Предлог пројекта из СИАП-а

Овим документом дат је предлог пројекта из предмета Системи за истраживање и анализу података. Документ садржи кратак опис теме пројекта, дефиницију и мотивацију за одабрану тему. Такође наведени су и укратко описани радови који се баве решавањем истог проблема као и овај рад разним методологијама. Затим описани су скуп података, методологије решавања проблема, методи евалуације, софтвер који ће бити коришћен и на крају угрубо план којим ће имплементације тећи.

Дефиниција пројекта

Тема рада јесте предикција квалитета вина. Идеја овог пројекта јесте да процени квалитет вина на скали од 0 до 10 на основу примљених података о истом.

Мотивација

Љубитељи вина би могли да уживају у винима без потребе да знају ишта о њима. Куповина вина лако може бити пун погодак или промашај ако нисмо експерти који процене врше на основну године бербе, цијене и разних других параметара. Циљ овог пројекта је повећање успешности проналаска што бољег вина не ослањајући се на соптсвено или на експертско знање о винима.

Релевантна литература

Први рад:

[1] Yogesh Gupta (December 2017) *Selection of important features and predicting wine quality using machine learning techniques*

Линк до рада

Задатак: Уз помоћ линеарне регресије, неуронских мрежа и SVM-а одредити зависност квалитета вина на основу 11 физичких и хемијских карактеристика.

Методологија: Због великих разлика у појединим вредностима атрибута, те већег утицаја појединих на предикцију, вршено је предпроцесирање. Предпроцесирање је извршено дељењем вредности сваког атрибута максималном вредношћу тог атрибута у скупу података. Уз помоћ линеарне регресије одређена је зависност атрибута "оцена квалитета" у односу на осталих 11 атрибута. На овај начим могуће је одабрати погодне атрибуте за предикцију квалитета. Неуронске мреже и SVM коришћени су за предикцију квалитета на основу свих, и на основу најпогоднијих атрибута који су одабрани у претходном кораку.

Скуп података: За скуп података коришћен је скуп вина који се садржи од 4898 узорака белог, и 1599 узорака црвеног вина. Сваки од узорака се састоји од 12 атрибута који представљају поменуте физичке и хемијске карактеристике. Атрибути: фиксна киселост (fixed acidity), испарљива киселост (volatile acidity), лимунска киселина (citric acid), резидуални шећер (residual sugar), хлориди (chlorides), слободни сумпор диоксид (free sulfur dioxide), укупан сумпор диоксид (total sulfur dioxide), густина (density), ниво киселости/базичности (pH), сулфати (sulphates), алкохол (alcohol) и оцена квалитета (quality rating). Вредност атрибута оцене квалитета конструисана је на основу тестирања вина барем три сомелијера који су своју оцену дали на склаи од 0 (веома лош квалитет) до 10 (веома одличан) **Error! Reference source not found.**.

Евалуација: Решење је евалуирано поделом на тренинг, тест и валидациони скуп. Метрика коришћена за евалуацију није наведена али је предпоставка да је коришћена тачност (accuracy).

Резутати: У раду се показало да се прецизност класификације знатно повећала када се у процес класификације укључују само обележја која су се показала као релевантна линеарном регресијом. Такође, смањује се димензионалност скупа, што у многоме повећава брзину учења.

Закључак: Након читања рада закључили смо да би било добро вршити одабир релевантних обележја уз помоћ линеарне регресије. Што се тиче класификације, највероватније ћемо користити SVM и неуронске мреже, с тим да ћемо дати предност новијим и напреднијим приступима уколико будемо морали да бирамо. За евалуацију користићемо метод унакрсне валидације (K - fold validation), али није искључено да ћемо радити и поделу на тренинг/тест/валидациони скуп.

Други рад:

[2] Sunny Kumar, Kanika Agrawal, Nelshan Mandan (January 2020) *Red Wine Quality Prediction Using Machine Learning Techniques*

Линк до рада

Задатак: Предикција квалитета црвеног вине уз помоћ техника машинског учења, као што су *Random Forest*, *SVM* і *Naive Bayes*.

Методологија: Коришћени и међусобно су упоређени Naive Bayes алгоритам, SVM (метод подржавајућих вектора) и Random Forest техника.

Скуп података: За скуп података је коришћен скуп од 1599 узорака црвеног вина. Сваки узорак има 12 атрибута, који се поклапају са атрибутима који се налазе у скупу података овог пројекта. Квалитет је пресликан на вредности у опсегу од 3 до 8, квалитет 3 је лош, док 8 представља одличан квалитет **Error! Reference source not found.**.

Евалуација: Скуп података је подељен на 70% тренинг и 30% тест скуп. За евалуацију су коришћени различити параметри мере као што су *precision*, *recall*, *specificity*, *f-measure*, *accuracy* и *misclassification*.

Резутати: У овом раду се SVM метод показао као најбоље решење. Показује добре перформансе на овом релативно малом скупу података.

Недостатак: Није споменуто претпроцесирање података, у виду нормализације или стандардизације, које често побољша модел, јер не дозвољава превелик утицај неких обележја на основу њихових вредности.

Закључак: Након прочитаног рада, за класификацију ћемо користити SVM метод, показао се најбоље у решавању овог проблема над скупом података који представља подскуп нашег (у случају овог рада само црвена вина), такође ћемо користити исте параметре мере за евалуацију, план је користити унакрсну валидацију, али могуће да ћемо испробати и 70% - 30% поделу скупа.

Трећи рад:

[3] Dahal, K. , Dahal, J. , Banjade, H. and Gaire, S. (2021) *Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics*

Линк до рада

Задатак: Предикција квалитета вина коришћењем алгоритама машинског учења.

Методологија: Коришћени су *Ridge Regression*, SVM, *Gradient Boosting Regression* (GBR) и *Artificial Neural Network* (ANN). Пре употребе наведених методологија извршено је претпроцесирање података, како би модели били ефективнији, у овом случају коришћена је стандардизација.

Скуп података: За скуп података је такође употребљен подскуп скупа података који ће се користити при изради пројекта, то јест посматрају се само узорци црвеног вина (1599 узорака). Обележја се поклапају и квалитет је пресликан на опсег од 3 до 8 **Error! Reference source not found.**.

Евалуација: За евалуацију су коришћени различити параметри мере као што су *Mean Squared Error* (MSE), *Mean Absolute Percentage Error* (MAPE) и коефицијент корелације (R). За одређивање вредности хиперпараметара извршена је унакрсна валидација (*10-fold cross-validation*)

Резутати: У овом раду се GBR показао као најбољи приступ, неуронска мрежа је подбацила у прецизности највероватније због релативно малог и некомплексног скупа података. За бољи рад модела извршена је стандардизација, како нека обележја због својих вредности не би преузимала примат над другима.

Закључак: Након прочитаног рада, у оптицај долази имплементација GBR модела као најбоље показаног, и имплементација неуронских мрежа јер користимо доста већи скуп података (6497 узорака). Користићемо унакрсну валидацију, навероватније *5-fold cross-validation*.

Остале референце

Напомена: пошто сви радови поменути, укључујући и овај, користе исти скуп података или његов подскуп (нпр. само црвена вина), стављен је само један линк.

[4] Линк до скупа података

Скуп података

Скуп података везан за варијанте црног и бијелог португалског вина Винхо Верде. Скуп се састоји од 6497 вина са припадајућим обележјима. Свако вино је описано са 13 обележја, а то су: тип вина (црно 25% или бело 75%), фиксна киселост (*fixed acidity*, минимална вредност: 3.9 - максимална: 15.9, просек: 7.22), испарљива киселост (*volatile acidity*, 0.08 - 1.58, 0.34), лимунска киселина (*citric acid*, 0 – 1.66, 0.32), преостали шећери (*residual sugar*, 0.6 - 65.8, 5.44), хлориди (*chlorides*, 0.01 - 0.61, 0.06), слободни сумпор диоксид (*free sulfur dioxide*, 1 – 289, 30.5), укупни сумпор диоксид (*total sulfur dioxide*, 6 – 440, 116), густина (*density*, 0.99 - 1.04, 0.99), pH (2.72 - 4.01, 3.22), сулфати (*sulphates*, 0.22 - 2, 0.53), алкохол (*alcohol*, 8 – 14.9, 10.5) и на крају излазни податак квалитет (*quality*) намапиран на распон вредности од 0 до 10. Детаљнији опис скупа података и обележја се може пронаћи у **Error! Reference source not found.**.

Методологија

Пре свега извршићемо претпроцесирање података са нормализацијом и стандардизацијом, а затим ћемо користити:

- вештачке неуронске мреже
- методе линеарне, полиномне регресије
- GBR (Gradient Boosting Regression)
- SVM

Метод евалуације

За евалуацију класификације ће се користити F1 меру, док за регресију ће се употребити RMSE.

Софтвер

Пројекат ће бити имплементиран у *Python* програмског језика унутар *Visual Studio Code* развојног окружења.

План

План рада на овом пројекту обухвата:

- претпроцесирање података
- креирање модела
- испробавање модела

Тим

Тим чине: **М**арко **Б**јелица (R2 10/2022) и **В**ељко **Т**ошић (R2 4/2022).