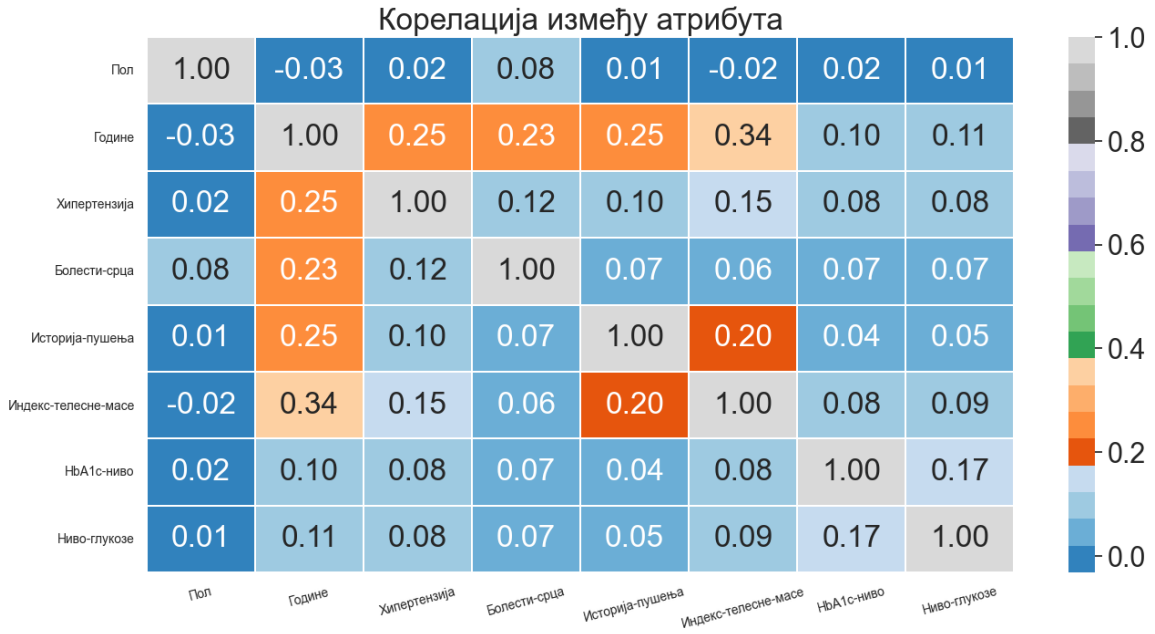


Слика 3.9.1. Дистрибуција нивоа глукозе код пацијената са и без дијабетеса

Повишени ниво глукозе, поготову у тренуцима гладовања или након конзумације шећера може указати на поремећај регулације глукозе и повећати ризик развоја дијабетеса, као што се може видети на слици 3.9.1. Редовно праћење нивоа глукозе у крви игра важну улогу и дијагнози и регулисању дијабетеса.

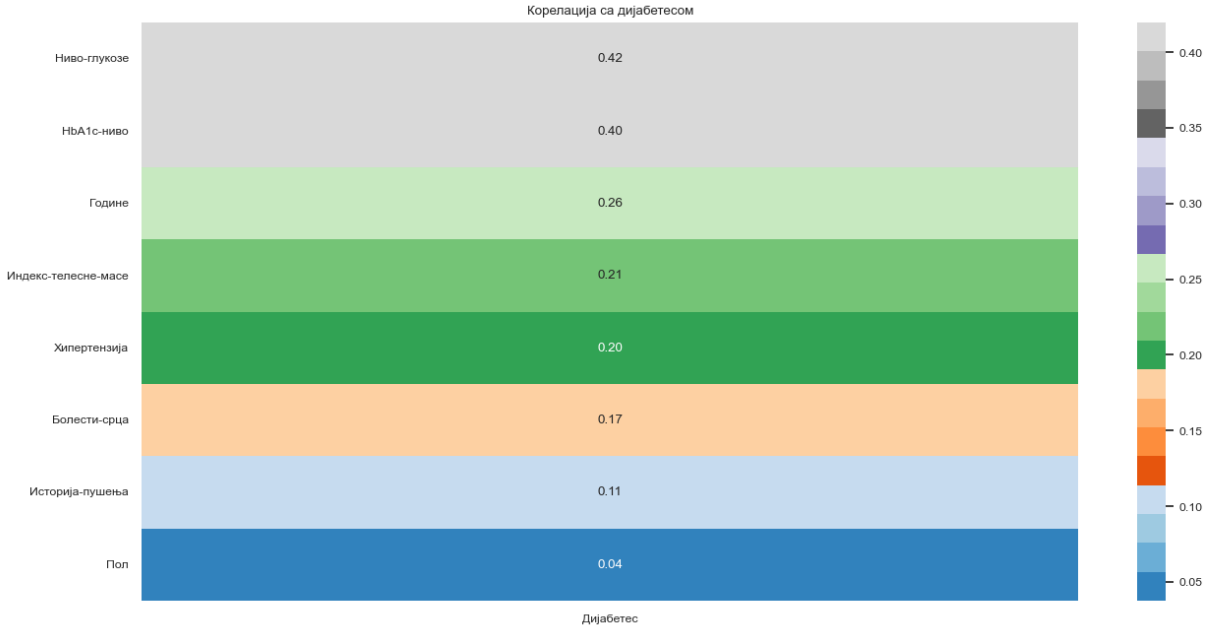
3.10. Закључак утицаја атрибута

Након детаљне анализе атрибута, са медицинске аспекта, можемо истражити њихову међусобну корелисаност на слици 3.10.1. Уочавамо да су године уско повезане са хипертензијом и болестима срца, за развијање којих значајно и расте шанса у каснијем животном добу. Најизраженија корелација примећује се између година живота и индекса телесне масе, што се може објаснити успоравањем стила живота. Такође примећујемо и повезаност историје пушења са годинама и индексом телесне масе, што поново указује на повезаност лоших животних навика, и већу учесталост истих у каснијем животном раздобљу.



Слика 3.10.1. Матрица корелације атрибута

Сада ћемо анализирати повезаност атрибута са дијагнозом дијабетеса и одабрати одговарајући подскуп атрибута за будућу класификацију.



Слика 3.10.2. Корелација атрибута и дијагнозе дијабетеса

Са слике 3.10.2. уочавамо да иако код жена постоји повећана склоност развоју дијабетеса типа 2, веза овог атрибута са дијабетесом није велика. Стога у даљем истраживању нећемо разматрати његов утицај на резултате класификације. Такође, због ниске корелације између историје пушења и дијабетеса, отклањамо и овај атрибут из улазног скупа атрибута за наше алгоритме.

4.

АЛГОРИТМИ КЛАСИФИКАЦИЈЕ

У овом поглављу биће представљена три алгоритма која ће бити централна у процесу класификације и дијагностиковања дијабетеса. Сваки од ових алгоритама представља јединствени приступ решавању проблема дијагностике дијабетеса. Изнећемо њихове теоријске основе, предности и недостатке.

4.1. Логистичка регресија

Логистичка регресија представља математички модел којим се описује веза између атрибута и категоричке зависне променљиве. Она враћа вероватноћу да дати скуп атрибута, са специфичним вредностима, припада свакој од категорија. У оквиру овог истраживања, примењена је бинарна логистичка регресија, јер је потребно утврдити да ли особа пати од дијабетес или не. Овај модел се ослања на употребу логистичке функције како би повезао улазне атрибуте и пружио вероватноћу припадности одређеној категорији.

Централни елемент логистичке регресије је логистичка или сигмоидна функција. Ова функција трансформише линеарну комбинацију улазних атрибута у опсег [0,1], чиме омогућава третирање резултата као вероватноћу испуњења неког догађаја. Сигмоидна функција се математички изражава следећом формулом:

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad (4.1.1.)$$

Где z представља линеарну комбинацију коефицијената и улазних атрибута, тј. линеарни модел добијен следећом формулом:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \qquad (4.1.2.)$$

Модел логистичке регресије за *n* независних атрибута може се математички изразити на следећи начин:

$$P(Y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \qquad (4.1.3)$$

Где је $P(Y = 1)$ вероватноћа да се догађај који се прати, у овом случају дијагноза дијабетеса, оствари путем логистичке регресије, а $\beta_0,\beta_1,\cdots\beta_n$ су коефицијенти регресије. Добијена сменом

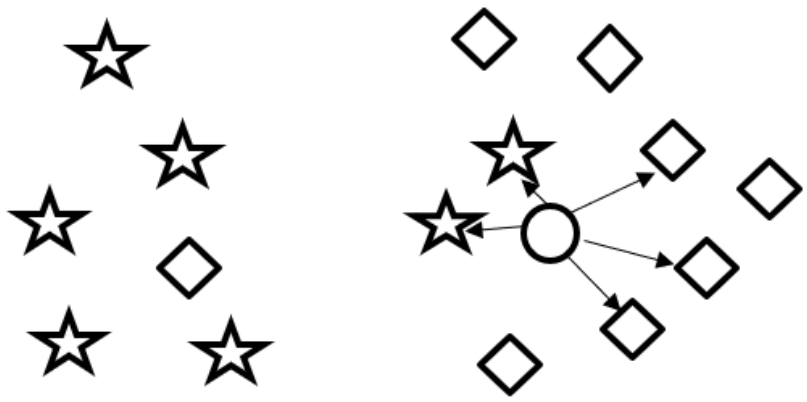
линеарног модела у формулу сигмоидне функције. Логаритам шансе, однос вероватноћа да догађај припада једној класи у односу на другу математички се изражава као:

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \qquad (4.1.4)$$

У овом истраживању коришћена је *scikit-learn* библиотека, програмског језика *python,* која користи оптимизацију градијентног спуста приликом проналажења оптималних вредности коефицијената у моделу логистичке регресије. Иницијализује почетне вредности коефицијената насумично. Користећи ове почетне коефицијенте у комбинацији са одабраним тренинг скупом улазних података, израчунава логаритам шансе. Затим израчунату вредност трансформише помоћу логистичке функције, како би се добила припадност одређеној класи. Модел упоређује израчунате вредности са стварним класама у тренинг скупу података, како би израчунао губитак функцију. Итеративно се коригују коефицијенти са циљем минимизације губитак функције, и добијања прецизније класификације.

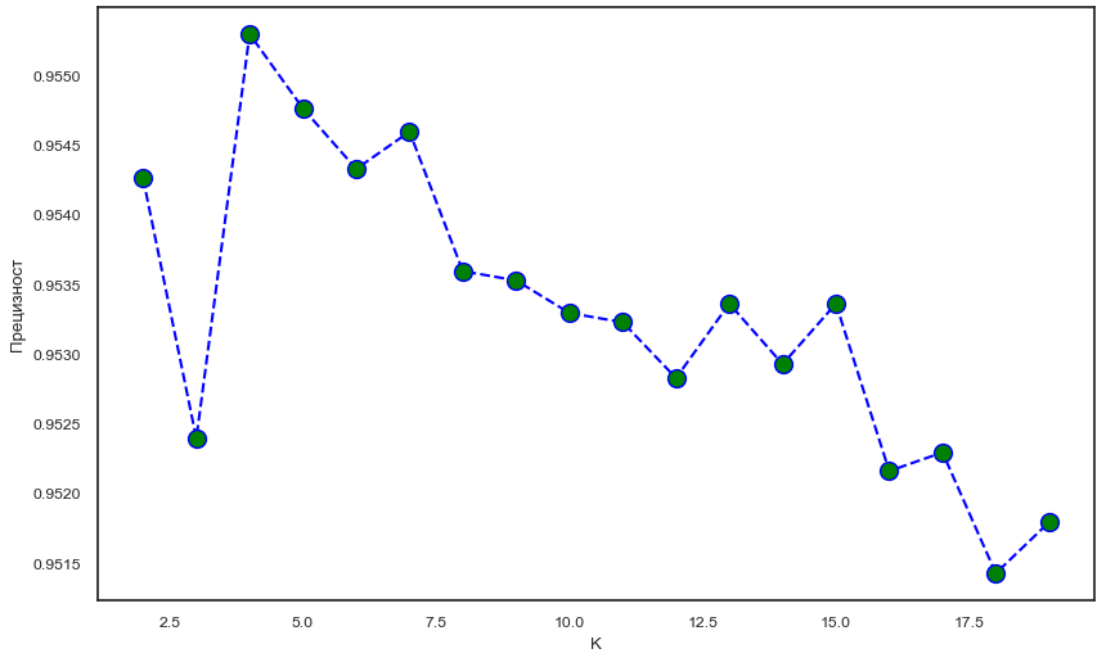
4.2. КНН – к најближих комшија

Алгоритам к најближих суседа скр. KNN (енгл. *k-nearest neighbors*) представља непараметарски метод, надгледног учења. Циљ КНН класификације је додељивање нових узорака одређеним класама на основу сличности са већ познатим тренинг скупом података. Користи метрику удаљености како би мерио сличности између узорака у простору атрибута. Ова метрика се користи за проналажење *к* најближих суседа за сваки нови узорак.



Слика 4.2.1. Визуелизација КНН алгоритма за к=3

Параметар *к* представља број најближих суседа који ће бити узети у обзир приликом класификације новог узорка. Одабиром мање вредности параметра, постоји ризик од грешке у класификацији услед присуства шума у подацима. На пример, ако бисмо применили КНН са к = 2 на узорку са слике 4.2.1., нови узорак (круг) би био класификован као класа А, односно као звезда, иако се у његовој околини налази више узорака класе Б, која представља његову стварну класу.

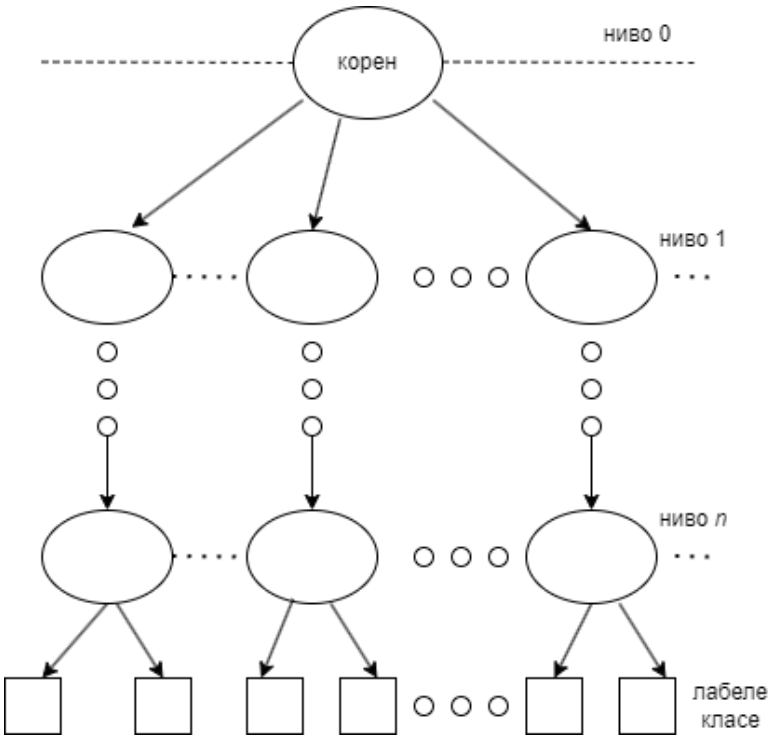


Слика 4.2.2. Прецизност КНН алгоритма за различите вредности параметра

Као што се види са слике 4.2.2., повећавање вредности параметра не доводи нужно до прецизније класификације. Модел постаје мање осетљив на локалне варијације у подацима, границе ће се простирати даље од тачке која се класификује, а то може довести до погрешне класификације. Поред варијације параметра, могуће је варирати и избор метрике за мерење удаљености узорака у простору.

4.3. Стабло одлучивања

Стабло одлучивања представља алгоритам машинског учења који се базира на концепту доношења одлука у складу са условима и акцијама. Карактерише га јасна графичка структура, у облику стабла приказана на слици 4.3.1.



Слика 4.3.1. Прецизност стабла одлучивања

У стаблу одлучивања, корен представља почетни услов који одређује избор параметра на основу кога се врши раздвајање података. Гране стабла представљају различите услове који се примењују на податке, а листови садрже финалне одлуке, тј. излазне резултате алгоритма. Процес раздвајања на гране захтева меру информационог добитка. Информациони добитак је колико раздвајање на одређени начин помаже у смањењу неодлучности. За сваки могући услов врши се рачунање информационог добитка и одабира се најбољи пут за раздвајање, тј. стабло се изграђује тако да минимизује неодлучност и максимизује информациони добитак.

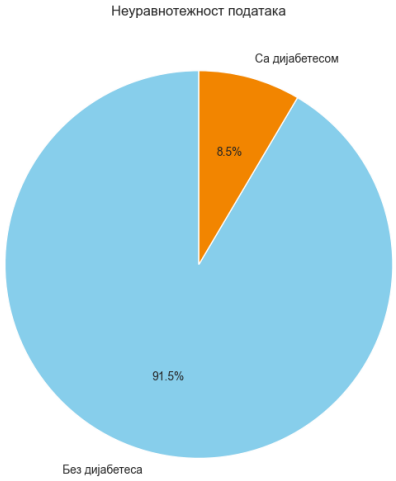
5.

ПРОБЛЕМ НЕУРАВНОТЕЖЕНОГ СКУПА ПОДАТАКА

У овом поглављу биће описан проблем неуравнотеженог скупа података, који се појављује у бази коришћеној у овом истраживању. Овај проблем представља учесталу појаву у медицинским базама података. Прво ће бити дат увид у приступе решавања самог проблема, а потом ће бити објашњене и основе алгоритама за решавање неуравнотежености података примењених у овом раду.

5.1. Опис проблема

При обради означених база података, често се сусрећемо са ситуацијом где је једна од класа доминантнија у односу на остале. Ово је учестала појава у медицинским базама података, где је учесталост узорака неке болести знатно нижа, у поређењу са учесталошћу здравих пацијената [5]. Са слике 5.1.1. можемо приметити да је такав случај и са базом података коришћеном у овом раду. Узорци пацијената са дијабетесом представљају само 8,5% целе базе података. Оваква неравнотежа у подацима може довести до проблема јер алгоритми машинског учења често имају тенденцију да се фокусирају на већинску класу, која у овом случају представља здраве пацијенте. Ово може резултирати тиме да алгоритми не науче на адекватан начин да разликују узорке мање класе, која у медицинским базама често представља тражену дијагнозу.



Слика 5.1.1. Приказ неуравнотежености базе података

Постоје два приступа решавању овог проблема, смањење већинског узорка и синтетичко повећање класе са мање репрезентативних података, како би се обезбедила једнакост или приближна једнакост броја узорака из различитих класа. Ово омогућава алгоритмима да боље генерализују и да постижу равнотежу између тачности класификације и способности откривања случајева за обе класе. Ова равнотежа има посебан значај у медицинским апликацијама где тачна идентификација стања пацијента има кључни утицај на исправну дијагнозу и адекватну терапију.

5.2. Случајно смањење узорка

Ова техника укључује случајно уклањање узорака већинске класе, на приближан број узорака као што се налази у другој класи како би се постигла равнотежа у бази података. Предност овог приступа је у једноставности и брзини примене, након које и само тренирање алгоритама траје сразмерно краће количини података која је на располагању. Међутим, овом техником може доћи до губитка значајних информација, избацивањем узорака који их садрже из скупа података. Ако је почетни скуп података већ мали, примена ове методе повећава ризик од преприлагођавања (енгл. *overfitting*) зато што може узроковати недостатак репрезентације свих потребних варијација података.

5.3. *SMOTE* и *ADASYN* алгоритам

*SMOTE (*енгл. *Synthetic Minority Over-sampling Technique)* и *ADASYN (*енгл. *Adaptive Synthetic Sampling)* су алгоритми за решавање проблема неуравнотежености класа у скуповима података. Ови алгоритми се користе како би се створили синтетички примери мањинске класе, помажући тако у побољшању перформанси модела класификације. Овим приступом не губе се информације као код смањења већинског узорка, али могу довести до преприлагођавања додавањем шума у виду велике количине генерисаних података, уколико мањинска класа има мали број узорака и потребна је велика количина синтетичких података да би дошло до уравнотежености класа.

5.3.1. SMOTE

Овај алгоритам генерише примере за мањинску класу тако што узима случајне узорке те класе и комбинује их са њихових к-најближих суседа. Параметар за КНН у примени за овај алгоритам је обично постављен на вредност 5. Креира нови синтетички примерак тако што комбинује атрибуте случајног узорка са атрибутима његових суседа. На овај начин настаје

синтетички примерак који је близак са стварним примерима из мањинске класе, али ипак није идентичан већ постојећем узорку из базе података. Примена овог алгоритма може побољшати перформансе класификационих модела тако што побољшава равнотежу класа и смањује склоност модела према већинској класи.

5.3.2. ADASYN

ADASYN представља побољшање у односу на *SMOTE,* које се прилагођава расподели података. Не користи фиксан број суседа, већ га динамички прилагођава за сваки пример у мањинској класи на темељу његове густине. Густина се рачуна као број примера исте класе који се налазе унутар одређеног радијуса око сваког узорка. Прилагођавање на основу густине се врши и за број генерисаних синтетичких примера мањинске класе. Примери мањинске класе са мањом густином, који се налазе у ретким деловима са малом количином реалних узорака, ће имати већи број синтетички генерисаних података на основу њихових атрибута и атрибута њихових суседа. На овај начин се добије уравнотежена база података, где су узорци мањинске класе равномерно распоређени у простору својих атрибута.

6.

Резултати

У овом поглављу биће представљени кључни резултати нашег истраживања. Анализираћемо перформансе различитих модела класификације на описаном скупу података у поглављу 5. Приказати њихове резултате коришћењем одговарајућих метрика. Такође, детаљно ћемо приказати приступ решавању проблема неуравнотежености базе података и разлику у раду алгоритама на оригиналној бази података и на њеној уравнотеженој варијанти.

6.1. Метрика

За евалуацију резултата класификације коришћене су три различите метрике:

- Прецизност
- Одзив
- Ф1-мера

Свака од ове три метрике пружа јединствени увид у перформансе класификационих модела, омогућавајући дубље разумевање њихових способности и недостатака. Висока прецизност значи да модел прави мало лажно позитивних предикција, што је од суштинског значаја у медицинским апликацијама како бисмо избегли нетачне дијагнозе и непотребне медицинске поступке. Док је висок одзив кључан за правовремено откривање стварних случајева болести. Ф1-мера узима у обзир и вредност прецизности и одзива, и рефлектује целокупну ефикасност модела.

6.1.1. Прецизност

Прецизност је метрика која процењује тачност позитивних предвиђања модела. Даје информацију о томе колико је модел склон тачном класификовању позитивних примера. Рачуна се по следећој формули:

$$\mathrm{P} = \frac{Tp}{Tp+Fp} \qquad (6.1.1.)$$

где Tp представља број правих позитивних вредности, тј. узорака које је модел класификације означио као позитивне, и они то заиста и јесу, а Fp представља број лажно позитивних, модел је негативне вредности погрешно класификовао.

6.1.2. Одзив

Одзив је метрика која процењује способност модела да тачно идентификује све стварно позитивне примере. Пружа информацију о томе колико је модел ефикасан у откривању позитивних примера. Рачуна се по следећој формули:

$$O = \frac{Tp}{Tp+Fn} \tag{6.1.2}$$

где Tp представља број правих позитивних вредности и Fn представља број лажно негативних вредности.

6.1.3. Ф1-мера

Ф1-мера комбинује прецизност и одзив модела у једну вредност, и представља њихову хармоничну средину. Користи се када је потребно узети у обзир равнотежу између тачности позитивних предвиђања и способности модела да идентификује све позитивне примере. Рачуна се по формули 6.1.3. где P представља вредност прецизности, а O одзива.

$$F1 = 2 * \frac{P*O}{P+O} \tag{6.1.3.}$$

Ако модел има високу прецизност и низак одзив, то значи да алгоритам даје тачне позитивне предикције (правилно идентификује позитивне примере) приликом доношења одлука, али пропушта многе стварне позитивне примере, тј. не идентификује их. Сусреће се у сценарију када је алгоритам врло опрезан и прави мало лажно позитивних предикција (Fp) али зато пропушта многе стварно позитивне примере (Fn). Овакав резултат може бити прихватљив код неких случаја, али приликом дијагностиковања дијабетеса он значи да неким пацијентима није дата дијагноза дијабетеса, иако болују од те болести.

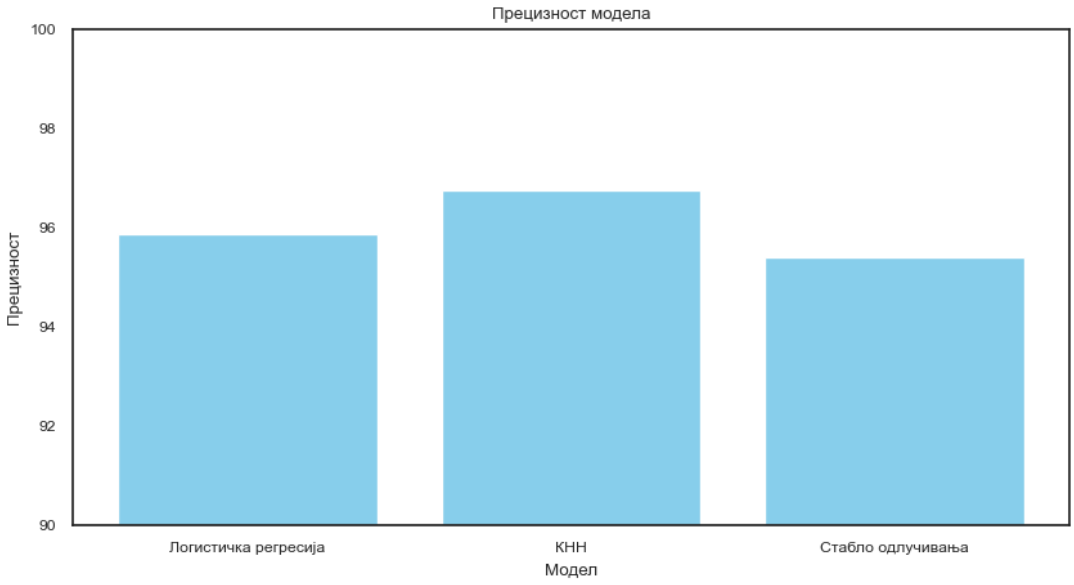
6.2. Опште информације о процесу тренирања и тестирања

Анализом базе података представљеном у трећем поглављу, одлучено је да се из улазних атрибута изоставе пол и историја пушења због слабе вредности корелације са излазним атрибутом класификације, тј. дијабетесом. Њихова вредност занемарљиво мало утиче на резултат саме класификације. Извршена је стандардизација вредности атрибута на опсег [0,1] како би се омогућило моделу да равноправно третира све улазне атрибуте, чиме се постиже боља генерализација и интерпретација резултата класификације. Извршена је подела улазних скупова података на тренинг скуп који се састоји од 70% почетног, и тест скуп. Имплементација је вршена у програмском језику Пајтон (енгл. *Python*) због своје

једноставности и широке примене у областима машинског учења. Коришћене су имплементације алгоритама за класификацију из библиотеке *sklearn* , као и *ADASYN* и *SMOTE* из библиотеке *imblearn* за потребе решавања проблема неуравнотежености базе података. Подаци из базе се налазе записани у *csv* фајлу.

6.3. Резултати пре решавања неуравнотежености базе

Након извршене селекције атрибута, улазни скуп податак се састоји од 8 атрибута и 100 000 узорака, код којих је 8500 од пацијената са дијабетесом. На слици 6.3.1. можемо видети прецизности алгоритама класификације на тест скупу добијеном поделом оваквих улазних података. Логистичка регресија је рачунала коефицијенте методом градијентног спуста у 2000 итерација. Док је код КНН алгоритма употребљена вредност параметра 9, добијена емпиријским путем на тренинг скупу ове базе података.



Слика 6.3.1. Приказ резултата на иницијалној бази података

Овакав приказ резултата рада алгоритама ствара нереалну перцепцију о њиховој ефикасности. Највећи изазов при оцењивању алгоритама за дијагностику дијабетеса је неуравнотеженост података, пошто има значајних разлика у броју података о пацијентима са дијагнозом дијабетеса у односу на пацијенте без дијагнозе. Ова неравномерна расподела може да доведе до лажних закључака о успешности алгоритама, јер алгоритми лакше постижу високу прецизност тако што једноставно предвиде да сви случајеви припадају већинској класи

(пацијентима без дијагнозе). Ово, међутим, не значи неопходно да су алгоритми добри у дијагностиковању дијабетеса код нових пацијената, пошто су пропустили многе реалне случајеве дијабетеса код пацијената из мањинске групе. Због тога је од суштинског значаја узети у обзир и друге метрике као што је одзив.

Табела 6.3.1. Резултати класификације на иницијалној базом

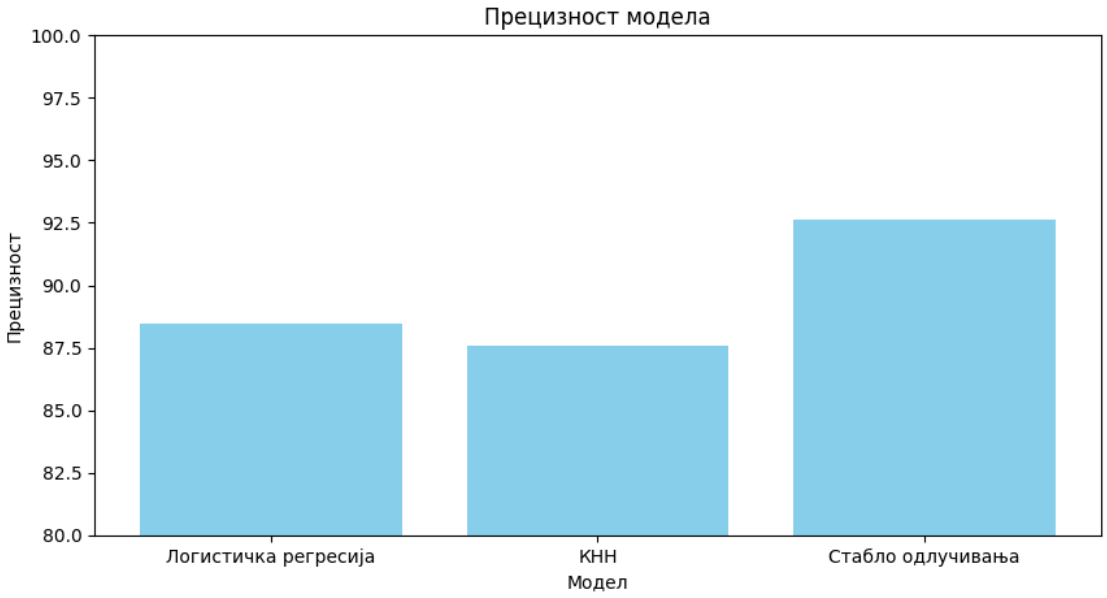
Алгоритми	Лабела	Прецизност[%]	Одзив[%]	Ф1-мера[%]
Логистичка регресија	0	96	99	98
	1	87	60	71
КНН	0	97	100	98
	1	96	65	77
Стабло одлучивања	0	98	97	97
	1	73	73	73

На основу добијених резултата из табеле 6.3.1. видимо јасну разлику у детекцији негативних и позитивних узорака. Све три метрике, сваког од алгоритама имају јако високе вредности за негативне податке, који представљају 91,5% улазне базе података, док њихова прецизност опада приликом детекције позитивних узорака, а значајна разлика се примећује код одзива алгоритма. Из свега наведеног закључујемо да су алгоритми пропустили да детектују велику количину позитивних података, враћају велики број лажних негатива, и да не препознају на прави начин разлике између података о здравим пацијентима и онима са дијагнозом.

6.4. Резултати на уравнотеженој бази података

База података коришћена у овом раду има изражен проблем неуравнотежености. Већинска класа је више од 10 пута богатија подацима, него мањинска. Као улазна база за логистичку регресију, употребљена је копија изворне базе код које је већинска класа насумично смањена на 11000 података, на овај начин и даље се база састоји од већег броја узорка здравих пацијената али је та разлика знатно смањена. За припрему улаза у КНН алгоритам и стабло одлучивања прво је извршено случајно смањење узорка оригиналне базе података, тако да се разлика у величини већинске и мањинске класе смањи на 2 пута. А након тога је извршено генерисање синтетичких узорака мањинске класе коришћењем *ADSAYN* алгоритма, као унапређене верзије *SMOTE* алгоритма. Прво је извршено смањивање разлике у величинама класа, зато што превелика количина синтетичких података може довести до лошег тренирања алгоритама и самим тим лошијих резултата предикције. Прецизност детекције модела након примене ова два приступа решавања проблема неуравнотежености можемо

видети на слици 3.4.1. Параметар логистичке регресије је остао непромењен, док је у случају КНН алгоритма емпиријски на новом редукованом скупу података одабрана вредност 11, за параметар к.



Слика 6.4.1. Приказ резултата на уравнотеженој бази података

Са слике 6.4.1. примећујемо пад прецизности алгоритама, у односу на слику 6.3.1. Оваква прецизност даје реалнију слику одзива алгоритама у класификацији нових случаја дијабетеса. Више не постоји та изражена неуравнотеженост базе података, па алгоритми не добијају високу прецизности простим класификовањем већине података у већинску класу.

Табела 6.4.1. Резултати класификације на уравнотеженој бази

Алгоритми	Лабела	Прецизност[%]	Одзив[%]	Ф1-мера[%]
Логистичка регресија	0	89	91	90
	1	88	85	86
КНН	0	98	83	90
	1	75	96	84
Стабло одлучивања	0	97	92	94
	1	85	95	90

Из табеле 6.4.1. можемо приметити да након балансирања базе података, одзив за позитивне вредности сва три алгоритма је порастао, у односу на резултате из табеле 6.3.1. Ово повећање одзива указује на то да алгоритми реагују боље и успешније идентификују позитивне примере, што смањује број лажно негативних класификација. У контексту медицинске дијагнозе, ово

представља значајан напредак, пошто смањује шансе да се случајно пропусте стварни случајеви дијабетеса.

У случају КНН алгоритма може се приметити пад у прецизности класификације позитивних узорака. Ово сугерише да је у ситуацији са смањеним бројем узорака о здравим пацијентима долази до давања лажно позитивне класификације, односно дијагнозе дијабетеса, али је одзив знатно већи и смањен број пропуштених позитивних дијагноза, код болесних пацијената. Све наведено у контексту коришћења овог алгоритма приликом дијагностиковања дијабетеса доводи у питање да ли је боље дати лажно позитивну или лажно негативну дијагнозу.

Код друга два алгоритма уочава се значајан напредак у дијагностиковању мањинске класе у свим релевантним метрикама, праћен благим смањењем перформанси при класификацији здравих пацијената. Ови резултати пружају реалнију слику о раду ових алгоритама. Детаљније, резултати указују на значајан пораст одзива за мањинску класу, што значи да су ови алгоритми постали много бољи у идентификацији стварних позитивних случаја дијабетеса. Овај пораст одзива иде у корист пацијената са дијабетесом, јер се смањује вероватноћа да ће њихови случајеви остати неоткривени. Са друге стране, благи пад у перформансама при класификацији здравих пацијената, може указивати на то да алгоритми постају опрезнији и мање склони давању лажно позитивних дијагноза.

7. ЗАКЉУЧАК

Циљ овог рада је био анализа нове базе података, њених атрибута, предности и недостатка у односу на широко коришћену базу *PIMA* у претходним истраживањима на ову тему. Примењени су и тестирани други алгоритми за дијагностиковање дијабетеса машинским учењем. Дат је и преглед различитих приступа решавању неуравнотежености базе и његова комбинација са испитаним алгоритмима класификације.

Значај проучавања оваквих проблема, лежи у аутоматизацији дијагностиковања истих. Аутоматизовани дијагностички алати омогућавају брже препознавање дијабетеса код пацијената. То значи да се лечење може отпочети раније, чиме се смањује ризик од компликација и побољшава здравствено стање пацијената. Ово истраживање представља класичан пример примене вештачке интелигенције у медицинским апликацијама, чиме се отварају врата за разноврсне могућности и унапређења здравствене неге. Кроз оваква нова истраживања, можемо идентификовати недостатке у постојећим методама дијагностиковања дијабетеса и радити на њиховом систематичном побољшању. Овај приступ омогућава медицинским професионалцима боље разумевање проблема и употребу технологије за унапређивање дијагноза и третмана.

Поред тога, истраживање такође указује на потребу за прикупљањем балансиране базе података која ће представљати адекватан темељ за обуку алгоритама. Неравнотежа у подацима може значајно утицати на перформансе модела, стога је скуп података који тачно одражава све стварне варијације резултата пацијената са и без дијагнозе од суштинског значаја. А опет да садржи равнотежну количину узорака пацијената, припадника обе класе.

Резултати овог истраживања су показали да је стабло одлучивања исплативо размотрити као преферирани алгоритам за дијагностиковање дијабетеса, посебно када се користи на балансираној бази података. Овај алгоритам показује конзистентно високе вредности прецизности и одзива за обе класе, што га чини поузданим за широк спектар пацијената. Међутим, важно је напоменути да постоје специфичне подкатегорије где су други алгоритми показали боље перформансе. Ово указује на потребу за приступом базираним на више алгоритама, где се одређени алгоритам може применити на основу карактеристика пацијената или специфичних захтева за дијагностику.

Одлука о избору алгоритама за дијагностиковање дијабетеса треба бити пажљиво разматрана у зависности од карактеристика података и потреба пацијената. Стабло одлучивања се издваја као солидан избор за општу употребу, док се други алгоритми могу применити селективно у одређеним контекстима. Даље истраживање и прилагођавање алгоритама могу допринети унапређењу процеса дијагностике дијабетеса и побољшању здравствене неге пацијената.

Литература

[1] A. Mujumard, Dr. Vaidehi V. „Diabetes prediction using Machine Learning Algorithms“, *International Conference on recent trends in advanced computing (ICRTAC)* 2019.

[2] Sakshi Gujral „Early Diabetes Detection using Machine Learning: A Review“, *International Journal for Innovative Research in Science & Technology (IJIRST)*, March 2019.

[3] V. Anuja Kumari, R. Chitra „Classification Of Diabetes Disease Using Support Vector Machine“, *International Journal of Engineering Research and Applications (IJERA)*, March-April 2013.

[4] F. Mercaldo et al, „Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques“, *International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES2017)*, September 2017.

[5] M.T. García-Ordás at al, „Diabetes detection using deep learning techniques with oversampling and feature augmentation“, *Computer Methods and Programs in Biomedicine*, January 2021.

[6] International Diabetes Federation [Online]. Available: https://idf.org/ (18.09.2023.)

[7] D. Haire-Joshu et al, „Smoking and Diabetes“, *Diabetes Care Magazine*, 22(11): 1887-1898, Nov. 1999.

[8] Centers for Disease Control and Prevention Available: https://www.cdc.gov/healthyweight/assessing/bmi/ (17.09.2023.)

[9] K.M.V. NARAYANet al, „Effect of BMI on Lifetime Risk for Diabetes in the U.S.” *Diabetes Care Magazine*, 30(6): 1562–1566, June 2007.

Списак слика

СПИСАК ТАБЕЛА