Klasifikacija zdravlja fetusa na temelju kardiotokografije

Valentina Križ, Jelena Kurilić, Lucija Valentić

***Sadržaj*—U ovom radu bavimo se klasifikacijom zdravlja fetusa na temelju kardiotokograma preuzetih s UCI repozitorija. Zbog nebalansiranosti podataka proučavamo nekoliko vrsta oversampling-a koje kombiniramo s različitim metodama strojnog učenja. Naš cilj je pronalazak alata koji će biti koristan u detekciji mogućeg patološkog stanja.**

I. UVOD

Fetalna patnja (*fetal distress*) je prilično slabo definiran medicinski pojam koji se odnosi na zdravstvene probleme nerođenog djeteta u trećem tromjesečju trudnoće. U većini slučajeva, fetalna patnja uključuje nedostatak kisika što dovodi do izvođenja hitnog carskog reza. Kako bi se to spriječilo ili pravovremeno liječilo koristi se dijagnostička metoda praćenja stanja fetusa zvana kardiotokografija (CTG). Kardiotokografija podrazumijeva grafički prikaz aktivnosti srca ploda i aktivnosti mišića zida maternice tokom trudnoće i tokom samog porođaja.

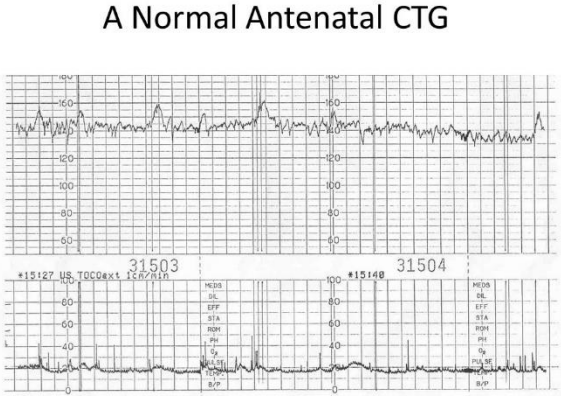

Figure 1: Primjer normalnog kardiotokograma

II. PODACI

Koristimo *„Cardiotocography"* skup podataka preuzet s UCI repozitorija za strojno učenje [1]. Skup podataka se sastoji od 2126 automatski procesiranih primjeraka CTG snimki. Svaka snimka je predstavljena pomoću 21 varijable (Table 1). Sve snimke su podijeljene u 3 klase (normalnu, sumnjivu i patološku) od strane stručnjaka. Također, postoji mogućnost podjele snimaka u 10 klasa koje preciznije određuju stanje fetusa, ali ta podjela se rijetko koristi.

Symbol	Description
LB	baseline value (SisPorto)
AC	accelerations (SisPorto)
FM	fetal movement (SisPorto)
UC	uterine contractions (SisPorto)
ASTV	percentage of time with abnormal short term variability (SisPorto)
mSTV	mean value of short term variability (SisPorto)
ALTV	percentage of time with abnormal long term variability (SisPorto)
mLTV	mean value of long term variability (SisPorto)
DL	light decelerations
DS	severe decelerations
DP	prolongued decelerations
DR	repetitive decelerations
Width	histogram width
Min	low freq. of the histogram
Max	high freq. of the histogram
Nmax	number of histogram peaks
Nzeros	number of histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency: -1=left asymmetric; 0=symmetric; 1=right asymmetric

Table 1: Opis svih CTG obilježja

III. DOSADAŠNJA ISTRAŽIVANJA

U cilju očuvanja zdravlja fetusa, bitna je brza interpretacija CTG snimki, pa je CTG dataset često korišten u radovima vezanim za strojno učenje. U nastavku navodimo neke od radova u kojima je korišten spomenuti dataset. Diksriminantna analiza

(DA), stabla odlučivanja (DT) i umjetne neuronske mreže (ANN) su korišteni u radu iz 2012. godine autora Huang. Postignuta je točnost od 82.1%, 86.36% i 97.78%, respektivno [2]. 2012. je također Sundar pomoću neuronskih mreža postigao točnost od 80% [3]. Yılmaz and Kılıkçıer 2013. koriste metodu potpornih vektora (SVM) i dobivaju točnost 91.62% [4]. Ocak i Ertunç (2013.) klasificiraju podatke u 2 klase (normalnu i patološku) pomoću prilagodljivog neuro-neizrazitog sustava (ANFIS) i time točno klasificiraju 96.6% patoloških stanja i 97.2% normalnih stanja [5]. 2014. Karabulut uspoređuje 6 metoda strojnog učenja: naivni Bayes (NB), radijalne mreže (RBN), Bayesove mreže (BN), SVM, ANN i DT bez i s AdaBoost metodom. Točnost NB, RBN, BN i DT je 87.39%, 87.67%, 92.61% i 95.01% respektivno, a ostalim metodama nije postignut značajan napredak [6]. Korištenjem metode najbližih susjeda (k-NN) i slučajnih šuma (RF) Şahin i Subasi postižu točnost od 98.4% i 99.18% [7]. Performanse ANN i ELM (Extreme Learning Machine) su uspoređene u radu autora Cömert i dobivena je veća točnost pomoću ELM-a (93.42%) [8]. Arif predlaže upotrebu slučajnih šuma i dobiva točnost od 93.6% [9]. Kamath i Kamat primjenjuju istu metodu za podjelu u 10 klasa i dobivaju točnost preko 87% [10].

IV. KORIŠTENI ALATI

Eksploratornu analizu, treniranje modela i analizu dobivenih rezultata radile smo u Python-u koristeći Jupyter Notebook okruženje. Uz standardne Python-ove biblioteke (`scikit-learn`, `pandas`, `numpy`, `matplotlib`), koristile smo i funkcije `SMOTE`, `BorderlineSMOTE` i `ADASYN` za oversampling iz biblioteke `imbalanced-learn` [11].

V. PRISTUP PROBLEMU

A. Eksploratorna analiza

Eksploratornom analizom dataseta dolazimo do zaključka da je dani problem nebalansiran klasifikacijski problem. Za evaluaciju naših modela stoga, uz točnost, koristimo osjetljivost i konfuzijsku matricu. Budući da se bavimo problemom medicinske prirode i cilj nam je detektirati patološke slučajeve, naglasak stavljamo na osjetljivost (omjer broja prepoznatih patoloških primjeraka i ukupnog broja patoloških primjeraka).

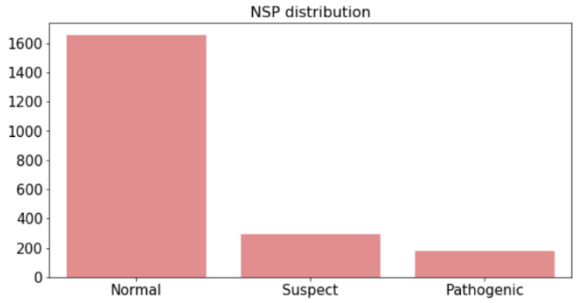

Figure 2: Distribucija klasa u CTG datasetu

Iz korelacijske matrice (Figure 3) uočavamo koreliranost značajki Min i Width (negativna koreliranost -0.89), te značajki Mean i Median (pozitivna koreliranost 0.94). U slučaju jake koreliranosti dviju varijabli, preporuča se izbaciti jednu od njih [12]. Budući da su Mean i Width jače korelirane s ciljnom varijablom, odlučujemo izbaciti značajke Min i Median.

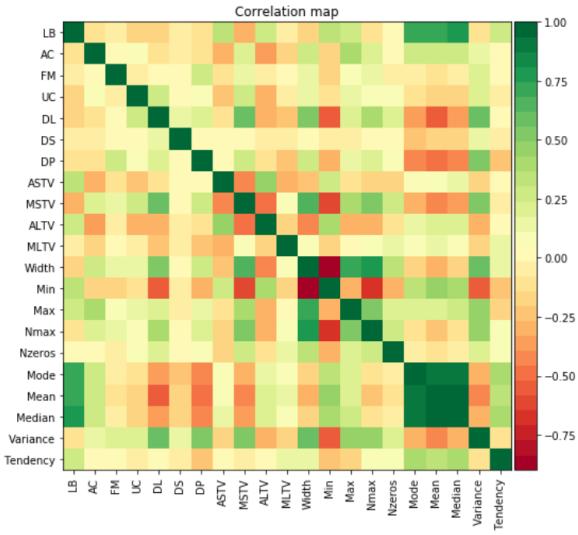

Figure 3: Matrica koreliranosti obilježja

B. Podjela primjera i odabir modela

Radi korištenja oversampling metoda prilikom treniranja, podatke dijelimo na samom početku na train i test skupove u omjeru 80:20, a stratifikacijom osiguravamo da oba podskupa dobro reprezentiraju cijeli skup podataka. Zatim kombinacijom `GridSearchCV`-a i `imbalanced-learn`

`Pipeline`-a istovremeno tražimo optimalne parametre modela i treniramo model pomoću 5-fold unakrsne validacije s primjenom oversampling metoda. Svaki model prvo treniramo bez oversampling-a, a zatim i s 3 vrste oversampling-a (SMOTE, BorderlineSMOTE i ADASYN).

Od modela strojnog učenja odabiremo XGBoost, SVC i Random forest. XGBoost je veoma popularan model koji se često koristi za nebalansirane podatke, a SVC i Random forest odabiremo jer su u dosadašnjim istraživanjima dali veoma dobre rezultate.

C. Parametri modela

Pomoću GridSearchCV-a određujemo optimalne parametre za pojedine modele. Za XGBoost optimiziramo sljedeće parametre: `min_child_weight` $\in \{1, 5, 10\}$, `gamma` $\in \{0.5, 1, 1.5, 2, 5\}$, `subsample` $\in \{0.6, 0.8, 1.0\}$, `colsample_bytree` $\in \{0.6, 0.8, 1.0\}$ i `max_depth` $\in \{3, 4, 5\}$. Optimizirani parametri za Random forest su `n_estimators` $\in \{16, 32, 64, 128\}$ i max_features $\in \{1, 2, 3, \ldots, 19\}$. Za SVC optimiziramo parametar `c` $\in \{1, 10, 100, 1000\}$ uz `linear` kernel, te parametre `c` $\in \{1, 10, 100, 1000\}$ i `gamma` $\in \{0.001, 0.0001\}$ uz `rbf` kernel.

D. oversampling

Algoritmi strojnog učenja najčešće imaju problema s nebalansiranim podacima jer su tada skloni pridružiti većinu podataka dominirajućoj klasi. Stoga se često koriste undersampling i oversampling. Zbog relativno malog broja primjeraka mi se odlučujemo za oversampling pomoću kojeg se povećava broj primjeraka koji pripada manje zastupljenoj klasi i time postiže ravnopravnija klasifikacija.

Na slici 4 prikazan je primjer nebalansiranog skupa podataka, dok slike 5, 6 i 7 prikazuju iste podatke uz primjenu različitih oversampling metoda (SMOTE, BorderlineSMOTE i ADASYN respektivno).

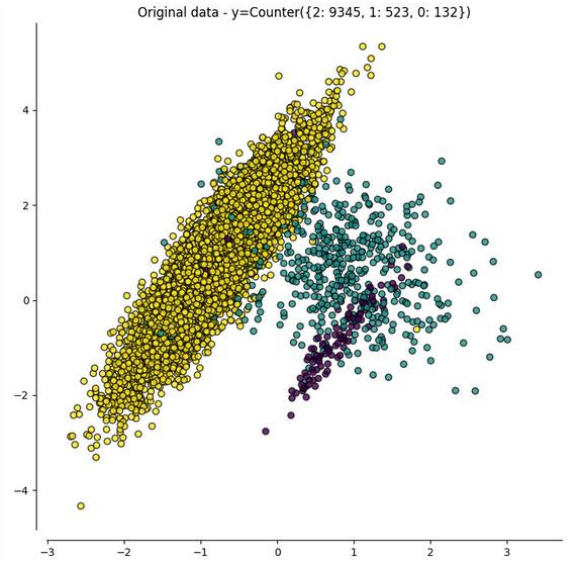

Figure 4: Originalni podaci

SMOTE (Synthetic Minority Over-sampling Technique) stvara nove primjerke tako što najprije odabere jedan od postojećih primjeraka te jedan od njegovih *k* najbližih susjeda i zatim konstruira novi primjerak na slučajno odabranoj poziciji na vektoru koji spaja ta dva primjerka.

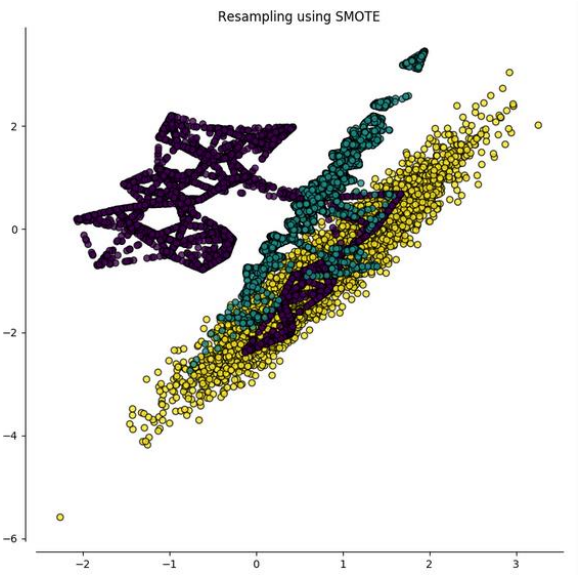

Figure 5: Podaci dobiveni uz SMOTE

BorderlineSMOTE stvara nove podatke jednako kao i SMOTE, ali samo na granici između manje i više zastupljene klase.

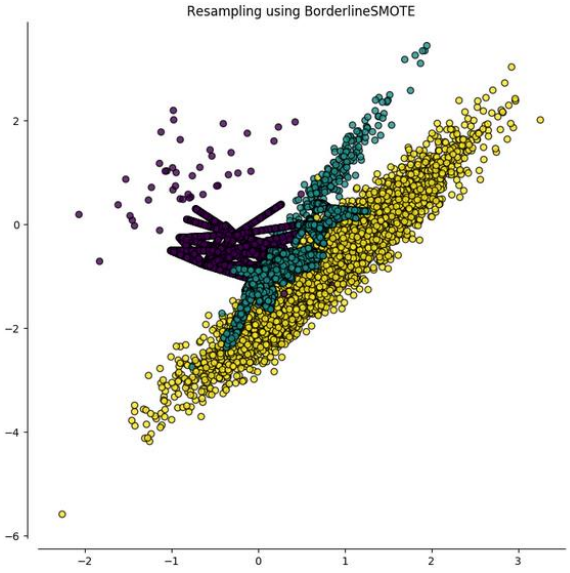

Figure 6: Podaci dobiveni uz BorderlineSMOTE

ADASYN (Adaptive Synthetic) oversampling tehnika radi slično kao i BorderlineSMOTE, ali ne konstruira nove primjere samo na granicama klasa nego i u blizini primjeraka manje zastupljene klase koji su okruženi primjercima više zastupljene klase.

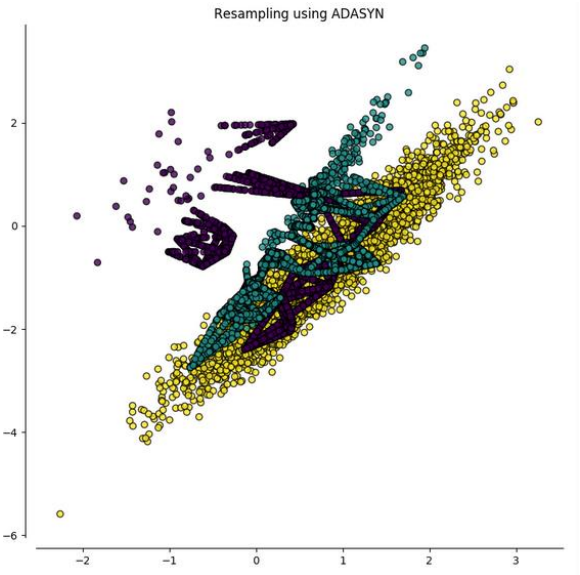

Figure 7: Podaci dobiveni uz ADASYN

VI. REZULTATI

A. SVC

SVC model pokazuje se kao najlošiji od svih isprobanih modela. U varijanti bez oversampling-a dobivamo točnost od 90.61%, ali osjetljivost je znatno niža – 74.29%. Od svih isprobanih oversampling varijanti, najbolje rezultate uz SVC dobivamo koristeći ADASYN oversampling. Točnost tog modela je nešto manja od modela bez oversampling-a (87.32%), ali je osjetljivost bolja i iznosi 91.43%. Na slici 8 prikazane su konfuzijske matrice za SVC bez oversampling-a i SVC uz ADASYN oversampling.

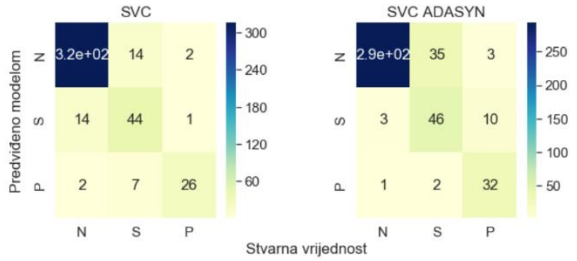

Figure 8: Konfuzijske matrice za SVC

B. XGBoost

XGBoost se očekivano pokazuje najboljim modelom. Čak i s modelom bez oversampling-a dobivamo zadovoljavajuće rezultate – točnost 93.43% i osjetljivost 91.43%. Ipak, koristeći modele s oversampling-om dobivamo bolje rezultate, a najbolje uz ADASYN oversampling – točnost 93.66% i osjetljivost 94.29%. Na slici 9 prikazane su konfuzijske matrice za XGBoost bez oversampling-a i XGBoost uz ADASYN oversampling.

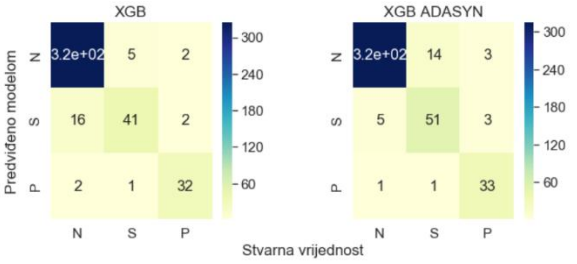

Figure 9: Konfuzijske matrice za XGBoost

Na kraju provjeravamo važnost pojedinih značajki u XGBoost modelu s ADASYN oversampling-om i značajke AC (accelerations) i ALTV (percentage of

time with abnormal long term variability) pokazuju se najvažnijima (Figure 10). Long term variability pokazuje koliko se međusobno razlikuju otkucaji srca u zadanom interval i u mnogim medicinskim radovima je istaknut kao važna komponenta pri određivanju zdravlja fetusa.

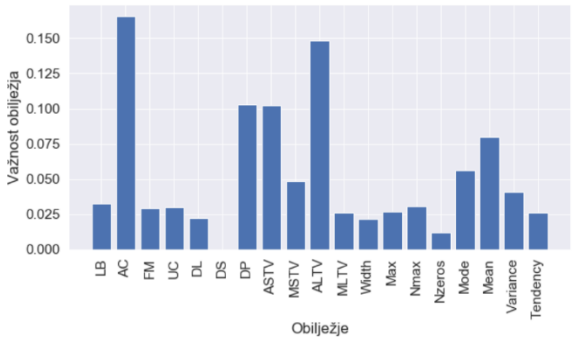

Figure 10: Važnost obilježja uz XGBoost s ADASYN oversampling-om

C. Random Forest

Zadovoljavajuće rezultate dobivamo koristeći Random Forest model. Bez oversampling-a dobivamo točnost 92.49% i osjetljivost 85.71%. Za razliku od prethodnih modela, uz Random Forest se najboljim pokazuje BorderlineSMOTE oversampling kojim dobivamo točnost 93.19% i osjetljivost 91.43%. Na slici 11 prikazane su konfuzijske matrice za Random Forest bez oversampling-a i Random Forest uz BorderlineSMOTE oversampling.

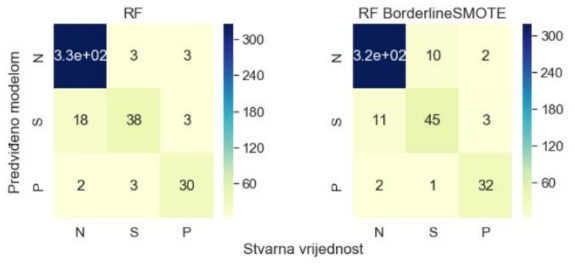

Figure 11: Konfuzijske matrice za Random Forest

Također promatramo važnost pojedinih značajki u Random Forest modelu uz BorderlineSMOTE oversampling (Figure 12) i dobivamo slične rezultate kao i kod XGBoost modela. Značajka ALTV se ponovno pokazuje iznimno važnom, a uz nju i značajka ASTV (percentage of time with abnormal short term variability) koja se također često ističe u medicinskim radovima.

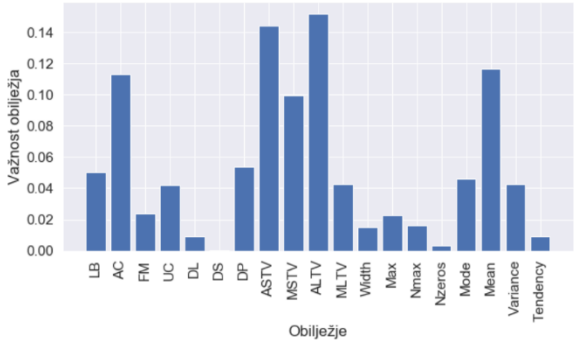

Figure 12: Važnost obilježja uz Random Forest s BorderlineSMOTE oversampling-om

D. Usporedba modela

Na kraju prikazujemo tablicu svih istreniranih modela s dobivenom točnošću i osjetljivošću radi ilustracije gore navedenih rezultata.

	-	SMOTE	BSMOTE	ADASYN
SVC	90.61	89.20	85.92	87.32
XGB	93.43	93.13	94.13	93.66
RF	92.49	92.25	93.19	92.25

Table 2: Točnost modela u postocima

	-	SMOTE	BSMOTE	ADASYN
SVC	74.29	88.57	85.71	91.43
XGB	91.43	91.43	91.43	94.29
RF	85.71	85.71	91.43	91.43

Table 3: Osjetljivost modela u postocima

VII. ZAKLJUČAK

Najbolje rezultate dobivamo koristeći XGBoost model koji se pokazuje dovoljno dobrim i bez oversampling-a, ali daje bolje rezultate uz ADASYN oversampling. Kod SVC i RF modela je pak vidljiva razlika u rezultatima bez i s korištenjem oversampling-a. Najveća razlika (17.14%) je vidljiva u SVC modelu gdje je osjetljivost porasla sa 74.29% na 91.43% zahvaljujući ADASYN oversampling-u. Slično je dobiveno u RF modelu, gdje je osjetljivost porasla s 85.71% na 91.43%. SMOTE oversampling se u našem slučaju pokazuje kao najlošija vrsta oversampling-a, dok ADASYN u prosjeku daje najbolje rezultate.

Promatrajući konfuzijske matrice možemo zaključiti da klasifikatori prilikom čijeg treniranja nije korišten oversampling većinu primjeraka označavaju kao Normal ili Suspect. To je normalna pojava u

slučaju nebalansiranih podataka, jer klasifikatori tada najčešće većinu podataka stavljaju u najzastupljeniju klasu. Koristeći oversampling prilikom treniranja izbjegavamo tu pojavu te klasifikatori češće smještaju primjerke i u manje zastupljenu klasu (Pathologic). Upravo to smo htjele postići budući da je u medicini bitnije točno dijagnosticirati patološka stanja.

VIII. DALJNJI RAD

U daljnjim istraživanjima točnost i osjetljivost modela mogla bi se popraviti korištenjem 10-fold unakrsne validacije umjesto ovdje korištene 5-fold unakrsne validacije. Također, optimalni parametri modela bi se mogli pronaći koristeći više različitih metoda uz veći raspon danih parametara. No, smatramo da bi se najveći napredak ostvario prikupljanjem većeg broja primjeraka budući da je trenutno dostupan skup podataka relativno malen (2126 primjeraka).

REFERENCES

[1] https://archive.ics.uci.edu/ml/datasets/cardiotocography

[2] M.-L. Huang and Y.-Y. Hsu, "Fetal distress prediction using discriminant analysis, decision tree, and artificial neural network," Journal of Biomedical Science and Engineering, vol. 5, p. 526, 2012.

[3] C. Sundar, et al., "Classification of cardiotocogram data using neural network based machine learning technique," International Journal of Computer Applications, vol. 47, 2012.

[4] E. Yılmaz and Ç. Kılıkçıer, "Determination of fetal state from cardiotocogram using LS-SVM with particle swarm optimization and binary decision tree," Computational and mathematical methods in medicine, vol. 2013, 2013.

[5] H. Ocak and H. M. Ertunc, "Prediction of fetal state from the cardiotocogram recordings using adaptive neuro-fuzzy inference systems," Neural Computing and Applications, vol. 23, pp. 1583-1589, 2013.

[6] E. M. Karabulut and T. Ibrikci, "Analysis of cardiotocogram data for fetal distress determination by decision tree based adaptive boosting approach," Journal of Computer and Communications, vol. 2, p. 32, 2014.

[7] H. Sahin and A. Subasi, "Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques," Applied Soft Computing, vol. 33, pp. 231-238, 2015.

[8] Z. Cömert, et al., "Cardiotocography signals with artificial neural network and extreme learning machine," in Signal Processing and Communication Application Conference (SIU), 2016 24th, 2016, pp. 1493-1496.

[9] M. Arif, "Classification of cardiotocograms using random forest classifier and selection of important features from cardiotocogram signal," Biomaterials and Biomechanics in Bioengineering, vol. 2, pp. 173-183, 2015.

[10] R. Kamath and R. Kamat, "Modeling fetal morphologic patterns through cardiotocography data: Decision tree-based approach," Journal of Pharmacy Research| Vol, vol. 12, p. 10, 2018.

[11] https://imbalanced-learn.readthedocs.io/en/stable/

[12] https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf