Titanic

- **Autori:** Maja Marija Barukčić, Mislav Beg, Roko Kokan
- **Opis:**
  - Kaggle dataset: *Titanic: Machine Learning from Disaster* https://www.kaggle.com/c/titanic
    - Mali skup pripremljenih primjera s 12 atributa (ciljni Survived: 0,1)
      - TEST: 891 (*missing values* u Cabin > Age > Embarked)
      - TRAIN: 418 (*missing values* u Cabin > Age)
  - Binarna klasifikacija s neodređenim vrijednostima atributa (engl. *missing values*)
    - Random Forest (RF)
    - Mjera uspješnosti: točnost (engl. *accuracy*)
- **Ocjena projektnog prijedloga:**
  - + prikladan opis problema i dobro poznatog podatkovnog skupa (DS)
  - + opisani programski paketi koji će se koristiti za izradu rješenja (python + ostatak…)
  - – površan opis dosadašnjih pristupa i istraživanja samo s fokusom na algoritme učenja (bez osvrta na dosadašnju uspješnost pobrojanih metoda u sklopu *Kaggle Kernels* ili mogućih objavljenih radova*)*
  - – metodologija: fokus samo na jednu metodu (RF), upitan način rješavanja neodređenih vrijednosti atributa, nedefinirana mjera uspješnosti
  - – očekivani rezultati: iskazani prema nedefiniranoj mjeri uspješnosti (0.8?)
  - – tehnički dojam i stil pisanja: ponavnjanje opisa skupa primjera iz opisa problema, vrlo slab popis literature
- **Komentari i prijedlozi:**
  - zbog odabira problema koji je vrlo dobro proučavan i za koji postoji mnoštvo rješenja predlaže se da se provede iscrpnija analiza od one predložene
  - provesti podrobnu eksploratornu analizu podatkovnog skupa
  - ustanoviti primjere koji su *outliers*
  - usporediti više modela za učenje i napraviti prikladanu evaluaciju (razmotriti više mjera uspješnosti) i odabir modela (holdout, CV)
  - za modele primjerice mogli bi se istražiti i naučiti DT, RF, xgboost
  - preporuča se da se detaljno prouče sve uvodne analize zasnovane na korištenju programskog jezika Python ovog podatkovnog skupa predstavljene na https://www.kaggle.com/c/titanic#tutorials
  - BONUS: opis načina za određivanje neodređenih vrijednosti atributa

Nedolazak na liječničke preglede

- **Autori:** Barbara Prkačin, Edi Ibriks
- **Opis**:
  - Kaggle dataset, Medical Appointment No Shows, https://www.kaggle.com/joniarroba/noshowappointments
  - Veliki skup primjera (30k) zakazanih medicinskih pregleda s 15 atributa (ciljni atribut Status: No.Show: 90731/Show.Up:209269)
  - Binarna klasifikacija na nebalansiranim podacima (engl. *imbalanced data*)
    - naivni Bayes, linearni SVM i logistička regresija (*scikit-learn*)
    - *Logarithmic loss*
- **Ocjena projektnog zadatka:**
  - + prikladan opis *Kaggle* problema
  - - površan opis dosadašnjih pristupa i istraživanja samo s fokusom na algoritme učenja (bez osvrta na dosadašnju uspješnost pobrojanih metoda u sklopu *Kaggle Kernels* ili mogućih objavljenih radova*)*
  - +/- metodologija je fokusirana samo na opis izabranih metoda učenja bez osvrta na značajke i problem nebalansiranih podataka (neuravnoteženog broja primjera po vrijednosti ciljnog atributa)
  - +/- očekivani rezultati su nedorečeni, mogli se se staviti u kontekst rezultata s *Kagglea*
  - – slab tehnički dojam i stil: literatura nije referencirana unutar teksta
- **Komentari i prijedlozi:**
  - Proučavanje osnovnih pristupa za nošenje s problemom neuravnoteženih podataka
    - 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset, http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/
  - Dodatna literatura [opcijonalno] (https://scholar.google.hr/)
    - Galar, Mikel, et al. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.4 (2012): 463-484. PDF
  - Korisni alati:
    - https://github.com/scikit-learn-contrib/imbalanced-learn

Otkrivanje web-stranica za krađu identiteta

- **Autori**: Tin Mavračić, Goran Vuković, Marko Dominković
- **Opis:**
  - UCI ML repository: *Phishing Websites Data Set*, https://archive.ics.uci.edu/ml/datasets/Website+Phishing
  - Binarna klasifikacija (sa potencijalno nebalansiranim podacima)
    - Predložen k-NN, SVM I NN
    - Potrebno je razmotriti druge mjere uspješnosti osim predložene mjere za točnost klasifikacije u slučaju nebalansiranih podataka (vidi http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/)
- **Ocjena projektnog prijedloga**:
  - + dobar opis problema i dosadašnjih istraživanja uz odgovarajuće referenciranu literaturu.
  - – loša razrada metodologije samo sa izlistavanjem metode učenja, uz neodgovarajući odabir mjere uspješnosti
  - – nisu spomenuti očekivani rezultati (što se želi dobiti od projekta zasnovano na prethodnim istraživanjima)
- **Komentari i prijedlozi:**
  - U slučaju nebalansiranog skupa proučiti pristup osnovnih pristupa za nošenje s tim problemom
    - 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset, http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Otkrivanje prevara počinjenih kreditnim karticama

- **Autori:** Martina Alilović, Domagoj Galić, Tina Marić
- **Opis:**
  - @*Kaggle*: Credit Card Fraud Detection, https://www.kaggle.com/dalpozz/creditcardfraud
  - Binarna klasifikacija na jako nebalansiranom podatkovnom skupu
    - 92 prevara od ukupno 284,807 transakcija (0.172%)
    - anonimizirane transakcije - značajke primjera u skupu su 28 prvih PCA komonenti
    - Predložene metode učenja Naive Bayes?, SVM
    - Mjera uspješnosti: AUPRC
- **Ocjena projektnog prijedloga:**
  - – u prijedlogu se ne spominje izvor podatkovnog skupa (može se samo predpostaviti)
  - – nisu spomenuta dosadašnja istraživanja ni očekivani rezultati
  - – metodologija uključje korištenje samo dvije metode te neprimjereni pristup učenja za visoko nebalansirane podatke, bez osvrta na rješavanje tog problema iako je spomenuto korištenje prikladne mjere uspješnosti
  - – nije navedeno kako će se napraviti split na train/test kod tako nebalansiranih podataka
  - – za tehnički dojam: literatura nije referencirana u tekstu, pristup učenja modela opisan u sekciji podaci
- **Komentari i prijedlozi:**
  - Proučavanje osnovnih pristupa za nošenje s problemom neuravnoteženih podataka
    - 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset, http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/
  - Dodatna literatura [opcijonalno] (https://scholar.google.hr/)
    - Galar, Mikel, et al. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.4 (2012): 463-484. PDF
  - Korisni alati:
    - https://github.com/scikit-learn-contrib/imbalanced-learn
    - Pogledati cost sensitive classification u weki.

Klasifikacija izborom atributa (podtipovi leukemije)

- **Autori:** Barbara Prkačin, Edi Ibriks
- **Opis**
  - AML/ALL classification dataset
    - https://software.broadinstitute.org/software/cprg/?q=node/55
  - Binarna klasifikacija na tkz. HDLSS (*High Dimensional Low Sample Size*) podacima
    - 72 primjera (47 ALL i 25 AML) sa 7129 gena
    - Fokus na metode za izbor atributa (imlp. *Releif* algoritma u okruženju *Wekka*)
- **Ocjena projektnog prijedloga:**
  - – podatkovni skup nije referenciran niti dobro opisan
  - – nedostaje opis dosadašnjih istraživanja
  - – metodologija se fokusira samo na izbor atributa spominjanjem jednog pristupa (Releif), bez osvrta na druge metode za smanjenje dimenzionalnosti i bez spominjanja metoda za klasifikaciju
  - – tehnički dojam: nisu specificirana imena autora projektnog prijedloga, popis literature je nedostatan i literatura nije referencira u tekstu
- **Komentari i prijedlozi:**
  - Istražiti dosadašnje pristupe na odabranim podacima i problemu (koristiti https://scholar.google.hr/) i istražiti potencijalno dodatne podatkovne skupove:
    - http://eps.upo.es/aguilar/datasets.html
    - http://www.ntu.edu.sg/home/elhchen/data.htm
  - Proučavanje više različitih metoda za izbor atributa (feature selection) i smanjenje dimenzionalnost (dimensionality reduction)
  - Probati koristiti Random forest sa malo većom šumom.
  - BONUS: općenito proučiti problem i načine analize HDLSS (*High Dimensional Low Sample Size*) podataka – izdvojiti nekoliko dodatnih skupova osim predloženog prijedlogom

Binarna klasifikacija slika

- **Autori:** Mirjana Jukić-Bračulj, Karla Kanižaj, Tihana Britvić
- **Opis:**
  - Skup podataka tvrtke Microblink
  - Binarna klasifikacija sa nebalansiranim skupom podataka
    - 13268 slika (10935 slika jednadžbi i 2333 slika teksta)
    - Konvolucijske neuronske mreže (CNN)
- **Ocjena projektnog prijedloga:**
  - + dobro motiviran i opisan problem i podatkovni skup
  - – površno spomenuta dosadašnja istraživanja i pristupi za izabrani problem
  - + metodologija je primjereno izabrana i opisana s fokusom na end-to-end učenje
- **Komentari i prijedlozi:**
- Promijeniti naslov projekta u specifičniji: „Binarna klasifikacija slika na tekst i jednadžbu“
  - Razmotriti korištenje **CNN arhitektura** koje su se pokazale dobrim u praksi (s fokusom na one za ImageNet problem: VGG, Inscpetion, Xception i za bonus AlexNet u okruženju tensorflow/keras)
  - Razmotriti korištenje **istreniranih CNN modela** za ImageNet problem njihovih pretreniranjem i prilgodbom (engl. *finetuning*) za izabrani problem umjesto počinjanja učenja iz početka
  - BONUS: razmotriti utjecaj nebalansiranih podataka prilikom učenja CNN modela, https://www.kth.se/social/files/588617ebf2765401cfcc478c/PHensmanDMasko_dkand15.pdf

Super-rezolucija fotografija lica

- **Autori:** Magda Klarić, Tomislav Levanić
- **Opis:**
  - *CelebFaces Attributes Dataset* (CelebA)
    - 10177 različitih osoba koje su prikazane kroz 202599 slika lica, od kojih svaka ima zapisane vrijednosti za 40 binarnih atributa
  - Cilj je udvostručenje rezolucije slike (npr. 89 × 109 --> 178 x 218)
- **Ocjena projektnog prijedloga:**
  - + dobar opis, zanimljiv i specifičan problem i podaci
  - + dobar opis dosadašnjih istraživanja
  - +/- primjereno izabrana state of the art metodologija i dostatno opisana, no uz nedostatak određivanja kvalitete i uspješnosti rezultata
- **Komentari i prijedlozi:**
  - https://github.com/carpedm20/DCGAN-tensorflow

Grupiranje slika prema njihovom sadržaju

- **Autori:** Terezija Ćosić, Iva Sović, Martina Barišić
- **Opis:**
  - Mozgalo 2017: Analiza slike
    - https://www.facebook.com/events/1863873213853692/
    - 7000 slika u boji različitih veličina
  - Cilj proizvoljno grupirati slike prema sadržaju (u najširem značenju)
- **Ocjena projektnog prijedloga:**
  - + zanimljiv i težak originalan problem
  - – metodologija, cilj i očekivani rezultati su nejasno razrađeni
  - – tehnički dojama: literatura nije referencirana u tekstu i nečitljivost predložene mjere sličnosti
- **Komentari i prijedlozi:**
  - Ispitati ponašanje embedding metoda kao t-SNE na podatkovnom skupu u svrhu dobivanja određenog uvida
  - Ručno ispitati i označiti određeni podskup slika (10%)
  - Napraviti novu reprezentaciju svake slike proizvoljnim odabirom značajki visoke i niske razine (*high-level* and *low level features*), a ne samo analizirati odnos vektoriziranih slika zasnovanih samo na vrijednostima pikselima
    - *High-level features*:
      - dobivanje koristeći prethodno naučene CNN modele za klasifikaciju slika na ImageNet skupu
        - https://www.tensorflow.org/tutorials/image_recognition
        - https://github.com/BVLC/caffe/wiki/Model-Zoo
      - Detekcija rubova
      - *SIFT* značajke, http://image-net.org/download-features
    - *Low level features:* obradom slike (image processing) odrediti/konstruirati značajke (npr. udio određene boje na slici, broj boj) ili primjernom PCA i NMF
  - Na novom transformiranom podatkovnom skupu sa dobrom reprezentacijom sadržaja potom primjeniti određene pristupe grupiranja
  - S obzirom na metode opisane u prijedlogu, jedna mogućnost je i upotreba supervised modela nakon što su definirane kategorije (dobivene korištenjem metoda klasteriranja).

Quora Question Pairs

- **Autori:** Elena Petek, Dora Mifka, Nensi Babić, Lovre Grzunov
- **Opis:**
  - @Kaggle: Quora Question Pairs - Can you identify question pairs that have the same intent?
    - https://www.kaggle.com/c/quora-question-pairs
  - Binarna klasifikacija s tekstualnim podacima (priprema značajki korištenjem metoda za obradu prirodnog jezika – NLP)
  - Baseline model za usporedbu: RF
  - Predloženi modeli: NB, MaxEnt, SVM
  - Mjera uspješnosti: *Logarithmic Loss*
- **Ocjena projektnog prijedloga:**
  - +/– u projektnom prijedlogu metodologija ne opisuje jasno način pripreme značajki iz tekstualnih podataka, dok je u prezentaciji predstavljeno korištenje sintaksnog stabla
  - – nepostojeći očekivani rezultati
- **Komentari i prijedlozi:**
  - Korištenjem Tree/Graph Kenels za prilagodbu sintaksnog stable za problem binarne klasifikacije
    - Zhang, Wei, Peifeng Li, and Qiaoming Zhu. "Sentiment classification based on syntax tree pruning and tree kernel." Web Information Systems and Applications Conference (WISA), 2010 7th. IEEE, 2010. https://www.researchgate.net/publication/224176235_Sentiment_Classification_Based_on_Syntax_Tree_Pruning_and_Tree_Kernel
  - Preporuča se proučiti – drugi pristup pripreme značajki iz teksta:
    - *Word embeddings: https://en.wikipedia.org/wiki/Word_embedding*
      - http://blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove/
      - http://ahogrammer.com/2017/01/20/the-list-of-pretrained-word-embeddings/
      - https://www.tensorflow.org/tutorials/word2vec
      - https://en.wikipedia.org/wiki/Word2vec
      - https://www.slideshare.net/mlprague/tom-mikolov-distributed-representations-for-nlp
      - https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
    - When Are Tree Structures Necessary for Deep Learning of Representations? https://nlp.stanford.edu/pubs/emnlp2015_2_jiwei.pdf
    - Za otvaranje skupa podataka preporuča se korištenje alata Notepad++. U slučaju da su podaci preveliki, može se koristiti Vim editor.
    - Primjeri u datasetu su parovi, možda moguće iskoristiti tranzitivnost ukoliko postoji više od dva duplikata.

Daily News for Stock Market Prediction

- **Autori:** Domagoj Babić, Dario Zorko, Dinko Ždravac
- **Opis:**
  - @Kaggle: https://www.kaggle.com/aaron7sun/stocknews
  - Binarna klasifikacija s tekstualnim podacima i vremenskom komponentom
  -
- **Ocjena projektnog prijedloga:**
  - +/- dobar opis dosadašnjih istraživanja, no s nedostatkom osvrta na odabrani specifični *Kaggle* problem i dosadašnje pristupe njegova rješavanja osvtom na dostupna rješenja u Kaggle Kernels
  - – u metodologiji nije spomenut način obrade teksta niti metode za učenje klasifikatora
  - – nije skroz jasno kako i s kojim ciljem će se koristiti metode za nenadzirano učenje unutar projekta.
  - – tehnički dojam i stil: na nekoliko mjesta ponavljanje i nesuvislost argumenata, literatura nije referencirana unutar teksta
- **Komentari i prijedlozi:**
  - Osvrnuti se na analizu iz:
    - http://blog.kaggle.com/2016/10/27/open-data-spotlight-daily-news-for-stock-market-prediction-jiahao-sun/
    - https://www.kaggle.com/ndrewgele/omg-nlp-with-the-djia-and-reddit
  - Proučiti metode za Natural language processing (NLP):
    - https://opennlp.apache.org/
    - https://www.tutorialspoint.com/opennlp/index.htm
    - http://www.nltk.org/
  - Pripremiti i ocjeniti utjecaj korištenja podataka iz prošlog perioda za primjer danog vremenskog trenutaka - lagged data
  - Napraviti prikladnu metodologiju evaluacije i odabira više različitih klasifikacijskih modela nad određenim skupom značajki
  - OPCIONALNO: primjeniti jedan od pristupa odabira značajki
  - BONUS: naučiti i primjeniti xgboost za klasifikaciju

Kratkotrajna kretanja u cijenama dionica

- **Autori:** Ante Mijoč, Porin Ćustić
- **Opis:**
  - @Kaggle: INFORMS Data Mining Contest 2010
    - https://www.kaggle.com/c/informs2010#description
  - Binarna klasifikacija na vremenskim serijama
  - Predloženo korištenje logističke regresije
  - Mjera uspješnosti: Area Under the ROC Curve (AUC)
- **Ocjena projektnog prijedloga:**
  - + dobro opisan skup podataka
  - – metodologija ne opisuje značajke modela koji će se učtiti
  - – očekivani rezultati nisu opisani u kontekstu Kaggle-a
  - – tehnički dojam: literatura nedostatna i nije referencirana unutar teksta
- **Komentari i prijedlozi:**
  - Osvrnuti se na analizu pobjednika natjecanja:
    - https://blog.kaggle.com/2010/10/11/how-i-did-it-the-top-three-from-the-2010-informs-data-mining-contest/
  - Pripremiti i ocjeniti utjecaj korištenja podataka iz prošlog perioda za primjer danog vremenskog trenutaka - lagged data
  - Napraviti prikladnu metodologiju evaluacije i odabira više različitih klasifikacijskih modela nad određenim skupom značajki
  - OPCIONALNO: primjeniti jedan od pristupa odabira značajki
  - BONUS: naučiti i primjeniti xgboost za klasifikaciju
  - Može se isprobati korištenje KStar klasifikacijskog algoritma.

Predviđanje kretanja dionica na Zagrebačkoj burzi

- **Autori:** Ana Lukačić, Vedran Rukavina, Tin Deranja
- **Opis:**
  - Podatkovni skup: vlastito pripremljen koristeći Agram broker sustav
  - Analiza i predviđanje vremenskih serija visoke frekvencije (high frequency time series forecasting)
  - PSO-LS- SVM, LS-SVM vs neuronske mreže (NN)
- **Ocjena projektnog prijedloga:**
  - + spomenut je glavni članak koji služi kao inspiracija za projekt
  - – nisu opisane metode ni očekivani rezultati
- **Komentari i prijedlozi:**
  - Evaluirati više regresijskih modela s automatiziranim pristupom odabira najuspješnijeg skupa značajki
  - Ovdje bi vjerojatno bilo korisno imati domensko znanje oko vlasničke structure i slično…

Procjena broja iznajmljenih bicikala

- **Autori:** Mario Skočić
- **Opis:**
  - @Kaggle: Bike Sharing Demand
    - https://www.kaggle.com/c/bike-sharing-demand
  - Predviđanje vrijednosti vremenskih serija
  - Linearna regresija
  - Mjera uspješnosti: RMSLE (Root Mean Squared Logarithmic Error)
- **Ocjena projektnog prijedloga:**
  - + opis podataka i podataka uz provedenu eksploratornu analizu
  - - od metoda je spomenuta samo linearna regresija
  - - nepostojeća literature
  - –nepostojeći related work
- **Komentari i prijedlozi:**
  - Evaluirati više regresijskih modela s automatiziranim pristupom odabira najuspješnijeg skupa značajki