Klasifikacija zdravlja fetusa na temelju kardiotokografije

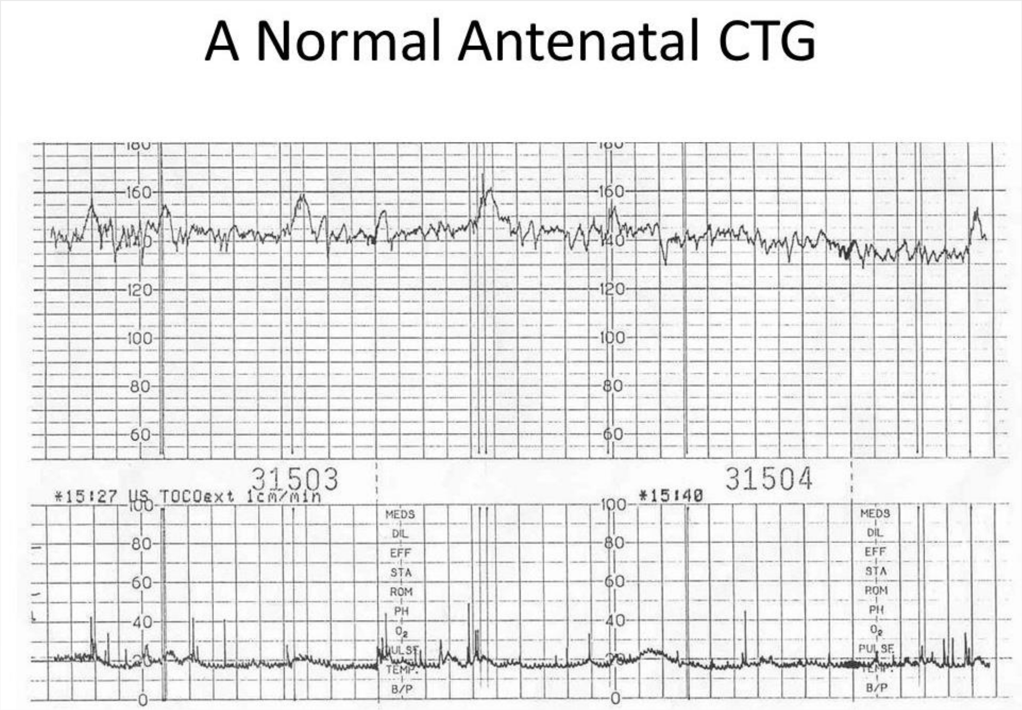
VALENTINA KRIŽ, JELENA KURILIĆ, LUCIJA VALENTIĆ

Strojno učenje
PMF-MO

Uvod

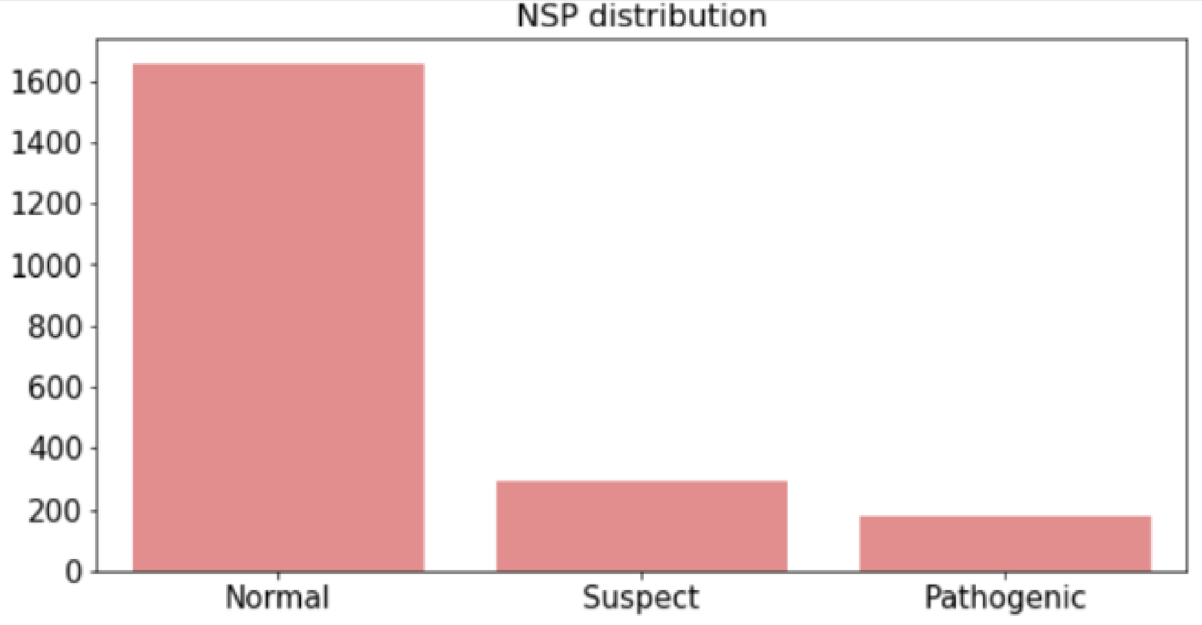
- fetalna patnja
  - zdravstveni problemi nerođenog djeteta u trećem tromjesečju trudnoće

- kardiotokografija
  - dijagnostička metoda praćenja stanja fetusa
  - grafički prikaz aktivnosti srca ploda i aktivnosti mišića zida maternice



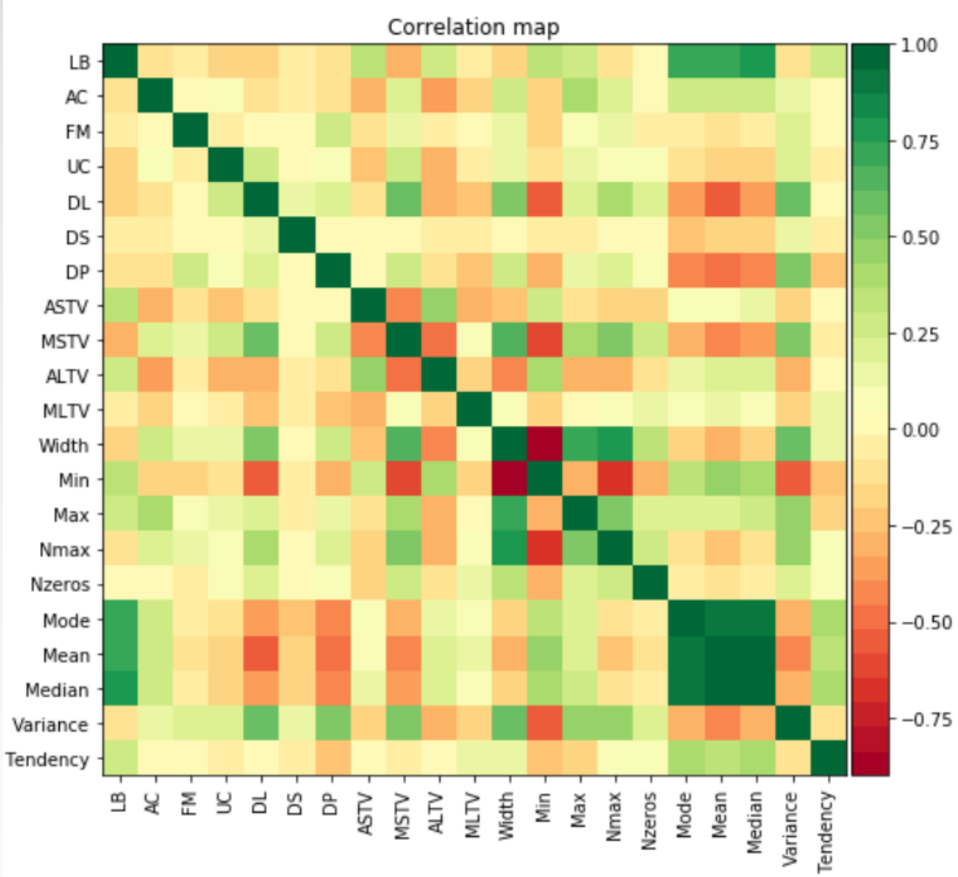
Podaci

- UCI „*Cardiotocography*“ skup podataka (2126 primjeraka automatski obrađenih CTG snimki)
- 21 varijabla, 3 klase (normalna, sumnjiva, patološka)



Nebalansirani klasifikacijski problem → oversampling

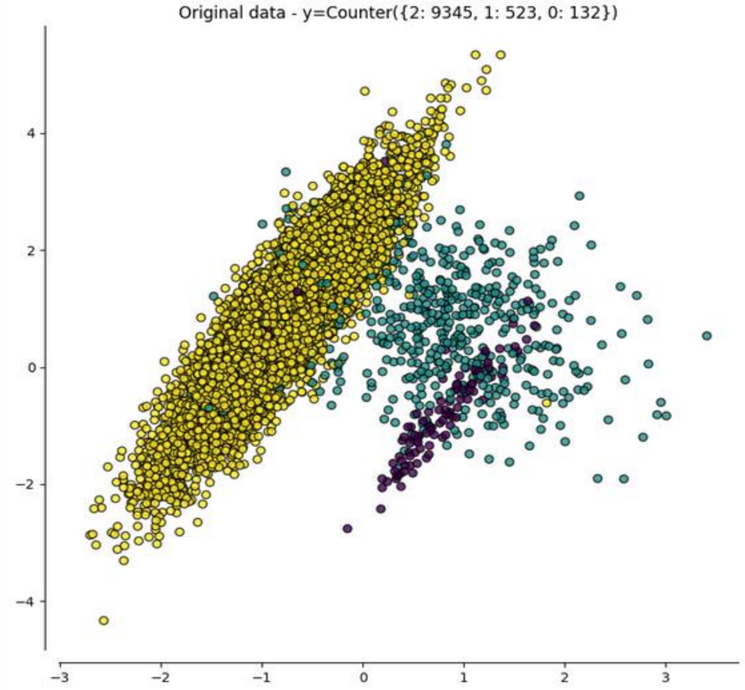
Varijable



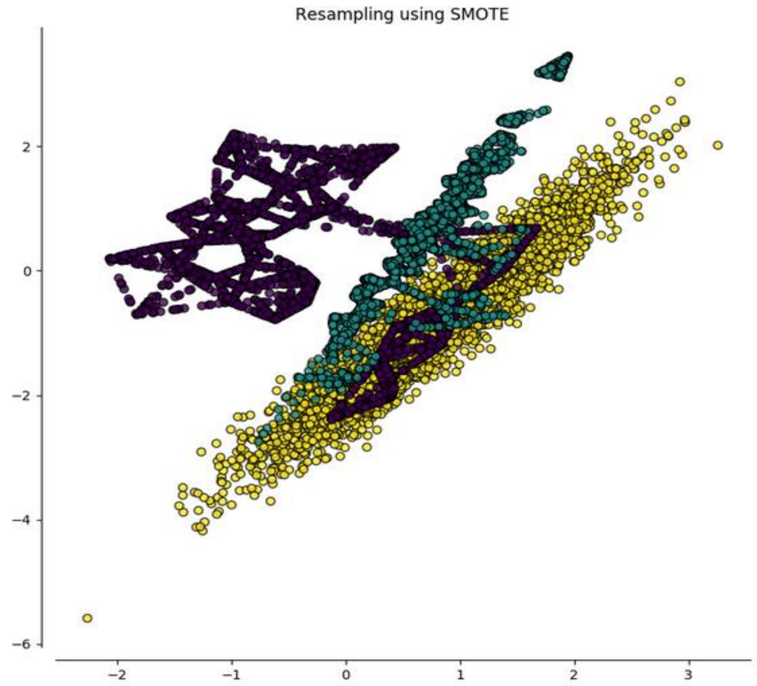
Symbol	Description
LB	baseline value
AC	accelerations
FM	fetal movement
UC	uterine contractions
ASTV	percentage of time with abnormal short term variability
mSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
mLTV	mean value of long term variability
DL	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

Podjela primjera i odabir modela

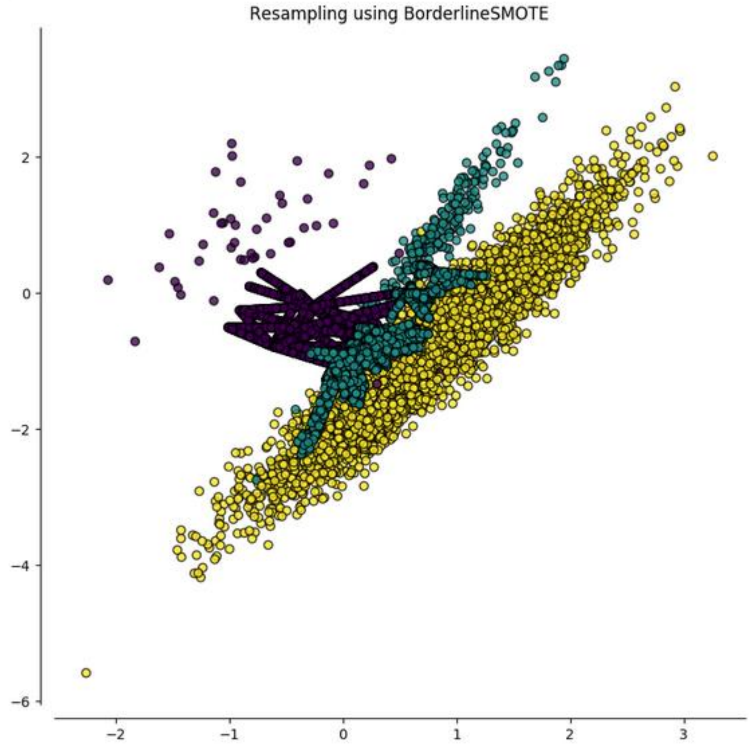
- train i test skup u omjeru 80:20 (stratificirana podjela)
- XGBoost, SVC, Random Forest
- GridSearchCV (5-fold unakrsna validacija) + imbalanced-learn Pipeline na train skupu
  - SMOTE, BorderlineSMOTE, ADASYN
  - parametri:
    - SVC - c ϵ {1, 10, 100, 1000}
      - gamma ϵ {0.001, 0.0001} (RBF kernel)
    - XGBoost - min_child_weight ϵ {1, 5, 10}
      - gamma ϵ {0.5, 1, 1.5, 2, 5}
      - subsample ϵ {0.6, 0.8, 1.0}
      - max_depth ϵ {3, 4, 5}
    - Random Forest - n_estimators ϵ {16, 32, 64, 128}
      - max_features ϵ {1, 2, 3, …, 19}
- testiranje na test skupu



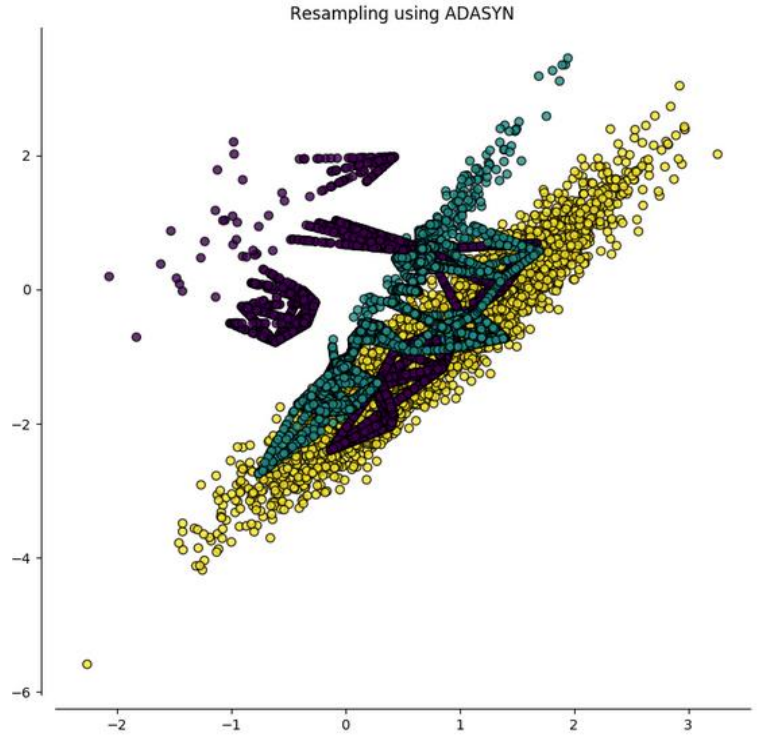
originalni
podaci



SMOTE



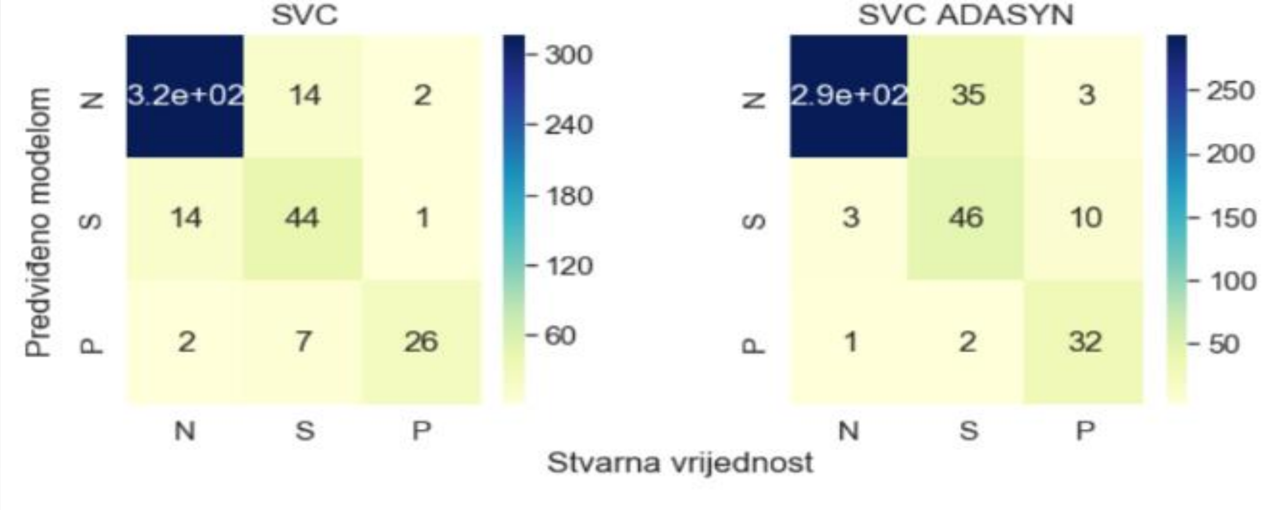
Borderline
SMOTE



ADASYN

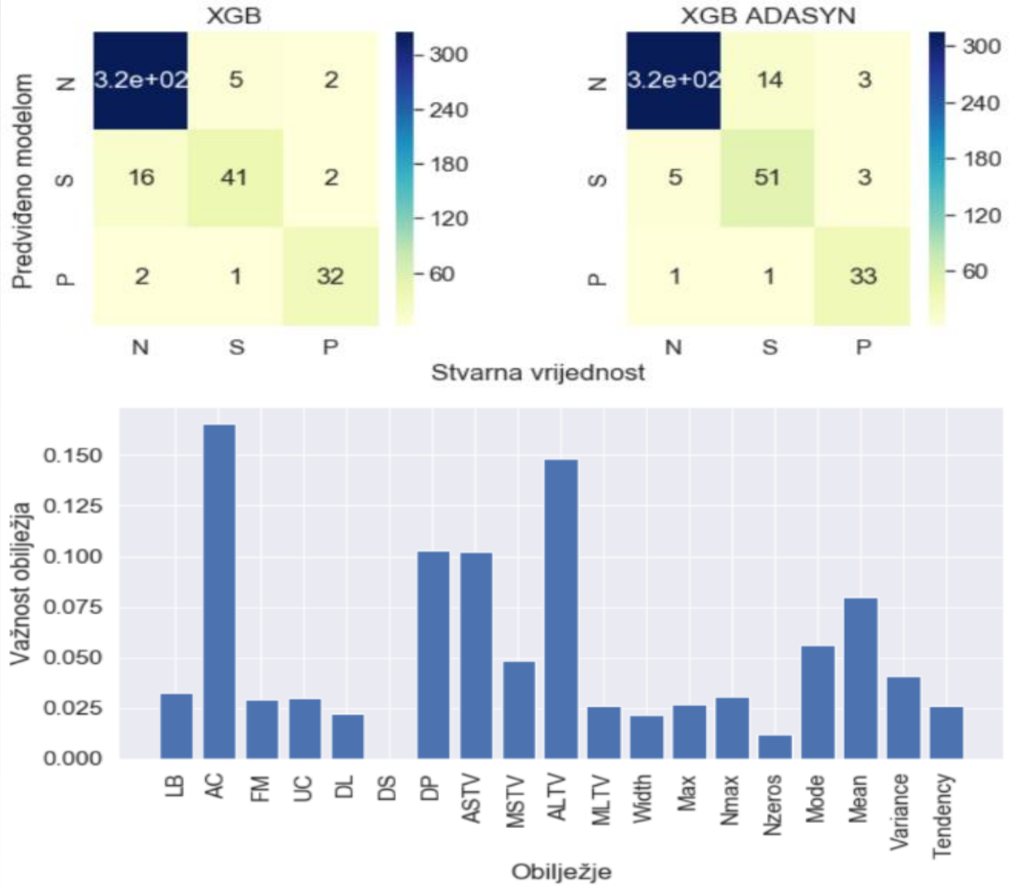
Rezultati - SVC

- najlošiji od svih isprobanih modela
- bez oversamplinga:
  - Točnost 90.61%
  - Osjetljivost 74.29%
- ADASYN oversampling:
  - Točnost 87.32%
  - Osjetljivost 91.43%



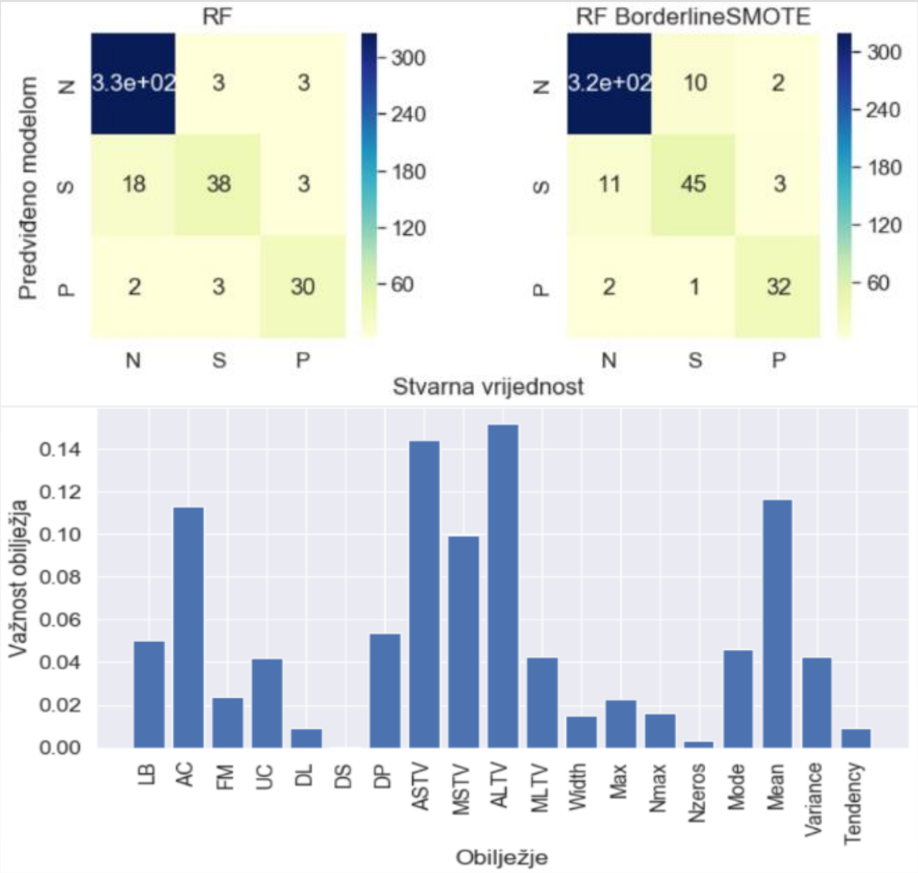
Rezultati - XGBoost

- najbolji od svih isprobanih modela
- bez oversamplinga:
  - Točnost 93.43%
  - Osjetljivost 91.43%
- ADASYN oversampling:
  - Točnost 93.66%
  - Osjetljivost 94.29%
- Najvažnije značajke:
  - AC (accelerations)
  - ALTV (percentage of time with abnormal long term variability)



Rezultati – Random Forest

- bez oversamplinga:
  - Točnost 92.49%
  - Osjetljivost 85.71%
- BorderlineSMOTE oversampling:
  - Točnost 93.19%
  - Osjetljivost 91.43%
- Najvažnije značajke:
  - ALTV (percentage of time with abnormal long term variability)
  - ASTV (percentage of time with abnormal short term variability)



Usporedba modela

	-	SMOTE	BorderlineSMOTE	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

Točnost modela u postocima

	-	SMOTE	BorderlineSMOTE	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

Osjetljivost modela u postocima

Zaključak

- XGBoost
  - uz ADASYN najbolji rezultati
  - dobri rezultati i bez oversamplinga
- SVC i Random Forest
  - uz oversampling značajan porast osjetljivost (i do 17.14%)
- u prosjeku najbolji ADASYN oversampling
- bez oversamplinga najčešće klasificirano kao Normal
- uz oversampling češće klasificiranje kao Pathologic (i Suspect)

Daljnji rad

- korištenje 10-fold unakrsne validacije
- veći raspon parametara pri određivanju optimalnih parametara
- prikupljanje većeg broja primjeraka

Literatura

[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.

Literatura

[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