Univerzitet u Beogradu
Elektrotehnički fakultet

Ива Перић, 2019/0392

Предвиђање вируса КОВИД-19 применом алгоритама за машинско учење помоћу ВЕКА софтвера

Пројекат из предмета Принципи модерних телекомуникација

Ментор:
проф. др Милан Бјелица

Београд, септембар 2022.

Сажетак

Рана дијагноза је кључна за спречавање развоја болести која може изазвати опасност на људске животе. КОВИД-19, који је заразна болест која је мутирала у неколико варијанти, постао је глобална пандемија која захтева да се дијагностикује што је пре могуће. Током израде овог пројекта, коришћено је неколико алгоритама машинског учења у изградњи модела за анализу и предвиђање присуства корона вируса помоћу скупова података о симптомима и присуства вируса. Ти алгоритми су: **J48 Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors и Naïve Bayes**. Алгоритми су примењени преко ВЕКА софтвера за машинско учење. Перформансе сваког модела су процењене коришћењем 10-струке унакрсне валидације и упоређивањем према главним мерама тачности, исправно или неисправно класификованим инстанцама, капа статистици, средњој апсолутној грешци и времену потребном за изградњу модела. Резултати показују да Support Vector Machine (SVM) надмашује друге алгоритме постижући тачност од 98,81% и средњу апсолутну грешку од 0,012.

Кључне речи: алгоритми, болест, КОВИД-19, LaTeX 2_ε, машинско учење, пандемија, предикција

Садржај

Списак слика

Списак табела

Глава 1

Увод

1.1 Предмет рада

Ковид-19 је заразна болест која утиче на респираторни систем особе. Људима са јаким имунитетом неће бити потребан посебан третман лечења, али то није случај са старијим особама, кардиоваскуларним болесницима, дијабетичарима, људима са респираторним обољењима, итд. КОВИД-19 се шири капљичним путем преко говора, кашљања и кихања, или чак додиривања неких контаминираних предмета или подручја. Светска здравствена организација (WHO) [1] навела је да често прање руку, дезинфекција, социјално дистанцирање, ношење маски и недирање лица могу заштитити особу од вируса. Око 10–15% пацијената са КОВИД-19 имају тешке симптоме. Неколико симптома који су најчешћи су грозница, сув кашаљ и умор, док су ређи симптоми главобоље, бол у грлу, дијареја, коњуктивитис, повређање и губитак мириса. Озбиљни симптоми су проблеми са дисањем, бол у грудима и губитак говора и покрета.

1.2 Мотивација

До 14. августа 2022. у свету је било 585.950.085 случајева КОВИД-19 и 6.425.422 умрлих. Како вирус наставља да се шири, може довести до повећаних потреба за болничким ресурсима и недостатка медицинске опреме и тестова на КОВИД-19. Ограничен приступ тестовима може ометати рану дијагнозу болести. Зато је неопходан систем предвиђања који има за циљ да утврди присуство вируса код особе.

1.3 Методологија

Алгоритми за класификацију машинског учења, скупови података и софтвер за машинско учење су алати за дизајнирање модела предвиђања КОВИД-19.

Машинско учење се може категорисати као supervised, unsupervised и учење са појачањем (reinforcement learning). Supervised машинско учење је приступ који обучава машину користећи означене скупове података, при чему су примери означени према класи којој припадају. Машина ће анализирати дате податке и на крају предвидети нове случајеве на основу информација које је научила из прошлих података. Машинско учење без надзора (unsupervised) учи само без присуства исправно означених података. Машина ће се хранити узорцима за тренинг, а посао машине је да одреди скривене обрасце из скупа података. За учење са појачањем, машина има за циљ да открије најприкладније акције кроз приступ покушаја и грешке и посматрање у окружењу. Сваки пут када машина успешно изврши задатак, биће награђена повећањем свог стања; у супротном ће бити кажњена смањењем свог стања, а овај приступ ће се понављати неколико пута док машина не научи како да правилно изврши одређени задатак.

Неколико метода машинског учења је коришћено у изградњи модела за предвиђање болести (коронарна артеријска болест, респираторна болест, рак дојке, дијабетес, деменција и болест масне јетре). У те сврхе коришћени су следећи алгоритми: J48 Decision Tree (**J48 DT**), Random Forest (**RF**), Support Vector Machine (**SVM**), K-Nearest Neighbors (**k-NN**), and Naïve Bayes (**NB**).

Неколико supervised метода машинског учења је коришћено у изградњи модела предвиђања болести и алгоритам може да ради другачије у зависности од скупа података. Овај пројекат има за циљ да изгради модел који може аутоматски да предвиди присуство КОВИД-19 код особе која користи J48 DT, RF, SVM, k-NN, и NB алгоритме, анализирајући симптоме КОВИД-19 користећи ВЕКА софтвер отвореног кода.

ВЕКА [2] је колекција алгоритама за машинско учење који подржава задатке рударења података пружајући широк спектар алата који се могу користити за претходну обраду података, класификацију, груписање, регресију, асоцијацију и визуелизацију.

Глава 2

Материјали и методе

2.1 Прикупљање података

За прикупљање података користила сам скуп података доступан у Kaggle-у под називом *COVID-19 Symptoms and Presence* [3]. Овај скуп података има 20 атрибута који су могући фактори у вези са добијањем вируса и 1 атрибут класе који одређује присуство КОВИД-19.

Attribute Name	Type	Percentage Level	Description
Breathing Problem	Nominal	10%	The person is experiencing shortness of breath.
Fever	Nominal	10%	Temperature is above normal.
Dry Cough	Nominal	10%	Continuous coughing without phlegm.
Sore Throat	Nominal	10%	The person is experiencing sore throat.
Running Nose	Nominal	5%	The person is experiencing a runny nose.
Asthma	Nominal	4%	The person has asthma.
Chronic Lung Disease	Nominal	6%	The person has lung disease.
Headache	Nominal	4%	The person is experiencing headache.
Heart Disease	Nominal	2%	The person has cardiovascular disease.
Diabetes	Nominal	1%	The person is suffering from or has a history of diabetes.
Hypertension	Nominal	1%	Having a high blood pressure.
Fatigue	Nominal	2%	The person is experiencing tiredness.
Gastrointestinal	Nominal	1%	Having some gastrointestinal problems.
Abroad Travel	Nominal	8%	Recently went out of the country.
Contact with COVID-19 Patient	Nominal	8%	Had some close contact with people infected with COVID-19.
Attended Large Gathering	Nominal	6%	The person or anyone from their family recently attended a mass gathering.
Visited Public Exposed Places	Nominal	4%	Recently visited malls, temples, and other public places.
Family Working in Public Exposed Places	Nominal	4%	The person or anyone in their family is working in a market, hospital, or another crowded place.
Wearing Masks	Nominal	2%	The person is wearing face masks properly.
Sanitation from Market	Nominal	2%	Sanitizing products bought from market before use.
COVID-19	Nominal	-	The presence of COVID-19.

Слика 2.1: *Описи карактеристика, процентуалног нивоа и атрибута скупа података о симптомима и присутности КОВИД-19 (преузето са сајта Kaggle)*

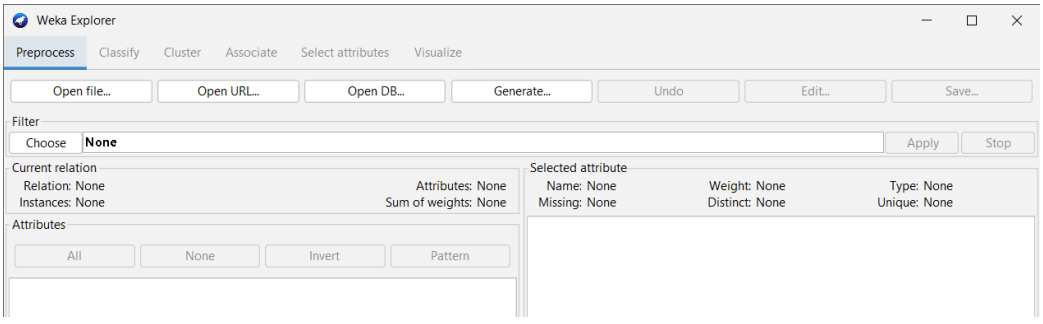
2.2 Обрада података

За обраду података користила сам ВЕКА софтвер за машинско учење и то Explorer и Knowledge Flow опције.

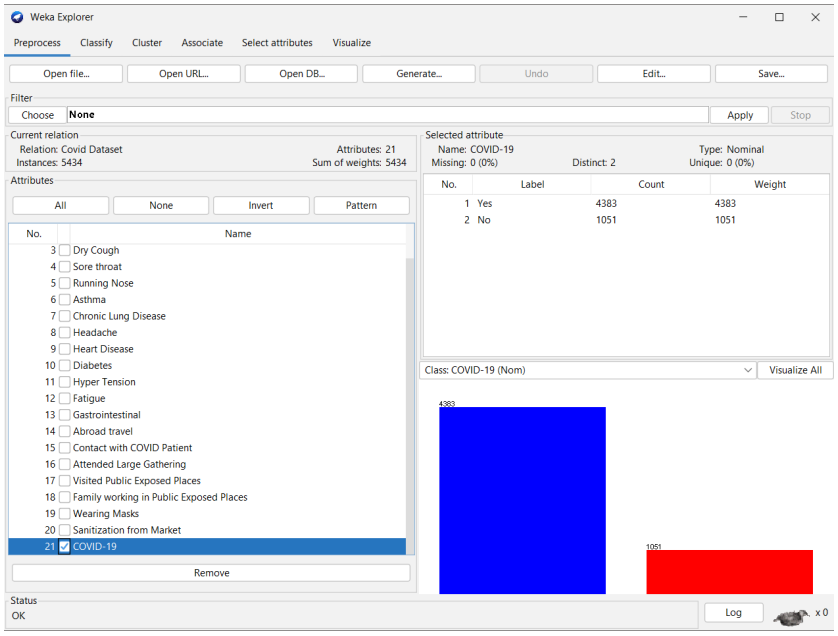
Слика 2.2: *Главна страница Веке*

Постоје различити формати докумената прихваћени у Веки, као што су arff, csv, JSON file, а скуп података о симптомима и присутности КОВИД-19 који сам користила је у csv формату, што олакшава увоз у софтвер и анализу.

Обрада података може се покренути кликом на open file и претраживањем локације скупа података. Када се скуп података увезе у Веку, биће приказана тренутна релација, атрибути, инстанце и збир визуелизација које се односе на скуп података.

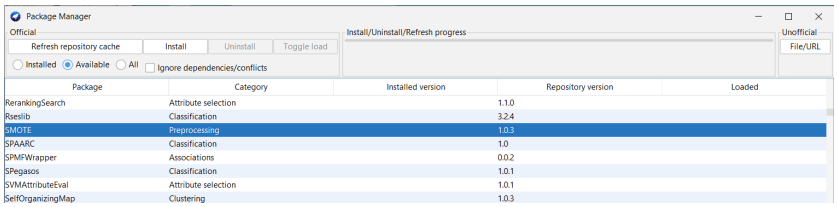


Слика 2.3: *Изглед почетног стања пре увоза података у апликацији Explorer.*

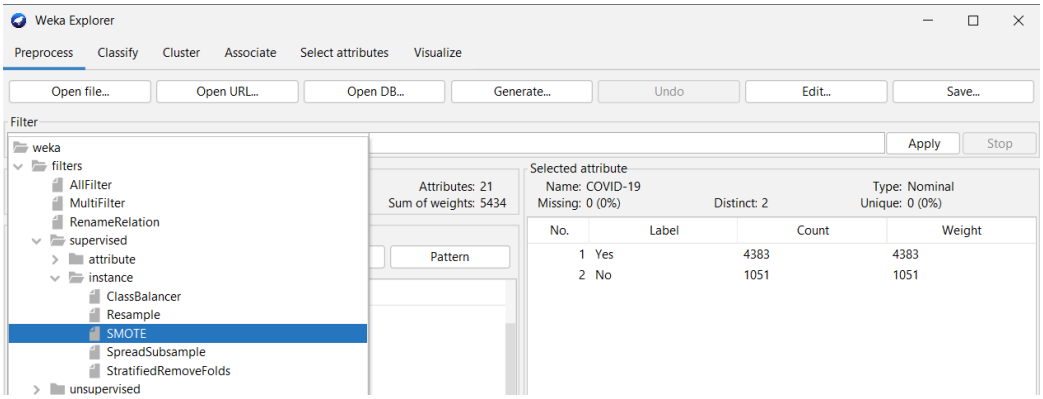


Слика 2.4: *Након одабира атрибута COVID-19.*

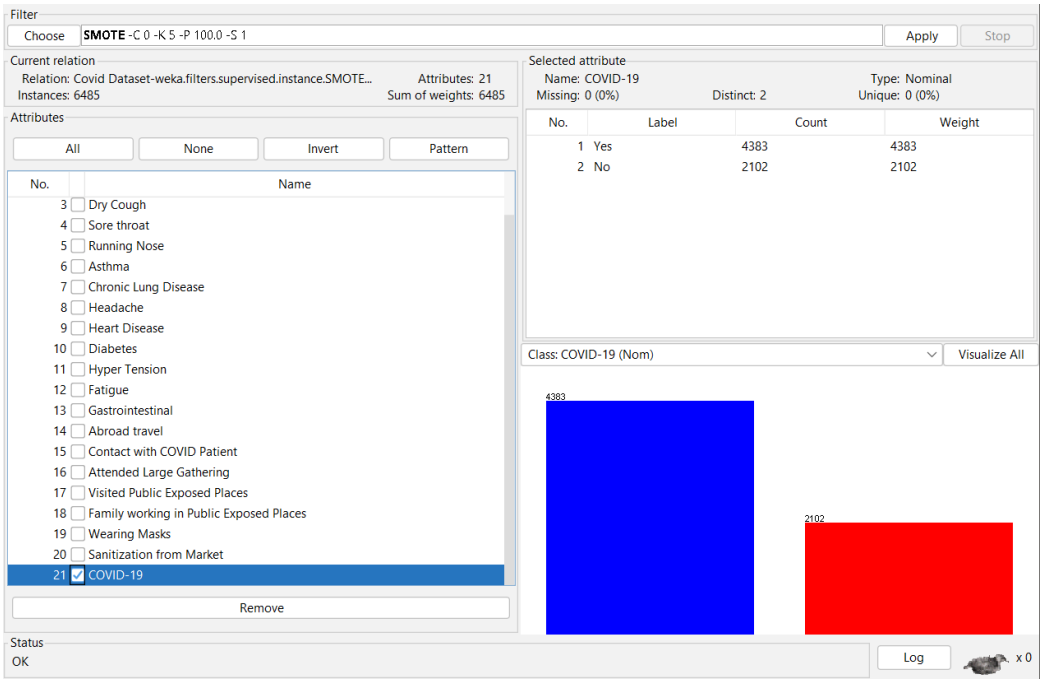
Када је атрибут изабран, name, missing, distinct и unique постаће видљиве. Кликом на атрибут COVID-19, појављује се графикон. За скуп података о симптомима и присутности КОВИД-19, класе имају неравнотежу класа 4:1. Коришћена је техника синтетичког предузорковања (SMOTE) да би се генерисале додатне инстанца за мањинску групу (класа означена са „Не“ која говори колико има негативних). SMOTE предузоркује мањинску класу генерисањем додатних синтетичких узорака и на тај начин ће се повећати класа са мање узорака. Комбиновање методе предузорковања мањинске класе и подузорковања већинске класе, или одсецања неких узорака у класи која садржи више узорака, даће боље перформансе класификатора него само подузорковање већинске класе.



Слика 2.5: *Инсталација SMOTE помоћу Package Manager-a у Tools менију.*



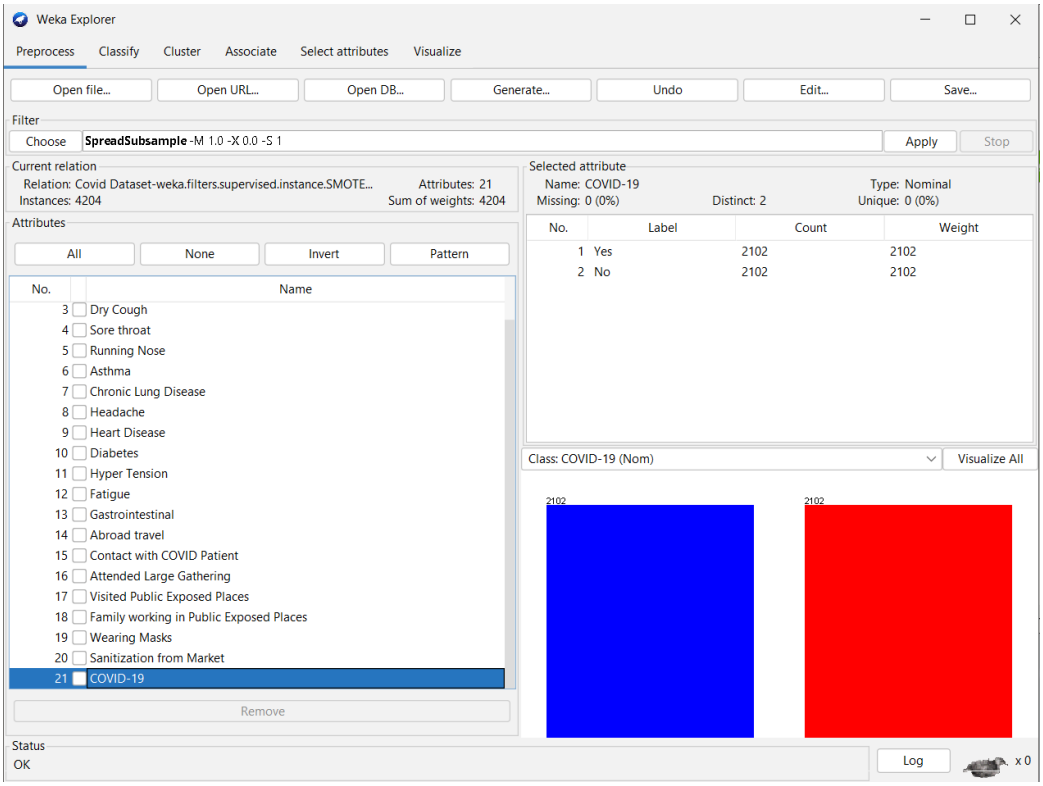
Слика 2.6: *Проналазак пакета и одабир филтера.*



Слика 2.7: *Стање графикона након SMOTE филтера.*

Након одабира SMOTE филтера (слика 2.7), црвена трака је повећана, што указује да су додати додатни узорци у класу „Не“. Слика приказује балансирање скупа података.

Сада када је мањинска класа добила свој број, скуп података је још увек неуравнотежен. Да би се то решило користимо Spread Subsample филтер(шири подузорак) како бисмо смањили број у већинској класи и изједначили га са мањинском класом. Он се такође налази у Filters->Supervised->Instance. Ажурирани бројеви класа у скупу података могу се видети на слици 2.8. Висине плаве и црвене траке су сада једнаке, што значи да обе класе имају исте вредности. Важно је користити уравнотежен скуп података како би класификатор био добро информисан о обе класе које треба предвидети, као и да се избегне пристрасност дистрибуције. Додатно је било потребно distributionSpread у поставци Spread Subsample поставити на 1.0.



Слика 2.8: *Скуп података након имплементације техника SMOTE и Spread Subsample.*

2.3 Моделовање

Након обраде података коришћењем техника SMOTE и Spread Subsample, неколико модела је направљено коришћењем WEKA Експлорер модула користећи различите supervised алгоритме машинског учења: J48 DT, RF, SVM, k-NN и NB. На картици Classify у WEKA Експлореру, бира се назив класификатора и десетоструку унакрсну валидацију. Да би се одредила најбоља конфигурација за сваки алгоритам, извршава се оптимизацију хиперпараметара изводећи неколико тренинга са истим скупом података за сваки алгоритам. Користи се десетоструко тестирање унакрсне валидације и величина серије од 100 за све експерименте.

За **J48 DT**, користи се 2 као минимални број инстанци по листу за све процесе тренинга. Користи се confidence factor for pruning, а unpruned (тачно или нетачно) као параметар који треба подесити.

За **RF**, поставиља се максимална дубина на нулу, односно неограничену дубину, а број итерација на 100, што је број стабала у шуми. Величина торбе (Bagsize) је параметар који се подешава да би се одредила најбоља величина торбе која ће дати добре резултате.

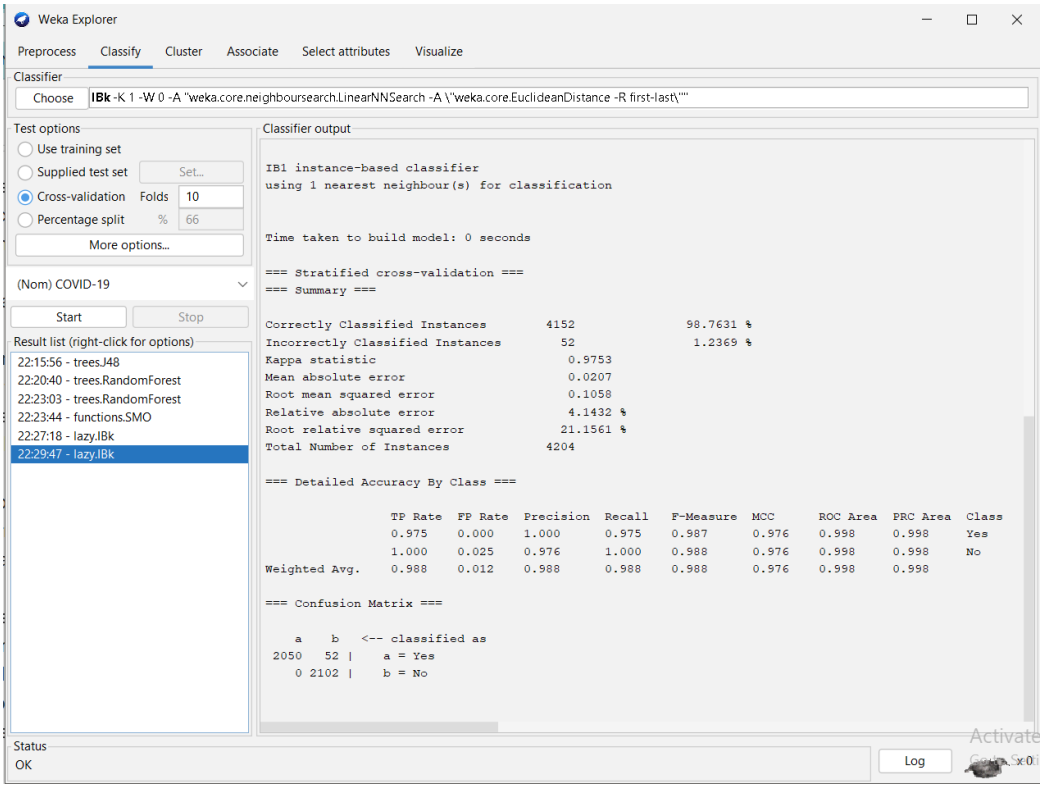
За **SVM** се користе хиперпараметри C и Kernel. C је параметар сложености, који одређује колико је флексибилан процес у одређивању линије која раздваја класе, а Kernel одређује како ће подаци бити раздвојени (линијом, закривљеном линијом, полигоналом итд).

Приликом изградње модела помоћу **k-NN** алгоритма, користи се функција удаљености Еуклидска удаљеност, а подешени хиперпараметри су KNN или број суседа за употребу и параметар Cross-Validate, који указује да ли ће се користити унакрсна провера у одређивању најбоље k вредности.

На крају, за **NB** алгоритам, хиперпараметри који се користе су да ли ће процес користити Kernel Estimator и Supervised Discretization, а не нормалну дистрибуцију нумеричких атрибута.

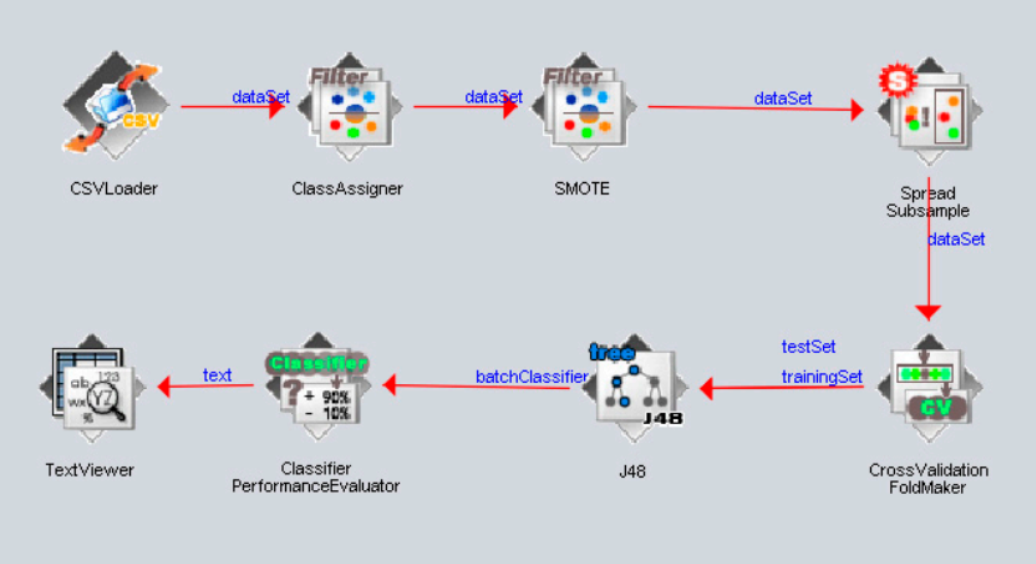
Сваки алгоритам пролази неколико експерименталних тренинга. За сваки тренинг, перформансе развијеног модела су приказане у одељку излаза класификатора. Пример картице класификатора са детаљима о развијеним моделима приказан је на слици 2.9. Када се кликне на дугме за одабир, биће приказани доступни класификатори алгоритма машинског учења. Биће представљено неколико фасцикли са листом алгоритама који ће се користити. Када је изабран жељени алгоритам машинског учења, потребно је попунити опцију теста.

Коришћена је 10-струка унакрсна валидација. Непосредно испод одељка са опцијама теста, изабран је атрибут класе или атрибут који треба предвидети. У овом случају, атрибут „COVID-19“ је атрибут класе. Да би се подесили хиперпараметри, мора се кликнути на ознаку поред дугмета за одабир како би се омогућило корисницима да унесу жељене конфигурације. На крају, потребно је кликнути на дугме за покретање да би се започео процес тренинга за изградњу модела.



Слика 2.9: *Детаљи у вези са перформансама развијеног модела приказани су у одељку излаза класификатора.*

Да би се обезбедио визуелни приказ изградње модела, осмишљен је дијаграм тока знања помоћу Knowledge Flow Module. Ток знања у изградњи модела приказан је на слици 2.10.



Слика 2.10: *Дизајн процеса од учитавања скупа података, до тренинга и тестирања помоћу алгоритама машинског учења и класификације перформанси помоћу WEKA Knowledge Flow модула.*

CSVLoader чита скуп података у формату раздвојеном зарезима, а затим атрибут треба означити као атрибут класе, ово се може урадити коришћењем додељивача класе (class assigner). Технике SMOTE and Spread Subsample су коришћене за балансирање скупа података, а да би се извршила унакрсна провера на тесту и скуповима за тренинг, мора се користити cross-validation fold maker. Следећи процес је J48 DT, који је један од supervised алгоритама машинског учења. Овде се knowledge flow наставља до евалуатора перформанси класификатора (classifier performance evaluator) да процени развијени модел, а резултати перформанси се могу проверити коришћењем Text Viewer-a. Овај ток знања ће се користити за све алгоритме коришћене овде, а сви резултати перформанси биће забележени како би се користили у упоредној анализи, што је следећа фаза пројекта.

2.3.1 Коришћени алгоритми за supervised машинско учење

J48 Decision Tree

Стабло одлучивања је алгоритам који производи графичку структуру налик стаблу, при чему су инстанце класификоване коришћењем

основног чвора који има тестни услов (нпр да ли особа има упалу грла или не) и грана које одређују одговор или ознаку. Дрво може бити или листни чвор или тестни чвор. Листни чвор представља класу којој припадају све инстанце. Ако инстанце припадају различитим класама, назива се тестни чвор, који се састоји од услова који се додаје вредности атрибута, а тестни чвор може даље бити представљен у два или више подстабала. Једна врста алгоритма стабла одлучивања је J48 DT, који је уобичајен и једноставан алгоритам стабла одлучивања који се користи у сврхе класификације; користи приступ *завади па владај*, који дели инстанце у подопсеге на основу вредности атрибута.

Random Forest

Као и DT, RF алгоритам такође производи стабло, али за овај алгоритам, неколико стабала ће бити генерисано из вредности случајних узорака у скупу података, а коначни резултат ће бити заснован на резултатима већине развијених стабала. RF доноси значајна побољшања у тачности класификације модела кроз изградњу група стабала које појединачно генеришу резултате, упоређујући те резултате и бирајући која класа је добила највише гласова.

Naïve Bayes

Наивни Бајесов класификатор је статистички supervised алгоритам машинског учења који предвиђа вероватноће чланства у класи. NB постиже високу тачност и брзину када се примени на велики скуп података, али такође функционише веома добро у малим скуповима података. NB се заснива на Бајесовој теореми која се може дефинисати на следећи начин:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

где се P(A) назива претходна вероватноћа, која означава вероватноћу да се A деси, а P(B) се назива маргинална вероватноћа, која означава вероватноћу да се деси B. Вероватноће A и B су биле независне вредности. Затим, P(A|B) се назива постериорна вероватноћа, што је вероватноћа да се A деси ако се B догодио. Најзад, P(B|A) се назива вероватноћа вероватноће, што означава вероватноћу да се B

деси ако је A тачно. Бајесова теорема [6] израчунава постериорну вероватноћу дељењем производа вероватноће и претходне вероватноће са маргиналном вероватноћом. Наивни Бајесов алгоритам не зависи од присуства других параметара и зато се назива наиван. Да би се постигла највећа вероватноћа међу израчунатим резултатима, коришћена је следећа формула:

$$A = argmax_A P(A) \prod_{i=1}^{n} P(B_i|A)$$

Support Vector Machine

Може постићи добре перформансе у генерализацији без потребе за претходним знањем или искуством. SVM алгоритам користи хиперраван која раздваја инстанце, стављајући исте класе у исту поделу док максимизира растојање сваке групе од хиперравни која дели.Хиперраван се користи за минимизирање грешака које настају при раздвајању инстанци према њиховим одговарајућим класама.

k-Nearest Neighbors

k-NN класификује дату инстанцу преко већине класа међу својим k-најближим суседима који се налазе у скупу података. Овај алгоритам се ослања на метрику удаљености која се користи за одређивање најближих суседа дате инстанце, а најчешће коришћена метрика је Еуклидска удаљеност [7], која се изражава следећом формулом: $x = (a_1, a_2, a_3 \ldots$

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} w_r \left(a_r(x_i) - a_r(x_j)\right)^2}$$

$a_n)$, n је број атрибута примера, a_r је r-ти атрибут и његова тежина се назива w_r, и $d(x_i, x_j)$ су два примера. Да би се израчунала ознака класе примера, користи се следећа формула: где је d_i пример помоћу

$$y(d_i) = \underset{k}{arg\ max} \sum_{x_j \in kNN} y(x_j, c_k)$$

којег ће алгоритам одредити класу којој припада, термин k_j је један од k-NN присутних у скупу података, а $y(x_j, c_k)$ указује да ли k_j припада класи c_k. Резултат једначине је класа која има највише чланова k-NN, а такође је и класа којој припада пример изнад. Еуклидско растојање се углавном користи као подразумевано растојање у k-NN класификацији или k-means груписању за одређивање „k closest points" примера.

2.3.2 Упоредна анализа

Извршавамо упоредну анализу перформанси различитих supervised алгоритама машинског учења користећи 10-струко тестирање унакрсне валидације, а важни критеријуми који се користе у овој фази су следећи.

Тачност/прецизност

Тачност је мерење свих тачно предвиђених инстанци у односу на укупна предвиђања направљена од стране модела Bagging Predictors, а сваки алгоритам може да ради другачије у односу на тачно класификоване инстанце. Прецизност израчунава однос тачно класификованих инстанци које су истинито позитивне (TP) и истинито негативне (TN) у односу на укупан број предвиђања, укључујући TP и TN и нетачна предвиђања односно лажно позитивна (FP) и лажно негативна (FN). Тачност се може израчунати коришћењем следеће формуле:

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

Штавише, главне мере тачности су укључене у упоредну анализу, као што су прецизност (precision), опозив (recall) и F-Measure [8]. Прецизност мери тачност предвиђања TP над свим предвиђеним позитивним вредностима тако што се TP подели збиром TP и FP. Према датом опису, прецизност значи колико је оних који су класификовани као позитивни на COVID-19 заправо позитивни на COVID-19, а може се израчунати помоћу ове формуле:

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

Опозив (recall) мери тачност предвиђања TP у односу на стварне позитивне инстанце у скупу података. Recall одговара на питање: од свих случајева који су позитивни на COVID-19, колико их је тачно предвидео модел? Проценат опозива се може добити на следећи начин:

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

Пошто precision и recall мере различите ствари, вредност F-Measure мери хармонију, баланс два критеријума. Насупрот томе, резултат F-Measure ће опасти ако се један критеријум побољша на рачун другог. Формула:

$$F - Measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Тачно и нетачно класификоване инстанце

Ове вредности су такође узете у обзир у упоредној анализи алгоритама машинског учења. Резултат тачно класификованих инстанци је збир TP и TN предвиђања; обрнуто, резултат нетачно класификованих инстанци је збир FP и FN предвиђања модела.

Kappa статистика

Cohen-ова Kappa статистика израчунава поузданост резултата између два оцењивача исте ствари; то је колико се оцењивачи случајно слажу. Нулта оцена значи да постоји насумична или мања сагласност између два оцењивача, а резултат може бити мањи од нуле, док оцена 1 указује на потпуно слагање. Може се израчунати коришћењем следеће формуле где је P_o вероватноћа слагања, а P_e вероватноћа случајног слагања између оцењивача:

$$K = \frac{P_o - P_e}{1 - P_e}$$

Средња апсолутна грешка (MAE)

За процену перформанси модела, MAE [9] се користи за мерење количине погрешних класификација или грешака у предвиђању модела. MAE је просек свих апсолутних грешака. Он одређује колико је блиска предвиђена вредност стварној вредности у скупу података. MAE се може добити по следећој формули:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

где n представља укупан број грешака, $\sum$ је симбол сумирања, x_i је предвиђена вредност, а k је стварна вредност.

Време потребно за израду модела

Време потребно за изградњу модела се изражава у секундама. Ова вредност приказује количину времена потребног за тренинг модела, што је неопходно да би се открило који модел ће бити најбржи.

2.3.3 Проналажење најбољег модела

Само тачност није довољна да се изабере најбољи модел који ће се користити, већ се морају узети у обзир и други резултати перформанси модела. Тачност најбоље функционише ако је скуп података симетричан или има близак или једнак број узорака по класи. Уместо упоредне анализе, критеријуми за проналажење модела који ће послужити као најприкладнији алгоритам машинског учења који ће се користити у изградњи предиктора присуства КОВИД-19 су следећи:

- Највећа тачност, прецизност, опозив и F-measure;
- Највише исправно класификоване инстанце;
- Најниже неисправно класификоване инстанце;
- Највиши kappa статистички резултат;
- Најнижа средња апсолутна грешка;
- Најкраће време потребно за изградњу модела

Глава 3

Резултати

3.1 Скуп података о симптомима и присутности КОВИД-19

Коришћен скуп података се састоји од 20 атрибута и 1 атрибута класе. Скуп података има 5434 инстанце, од којих 81% или 4383 случаја припада Да класи, што указује на то да особа има КОВИД-19, док 19%, или 1051 инстанца, припада Не. Пошто је скуп података неуравнотежен, примењен је SMOTE за прекомерно узорковање или повећање мањинске класе и техника Spread Subsample за подузорковање мањинске класе . У табели 3.1 постоје три генерисана скупа података. Први је оригинални скуп података. Након имплементације SMOTE-а, инстанце мањинске класе су удвостручене генерисањем синтетичких узорака, а генерисани скуп података је други скуп података у табели. На крају, примењена је техника Spread Subsample која подузоркује класу већине да би скуп података био избалансиран, који се налази у трећем скупу у табели. Имплементацијом техника SMOTE и Spread Subsample, скуп података је постао избалансиран, а затим коришћен за извођење процеса тренинга и тестирања како би се направио предиктор КОВИД-19.

Tabela 3.1: Резултати техника SMOTE и Spread Subsample.

Број	Техника	Инстанце Да класе	Инстанце Не класе
1	-	4383	1051
2	SMOTE	4383	2102
3	Spread Subsample	2102	2102

3.2 Резултати моделовања

3.2.1 Оптимизација хиперпараметара

Овде се упоређују supervised алгоритами машинског учења како би се утврдило који је најприкладнији алгоритам који ће се користити у развоју предиктора КОВИД-19. Међутим, оптималне перформансе алгоритма се могу постићи ако се најбоља конфигурација користи у процесу моделирања. Због тога се извршава оптимизација хиперпараметара како би се одредиле вредности на којима ће алгоритам бити најбољи. За процену перформанси модела, 10-струка унакрсна валидација и величина серије (batch siz) од 100 коришћени су за све експерименте. За алгоритам **J48 DT**, користили смо confidence factor и опције Unpruned у ВЕКИ.

Tabela 3.2: Резултати оптимизације хиперпараметара алгоритма J48 DT користећи 2 као минимални број инстанци присутних у листу.

Број тренинга	Confidence factor	Unpruned	Тачност
1	**0.25**	**True**	**98.57%**
2	0.50	True	98.57%
3	0.75	True	98.57%
4	0.25	False	98.45%
5	0.50	False	98.55%
6	0.75	False	98.55%

Коришћена је заједничка вредност за минимални број случајева који се могу наћи на сваком листу, а то су 2 инстанце. Изведено је шест експерименталних тренинга за DT: за прва три тренинга параметар Unpruned је постављен на True, а фактори поверења су били 0.25, 0.50 и 0.75, респективно. У овом експерименту, перформансе модела су биле 98,57% за све случајеве. Исте вредности фактора поверења коришћене су у последња три тренинга са параметром Unpruned постављеним на False, а перформансе модела су достигле 98,45% за фактор поверења 0.25 и 98,55% за 0.05 и 0.75. Зато Unpruned треба поставити на True, што значи да се не врши орезивање. Пошто су перформансе модела биле исте за све факторе поверења, изабран је 0,25, што је подразумевана вредност коју је поставила ВЕКА.

Следећи алгоритам је **RF** алгоритам; узима се 100 као број итерација, што одређује број стабала у случајној шуми. Максимална дубина је

постављена на нулу, што значи да не постоје границе са дубином.

Tabela 3.3: Резултати оптимизације хиперпараметара RF алгоритма коришћењем 100 итерација

Број тренинга	Bag Size	Тачност
1	**100**	**98.81%**
2	75	98.81%
3	50	98.79%

Хиперпараметар bag size је подешен тако да има вредности од 100, 75 и 50. За bag size од 100 и 75, перформансе су дале исти резултат, 98,81%, док је за bag size од 50 , перформансе су опале за 0,02%. Изабран је bag size од 100, што је подразумевана вредност коју је поставила ВЕКА.

Следећи алгоритам је **SVM** и коришћена је C вредност и Kernel хиперпараметри.

Tabela 3.4: Резултати оптимизације хиперпараметара SVM алгоритма коришћењем C вредности и Kernel хиперпараметара.

Број тренинга	C	Kernel	Тачност
1	1	Poly Kernel	95.48%
2	2	Poly Kernel	95.34%
3	3	Poly Kernel	95.22%
4	1	Normalized Poly Kernel	94.84%
5	2	Normalized Poly Kernel	95.34%
6	3	Normalized Poly Kernel	95.39%
7	**1**	**Pearson VII**	**98.81%**
8	2	Pearson VII	98.81%
9	3	Pearson VII	98.81%

Вредност C служи као параметар регуларизације, који контролише колико често ће класификатор избегавати грешке у класификовању узорака за тренинг. Коришћене C вредности су биле 1, 2 и 3. Коришћени кернели су били Poly Kernel, нормализовани Poly Kernel и Пирсонова VII функција. Изведено је девет тренинга, а према резултатима, C вредност од 1 је дала најбољи учинак за све коришћене Кернеле. Pearson VII је дао најбољу тачност међу коришћеним кернелима, од 98,81%, а пошто су перформансе модела биле исте за све коришћене C вредности, изабрана је C вредност од 1, што је подразумевана вредност коју је поставила ВЕКА.

За **k-NN**, коришћена функција удаљености била је Еуклидска раздаљина за све изведене тренинге, а подешаван је параметар KNN пареметар (или број најближих суседа који се користе у процесу), као и параметар Cross-Validate (да ли ће унакрсна провера бити обављена или не).

Tabela 3.5: Резултати оптимизације хиперпараметара k-NN алгоритма користећи Еуклидову удаљеност.

Број тренинга	KNN	Cross Validate	Тачност
1	**1**	**True**	**98.69%**
2	3	True	98.69%
3	7	True	98.69%
4	1	False	98.69%
5	3	False	97.57%
6	7	False	94.53%

Коришћене KNN вредности су биле 1, 3 и 7, а за параметар унакрсне провере вредности True или False. Тренинзи који су дали високу тачност били су тренинзи где је параметар унакрсне валидације био постављен на True. Када је KNN вредност била 1, чак и без унакрсне провере, тачност је и даље била иста. KNN вредности 3 и 7 без употребе унакрсне валидације дале су нижу тачност, са 97,57% и 94,53%, респективно. Пошто се класификатор показао подједнако добро на прва три спроведена тренинга, изабрана KNN вредност је 1 и параметар унакрсне провере постављен на True.

На крају, **Naïve Bayes** алгоритам, коришћени су параметри Use Kernel Estimator и Supervised Discretization.

Tabela 3.6: Резултати оптимизације хиперпараметара NB алгоритма коришћењем хиперпараметара Kernel Estimator и Supervised Discretization.

Број тренинга	Use Kernel Estimator	Supervised Discretization	Тачност
1	**False**	**False**	**93.98%**
2	True	False	93.98%
3	False	True	93.98%

Може се приметити да су сви тренинзи коришћењем Kernel Estimatora или Supervised Discretization дали исте резултате, 93,98%. Зато је

изабрана подразумевана вредност коју је поставила ВЕКА, а то је False за оба параметра.

За табеле 2.2–2.6, редови подебљани означавају конфигурацију коришћену у изградњи модела за предвиђање КОВИД-19 који је прошао упоредну анализу.

3.2.2 Резултати за компаративну анализу

Након фазе оптимизације хиперпараметара, одлучено је која је најбоља конфигурација за сваки алгоритам која ће дати најбоље резултате. За упоредну анализу, supervised алгоритми машинског учења који користе те најбоље конфигурације коришћени су у изградњи модела који ће предвидети присуство КОВИД-19 код особе. Развијени модели су процењени коришћењем технике 10-струке унакрсне валидације.

Tabela 3.7: Главне мере тачности, прецизности, опозива и F-мере за поређење перформанси сваког алгоритма.

Алгоритам	Тачност	Прецизност	Recall	F-Measure
J48 DT	98.57%	0.986	0.986	0.986
RF	**98.81%**	**0.988**	**0.988**	**0.988**
SVM	**98.81%**	**0.988**	**0.988**	**0.988**
k-NN	98.69%	0.987	0.987	0.987
NB	93.98%	0.940	0.940	0.940

Постоје два алгоритма за које су све вредности подебљане, што представља два алгоритма који су постигли највећу тачност међу коришћеним алгоритмима.

Током десетоструког процеса унакрсне валидације, време потребно за изградњу модела је забележено у секундама. Такође израчунавају се перформансе модела према исправно и неисправно класификованим инстанцама, капа статистици и средњој апсолутној грешци.

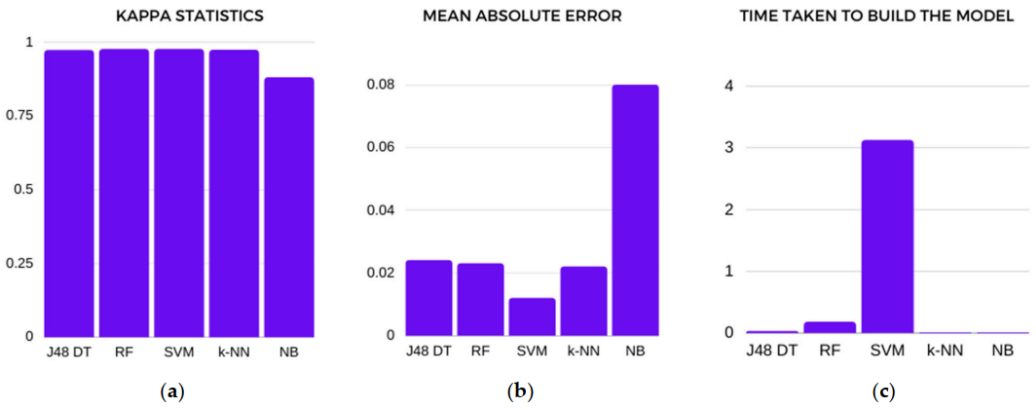
У овој табели приказани су резултати исправно и неисправно класификованих инстанци, капа статистика и средњи резултати апсолутне грешке на основу 10-струких резултата унакрсне валидације. Ови резултати су неопходни у процесу упоредне анализе и одређивања најприкладнијег алгоритма који ће се користити у изградњи предиктора КОВИД-19.

Tabela 3.8: Перформансе Supervised алгоритама за машинско учење са скупом података о симптомима и присутности КОВИД-19 помоћу подешених хиперпараметара.

Алгоритам	Исправно класифик инстанце	Неисправно класифик инстанце	Kappa	Средња апсолутна грешка	Време
J48 DT	4144	60	0.972	0.024	0.03
RF	**4154**	**50**	**0.976**	0.023	0.18
SVM	**4154**	**50**	**0.976**	**0.012**	3.12
k-NN	4149	55	0.973	0.022	**0.01**
NB	3951	253	0.880	0.080	**0.01**

3.3 Дискусија

На основу експерименталних резултата, RF и SVM су алгоритми машинског учења који се издвајају међу осталим supervised алгоритмима машинског учења.. Постигли су највећу тачност, као и највеће вредности у другим мерама тачности, као што су прецизност, опозив и F-мера. RF и SVM су такође били најбољи алгоритми у погледу исправно класификованих инстанци, а имали су најмање неисправно класификованих инстанци. Ове анализе говоре да су ови алгоритми који користе оптимизоване конфигурације били **најпоузданији алгоритми у погледу идентификације присуства КОВИД-19 код особе на основу датих симптома.**



Слика 3.1: *Графикони за капа статистику (a), средњу апсолутну грешку (b) и време потребно за изградњу модела (c).*

Посматра се капа статистика. Према резултатима, најучинковитији алгоритми су и даље били RF и SVM, са 0,976 капа резултата. J48 DT је достигао статистички резултат од 0,972 капа, док је KNN постигао 0,973. Најнижи капа резултат припада NB, 0,880. Коришћењем Cohen-ове kappa-е, компаративна анализа је постала поузданија у поређењу са само коришћењем тачности.

Разматрамо коришћење средње апсолутне грешке (MAE) за мерење грешке предвиђених класа у поређењу са стварним класама. На овај начин, може да се одреди колико добро ће развијени модел предвидети ознаке узорака у поређењу са стварним вредностима присутним у скупу података. Сматра се да је модел са ниским MAE ефикаснији од модела са вишим MAE. Најнижи MAE постигао је SVM алгоритам, са резултатом 0,012. Што је нижи резултат, то је мања шанса да ће класификатор направити грешке или погрешне класификације током предвиђања.

На крају, време потребно за изградњу модела. На основу резултата, k-NN алгоритам је био најбржи алгоритам за обуку модела класификатора, а следе NB и J48 DT. Аргоритмима RF и SVM, било је потребно више времена за тренинг: RF-у је требало 0,18 секунди, а SVM-у 3,12 секунди. Истовремено, ово су алгоритми који су се најбоље показали.

Све у свему, **SVM** алгоритам са одговарајућом конфигурацијом или подешавањем хиперпараметара јесте најприкладнији алгоритам који ће се користити у развоју предиктора КОВИД-19. SVM је постигао 98,81% за меру тачности и предвидео 4154 случаја од укупно 4204 случаја, погрешно класификујући само 50. SVM је такође постигао одличан резултат у капа статистици, на 0,976, што је најближе савршеном слагању од 1. За средњу апсолутну грешку, SVM је приказао грешку од 0,012, а то значи да развијени модел који користи овај алгоритам може да направи најмање могуће грешке у поређењу са другим алгоритмима. SVM и RF су били алгоритми који су радили најбоље, иако у смислу MAE критеријума, SVM је постигао веома низак резултат у поређењу са RF, са 0,023, што значи да је већа вероватноћа да ће RF извршити погрешне класификације. Међутим, RF има брже време обуке у поређењу са SVM-ом, тако да је RF други најбољи алгоритам који треба узети у обзир при изградњи предиктора КОВИД-19.

Глава 4

Закључак

Овај пројекат је имао за циљ да изгради модел предиктора присутности КОВИД-19 применом пет supervised алгоритама за машинско учење, укључујући J48 Decision Tree, Random Forest, K-Nearest Neighbors, Naïve Bayes и Support Vector Machine. Упоредна анализа је спроведена проценом перформанси модела у 10-струкој унакрсној валидацији преко ВЕКА софтвера за машинско учење. Резултати показују да је support vector machine (SVM) који користи универзални kernel Pearson VII најбољи алгоритам за машинско учење, који има тачност од 98,81% и 0,012 просечне апсолутне грешке. Алгоритам SVM је надмашио друге алгоритме у погледу тачности, прецизности, опозива, F-мере, исправно и неисправно класификованих инстанци, kappa статистичког резултата, средње апсолутне грешке и времена потребног за изградњу модела.

Потребно је напоменути и да је Random Forest (RF) други најбољи алгоритам који треба узети у обзир при изградњи предиктора присуства КОВИД-19, јер има исте мере тачности као оне које постиже алгоритам SVM, осим средње апсолутне грешке. Алгоритам RF се може узети у обзир у развоју модела високих перформанси са краћим временом обуке у поређењу са SVM.

Поред та два алгоритма, K-Nearest Neighbors је трећи најприкладнији алгоритам који се користи, јер такође може да изгради модел у кратком временском периоду у поређењу са другим алгоритмима. Затим, J48 стабло одлучивања је рангирано као четврто, а Naïve Bayes је рангиран као пети најприкладнији алгоритми који се разматра.

Ове информације могу бити значајне за лекаре, користећи развијени модел као помоћно средство у откривању присуства КОВИД-19 код особе

на основу декларисаних симптома. Поред тога, појединци који имају неке симптоме повезане са КОВИД-19 могу га користити да одреде могућност да буду позитивни или негативни приликом тестирања на КОВИД-19, да би се добила рана дијагноза болести. На овај начин може се помоћи у спречавању ширења заразне болести, умањујући опасност по људске животе. Развијени модел може се користити за прављење апликације са следећим предностима:

- Појединци могу лако да провере могућност добијања КОВИД-19 на основу симптома;
- Прелиминарна процена пацијената за лекаре;
- Помагање предузећима да ограниче физички контакт са клијентима који су можда оболели од КОВИД-19;
- Праћење да ли је особа развила симптоме КОВИД-19 док је била у изолацији;
- Заједница и влада могу користити ове информације као средство за смањење ширења вируса раним откривањем КОВИД-19.

Литература

[1] World Health Organization (WHO) https://www.who.int/health-topics/coronavirus

[2] Weka: https://www.cs.waikato.ac.nz/ml/weka/

[3] Kaggle база података: https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence

[4] The Importance of Training on Balanced Datasets: https://towardsdatascience.com/why-weight-the-importance-of-training-on-balanced-datasets-f1e54688e7df

[5] How to Use Classification Machine Learning Algorithms in Weka https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/

[6] Бајесова теорема https://www.britannica.com/topic/Bayess-theorem

[7] Euclidean Distance: https://sebastianraschka.com/faq/docs/euclidean-distance.html

[8] Accuracy, Recall, Precision, F-Score: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

[9] Absolute Error Mean Absolute Error (MAE): https://www.statisticshowto.com/absolute-error/

[10] Performance analysis of data mining algorithms for diagnosing COVID-19 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8719570/