UDC: 004.7 Info M str. 27-32

КВАНТИФИКАЦИЈА КАТЕГОРИЧКИХ АТРИБУТА ЗА ПРИМЕНУ АЛГОРИТАМА МАШИНСКОГ УЧЕЊА
ENCODING OF CATEGORICAL ATTRIBUTES FOR MACHINE LEARNING ALGORITHMS

Алекса Милосављевић, др Милан Вукићевић

РЕЗИМЕ: Овај рад истражује различите методе квантификације категоричких атрибута за примену алгоритама машинског учења. Категорички атрибути најчешће описују квалитативне карактеристике, често пристутне у реалним скуповима података, попут пола, боје, образовног нивоа, позиције у фирми и слично. Квантификација таквих атрибута је важна јер већина алгоритама машинског учења захтева нумеричке улазе. Игнорисање категоричких (квалитативних) атрибута може значајно да смањи количину информација у скуповима података и последично да утиче на квалитет изграђених модела. У овом раду ће бити описане и примењене неке од популарних метода за квантификацију попут One-Hot encoding-a, Target encoding-a, Count encoding-a и других. У оквиру експерименталне евалуације ове методе биће примењене над неколико скупова података који садрже различите типове категоричких атрибута, а затим ће се након примена алгоритама машинског учења сагледати и њихове перформансе. Резултати истраживања пружиће дубље разумевање метода за квантификацију и њихових ефеката на перформансе алгоритама.

КЉУЧНЕ РЕЧИ: Машинско учење, категорички атрибути, квантификација, перформансе алгоритама, класификација, припрема података

ABSTRACT: This paper explores various methods for encoding categorical attributes for the application of machine learning algorithms. Categorical attributes usually represent qualitative characteristics, commonly present in real-world datasets, such as gender, color, educational level, item category, and similar. Encoding such attributes is important because most machine learning algorithms require numerical inputs, and the lack of encoding of categorical attributes can significantly impact the efficiency and performance of the algorithm. The paper will present several methods for quantification, such as One-Hot encoding, Target encoding, Count encoding, and others. These methods will be applied to several datasets containing different types of categorical attributes, and after applying machine learning algorithms, their performance will be evaluated. The research results will provide a deeper understanding of encoding methods and their effects on algorithm performance.

KEY WORDS: Machine learning, categorical attributes, encoding, algorithm performance, classification, data preprocessing

1. УВОД

Категорички подаци се угрубо могу поделити на номиналне и ординалне. Номинални категорички подаци представљају категорије које имају чисто описни карактер. Између њих нема разлике у вредости. Пример номиналних података могла би бити боја аутомобила, поштански код, пол или бренд. Ординалне категоричке вредности са друге стране имају углавном јасан поредак. На пример, степен образовања где су могуће вредности „средња школа“, „основне студије“, „мастер“ или „докторат“. Без обзира што овакве вредности имају неки редослед, проблем постоји у прецизношћу дефинисања колике су разлике између постојећих категорија и како их представити нумерички.[1]

Категорички подаци, попут жанра неког филма или музике, марке неког производа или оцене неке услуге, често носе важне квалитативне информације али се не могу једноставно превести у квантитативни облик који је погодан за учење модела машинског учења. Због тога су развијене бројне технике за квантификацију категоричких варијабли које омогућавају моделима машинског учења да ефикасно тумаче и користе ове податке иако је већина алгоритама машинског учења дизајнирана да ради само са нумеричким подацима.[1]

Неки од проблема са којима се алгоритам машинског учења може суочити уколико се категорички атрибути не квантификују на адекватан начин су погрешно тумачење важности категорије, претренирање због ретких категорија, појављивање категорија у тест подацима које се нису појавиле у тренинг подацима, цурење информација и слично.

Кроз примену метода за квантификацију категоричких вредности над скуповима података биће извршена компаративна анализа тих метода и биће упоређени добијени резултати. Сврха је да се кроз резултате дође до закључка која од метода је најадекватнија у зависности од врсте категоричког атрибута и од података који он садржи.

На основу сврхе и циља овог рада, ово истраживање ће се водити следећим истраживачким питањима, кроз чије одговоре ће се доћи до тражених резултата и закључка:

1) Које врсте категоричких података су непходне како би се успешно спровео експеримент?
2) У зависности од скупа података која метода би била најадекватнија?
3) Каква трансформација се врши над категоричким атрибутом у зависности од методе која је примењена?
4) Који су резултати добијени након примене метода и која је дала најбољи резултат?

2. ПРЕГЛЕД ЛИТЕРАТУРЕ

У раду [2] истражује се утицај квантификације неколико различитих метода над подацима који се баве пре-

варама кредитним картицама. Скуп података поседује велики број редова и садржи више категоричких атрибута попут категорије продавца, типа картице, државе и слично. Експеримент је спроведен применом метода квантификације које се заснивају на концептима *Target*-а *Weight of Evidence*-а. Примењене су методе *Target encoder, M-estimate, Catboost encoding, Pozzolo* и *James-Stein* а алгоритми који су коришћени су *LightGBM, CatBoost, XGBoost.* Методе се пореде коришћењем мера евалуације попут прецизности, одзива, *F1* мере и површине испод *ROC* криве. Резултати показују да за *LightGBM* најбоље резултате дају *CatBoost* и *Weight of Evidence* методе, за *CatBoost* алгоритам најбоље резултате даје његова уграђена *CatBoost* метода док за *XGBoost* алгоритам најбоље резултате постигла је *Target encoding* метода. Као закључак аутори наводе да је *CatBoost* најбоља метода за квантификацију код проблема препознавања превара.

Рад [3] тестира ефекат 10 метода квантификације за алгоритме машинског учења: *Ordinal, One-Hot, Sum, Helmert, Backward Diference, Target, M-estimate, Leave One Out, CatBoost, James Stein*. За експеримент коришћено је пет алгоритама машинског учења (логистичка регресија, наивни Бајес, метода потпорних вектора, неуронске мреже и *XGBoost*). Методе су примењене над седам реалних скупова података, али и над вештачким подацима које аутори користе јер сматрају да се над таквим подацима прецизиније може проценити какав ефекат методе имају. Са вештачким подацима аутори избегавају неопходну припрему и чишћење података, јаснија је веза између зависне промењљиве и осталих и може се креирати произвољан број категоричких и нумеричких атрибута. Резултати на реалним скуповима података указују да методе квантификације могу имати различити утицај на перформансе алгоритама. Из резултата на реалним скуповима података извучени су следећи закључци:

- За неуронске мреже, методу потпорних вектора и логистичку регресију, избор методе квантификације није имао превелики утицај на разлику међу резултатима.
- За алгоритам Наивни Бајес лошији резултати постизани су са методама које увећавају скуп података.
- Најлошији резултати најчешће су добијани када су коришћене *Ordinal* и *M-estimate* методе.

Из резултата који су добијени са вештачким подацима извучени су ови закључци:

- Неуронске мреже, Наивни Бајес и метода потпорних вектора најлошије су радили са методама које увећавају скуп података, док је логистичка регресија радила боље.
- Међу најбољим резултатима најчешће се јављала метода *CatBoost*, а уз њу истакле су се и *Leave One Out* и *Ordinal* методе.
- Најлошији резултати су најчешће постизани уз *Backward difference encoder.*

Постоји још доста радова који се баве истраживањем разних метода квантификације у различитим сценаријима и над различитим подацима попут [4] где је фокус на поређењу *target-based* и *target-agnostic* група метода, или обиман рад [5] који пореди чак 32 методе над 50 скупова података. Рад [7] се бави проблемом рада са категоричким подацима који имају високу кардиналност, и предлаже методе *Gamma-Poisson* и *min-hash* методе као алтернативу *One-Hot encoding*-у.

Овај рад, има слично експерименталну поставку као претходно наведени радови где се пореде комбинације метода и алгоритама над скуповима података. Додатна пажња посвећена је проблему који се појављује у неколико наведених радова, а то је коришћење метода квантификације које увећавају скуп података у комбинацији са категоричким атрибутима високе кардиналности. У раду је приказана техника *binning* као решење и њен утицај на резултате експеримента.

3. МЕТОДЕ ИСТРАЖИВАЊА

Постоји велики велики број метода које се данас користе у науци о подацима. Оне функционишу на различите начине и зато је битно разумети начин на који оне врше трансформацију како би се одабрала најадекватинија за конкрентан случај. У овом раду истражује се шест метода квантификације.

One-Hot encoding је метода квантификације категоричких вредности која претвара категорије у бинарне векторе који алгоритам може да разуме. Због своје једноставности представља једну од најкоришћенијих метода квантификације. За сваку могућу вредност у неком скупу категорија *One-Hot encoding* креира нову колону која садржи вредности 1 и 0, тако да 1 означава присуство те категорије у неком реду, а 0 одсуство.[1]

Count encoding који се често назива *Frequency encoding* представља једноставну методу која квантификује категоричке вредности узимајући у обзир број њиховог понављања. Свака категорија се мења бројем понављања те категорије међу инстанцама у скупу података. Често се уместо броја мења и процентом који нека категорија чини од укупног броја опсервација.[1]

Target encoding је метода квантификације која припада групи енкодера која узима у обзир и излазну промењљиву при трансформисању категорија. Идеја иза ове методе је да се свака могућа вредност категорије замени са средном вредношћу излазног атрибута за ту категорију. Зато се често назива и *Mean encoding*.[1]

Weight of Evidence као и *Target encoder* припада групи метода које квантификују категорије ослањајући се на излазни атрибут. Ова метода се често користи у проблемима који имају небалансирани излазни атрибут, попут проблема који се баве кредитним ризиком. *Weight of Evidence (WoE)* представља однос удела „позитивних“ тј. исхода где је *target=1* и „негативних“ исхода излазног атрибута где је *target=0*.[30]

$$WOE = ln\left(\frac{\text{Дистрибуција "позитивних" исхода у категорији}}{\text{Дистрибуција "негативних" исхода у категорији}}\right)$$

Backward Difference encoding је метода која спада у групу *Contrast encoder*-а и која добро ради са ординалним категоријама. Она тренутни ниво категорије пореди са претходним нивоима (нивои категорије могу на пример бити *„low“*, *„medium“* и *„high“*) и креира k-1 нових бинарних колона за k категорија у атрибуту. Категорије се квантификују по следећој шеми, где *k* представља кардиналност категорије.[31]

Табела 1. *Шема за Backward Difference encoding*

	Контраст 1	Контраст 2	Контраст 3
Ниво	ниво 1 *vs.* ниво 2	ниво 2 *vs.* ниво 3	ниво 3 *vs.* ниво 4
Novice	-(k-1)/k	-(k-2)/k	-(k-3)/k
Contributor	1/k	-(k-2)/k	-(k-3)/k
Master	1/k	2/k	-(k-3)/k
Grandmaster	1/k	2/k	3/k

Catboost encoding је метода која се најшешће користи у оквиру *CatBoost* алгоритма али се може користити и засебно као *CatBoostEncoder*. Функционише по сличном принципу као и *Target encoder* али је побољшан тако да умањује могућност од цурења информација и претренирања. *CatBoost encoder* мења вредност категорије са просечном вредношћу излазног атрибута из претходних редова где се та категорија јавља. Квантификација се врши коришћењем следеће формуле:

$$\frac{TargetSum + \text{prior}}{FeatureCount + 1}$$

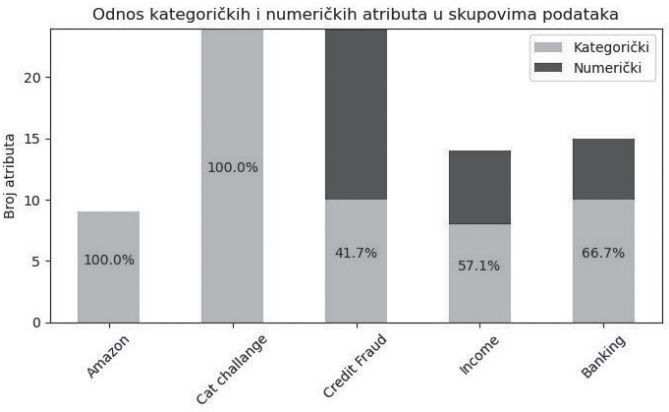
где *TargetSum* представља суму вредности излазне променљиве све до тренутне опсевације, *Prior* означава константну вредност која се рачуна као сума вредности излазне променљиве подељена са укупним бројем опсервација а *FeatureCount* означава број пута који се та категорија појавила до тренутне опсевације.[32]

4. ЕКСПЕРИМЕНТАЛНА ЕВАЛУАЦИЈА

У овом делу рада биће спроведена експериментална евалуација метода које се истражују. Експеримент је извршен коришћењем програмског језика *Python* и његових библиотека за манипулацију подацима и машинско учење, а окружење у оквиру којег је експеримент спроведен је *Jupyter Notebook*.

За спровођење експеримента користи се пет скупова података преузетих са платформе *Kaggle*. Сваки од скупова података поседује барем неколико категоричких атрибута. У скуповима се налазе више типова атрибута попут бинарних, номиналних и ординалних који носе информације о полу, позицији, локацији, занимањима и многим сличним атрибутима који могу носити важне информације. Ови скупови су изабрани зато што садрже велики број разноврсних категоричких атрибута, који имају различите нивое кардиналности, од ниске до високе.

На следећем графикону може се видети удео који категорички атрибуте чине од укупног броја атрибута у скуповима података који се користе у експерименту.



Слика 1. *Број атрибута у скуповима података и удео категоричких*

Amazon скуп података садржи податке о запосленима и њиховим карактеристикама попут њихове позиције, сектора којем припадају, ко им је надређени итд. Циљ је предвидети да ли одређеном запосленом, на основу његових карактеристика, треба дозволити приступ неком ресурсу или не и на тај начин аутоматизовати посао који мануелно морају да раде надређени. Сви атрибути у скупу података су категоричке вредности које су представљене нумеричким шифрама.

Categorical Feature Encoding Challenge скуп података је специјално направљен за истраживање метода квантификације категоричких атрибута и садржи само категоричке атрибуте. Садржи 23 атрибута од којих су неки бинарног, неки ординалног а неки номиланог типа. У овом скупу предвиђа се атрибут *target* који садржи вредности 1 и 0.

Credit Card Transactions Fraud Detection Dataset скуп података представља симулиране податке о правим и лажим трансакцијама са кредитним картицама. У скупу података налази се 23 атрибута заједно са излазним атрибутом. Од тога је 10 категоричких атрибута.. Атрибут који се предвиђа је *is_fraud* који говори да ли је трансакција превара или не.

Income classification скуп података садржи податке о особама који описују њихове демографске карактеристике. Излазни атрибут који треба предвидети говори о томе да ли особа има приходе веће или мање од 50 хиљада. Скуп података садржи 15 атрибута укључујући и атрибут који се предвиђа. Постоји 9 категоричних атрибута и 6 нумеричких. Атрибут који се предвиђа *income* садржи вредности <=50K и >50K које говоре о годишњим приходима.

Banking Dataset Classification скуп података садржи податке о карактериситикама клијената банке Португала. Банка жели да идентификује оне клијенте за које постоји већа вероватноћа да ће уложити у дугорочни депозит како би фокусирала своју маркетнг кампању на њих. Потребно је предвидети излазни атрибут који говори да клијент улаже у дугорочни депозит или не. Скуп података саджи 16 атрибута укључујући и излазни атрибут. Постоји 10 категоричких и 5 нумеричких атрибута. Излазни атрибут *у* саджи вредности *yes* и *no*.

Алгоритми машинског учења који се примењују у експерименту су три основна алгоритма, логистичка регресија, *k* најближих суседа и стабло одлучивања, као и два ансамбл алгоритма, случајне шуме (енг. *random forest*) и градијентно појачавање (енг. *gradient boosting)*. Као мера евалуације користиће се *F1-score* због тога што у свим скуповима података атрибут који се предвиђа има небалансирану дистрибуцију класа.

Како је фокус овог рада утицај метода квантификације на алгоритме машинског учења за сваки скуп података урађена је једноставна припрема података попут уклањања/попуњавања недостајућих вредности, скалирања и узорковања.

Свака метода квантификације је тестирана са сваким алгоритмом машинског учења над свих пет скупова података. Скупови података подељени су на скупове за тренинг и тест. За оптимизацију хиперпараметара и унакрсну валидацију коришћен је *GridSearchCV*. При учењу модела машинског учења, оптимизовани су следећи хиперпараметри:

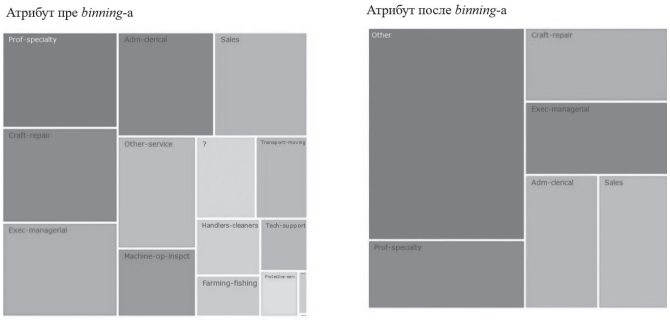
- За логистичку регресију оптимизују се следећи параметри: *solver: liblinear*, *penalty: l1*, *C: 0.01, 0.1, 0.8, 1.5*, *max_iter: 1000;*
- За стабло одлучивања: *criterion: gini, entropy*; *max_depth: 5, 10, 20;*
- За *k* најближих суседа: *n_neighbors: 3, 5, 10; weights: uniform, distance;*
- За случајне шуме: *n_estimators: 50, 100, 150; criterion: gini, entropy; max_depth: 5, 10, 20;*
- За градијентно појачавање: *n_estimators: 50, 80, 120; learning_rate: 0.01, 0.1; max_depth: 3, 5;*

5. МОДИФИКАЦИЈА ЕКСПЕРИМЕНТА – *BINNING*

С обзиром да се комбинује пет скупова података са методама квантификације и алгоритмима машинског учења, једна од главних препрека при извршавању овог експеримента је велика временска и меморијска захтевност. Томе највише доприноси то што неки скупови података садрже велики број категоричкх атрибута, од којих неки имају јако високу кардиналност. Висока кардиналност у комбинацији са методама квантификације које увећавају скуп података попут *One-Hot encoding*-а или *Backward difference encoding*-а могу значајно повећати количину података. Једна од метода којом се такви проблеми могу умањити је *binning,* односно груписање одређеног броја категорија у једну како би се умањила кардиналност.

Груписање категорија може се вршити на многе начине, попут коришћења доменског знања, груписања категорија узимајући излазни атрибут у обзир, по уделу неких категорија у скупу података итд. У овој модификацији експеримента категорије ће се груписати на основу њихове фреквенције, тачније категорије које покривају 70% укупне фреквенције су задржане док су остале категорије груписане под новом категоријом *Other*, или -1. Ова метода *binning*-а је посебно корисна у случајевима када је дистрибуција категорија у атрибуту таква да мањи број катеогрија обухвата већи део укупне фреквенције.

На наредној слици се може се видети пример категоичког атрибута пре и после *binning*-а где је кардиналност смањена са 15 категорија на шест.



Слика 2. *Атрибут occupation пре и после груписања*

6. РЕЗУЛТАТИ

Гледајући све скупове података најбоље резултате постиже *Target encoder* који је постигао најбољи резултат на 4 од 5 скупова података, најчешће са једним од два ансамбл алгоритма или са стаблом одлучивања. Уз њега добре резултате постиже и *Weight of Evidence* који најчешће даје најбољи резултат кад је упарен са *Gradient Boosting* алгоритмом, али се показао као најбоља метода у последњем скупу података где је најбољи резултат постигао са *K* најближих суседа. У просеку најлошији резултат постигнут је са *Count encoding*-ом, што је донекле и очекивано с обзиром на његову једноставност.

У другом скупу података добијени резултати су у просеку знатно лошији у поређењу са осталим скуповима података, али пружају дубљи увид у разлике у утицају метода квантификације. Ово је због тога што је овај скуп података вероватно нешто комплекснији и има доста више категоричких атрибута од осталих јер је намењен управо за истраживање квантификације категоричких података. У резултатима знатно одскачу *Target* и *Weight of Evidence encoder,* међутим у комбинацији са *K* најближих суседа дају доста лошији резултат него са осталим моделима. Због овога ове две методе имају и највећу стандардну девијацију. Остале методе дају нешто стабилније резултате у комбинацијама са моделима, где методе које проширују скуп података тј. *One-Hot encoding* и *Backward difference encoding* најбоље раде са логистичком регресијом а *Count* и *CatBoost* и најбоље раде са *Gradient Boost-ом.*

Још један битан аспект код метода квантификације је и њихов утицај на време извршавања и меморијску захтевност при извршавању експеримента. Убедљиво најдуже време извршавања било је код *One-Hot encoding-а* и *Backward difference encoding*-а, пре свега када се комбинују са *Gradient Boosting* алгоритмом. Дуго време извршавања у овим комбинацијама посебно је изражено на прва два скупа података у којим се налазе само категорички атрибути.

Модификација експеримента дала је релативно сличне резултате без неких приметних измена у односу између метода. Свакако најзначајнија промена је знатно смањено

време извршавања, из чега се може закључити да је *binning* у том контексту користан када постоји већи број категоричких атрибута или постоје атрибути са високом кардиналношћу. Уз ову методу свакако постоји ризик губитка информација груписањем категорија тако да треба пажљиво изабрати методу којом ће се категорије груписати.

Просечни резултат (*F1-score*) постигнут над свих пет скупова података за сваку од комбинацију метода-алгоритам може се видети на следећој табели, где су подебљани најбољи резултати који је сваки од алгоритама постигао са неком од шест метода:

Табела 2. *Просечни резултати комбинација над свих пет скупова*

Алгоритам / Метода	Логистичка регресија	Стабло одлучивања	*K* најближих суседа	Случајне шуме	Градијентно појачавање
One-Hot	0.8694	0.8612	0.8435	0.8393	0.8614
Count	0.8412	0.8527	0.8347	0.8549	0.8631
Target	**0.9202**	**0.9215**	**0.8529**	**0.9309**	**0.9293**
WOE	0.9068	0.8987	0.8501	0.9080	0.9123
Backward Difference	0.8666	0.8540	0.8381	0.8410	0.8633
CatBoost	0.8577	0.8445	0.8425	0.8572	0.8657

У табели се може приметити да у просеку најбољи резултат постигнут са *Target* методом за сваки алгоритам.

Просечни резултати за свих пет скупова података по методи могу се видети у наредној табели, сортирани од најбољег до најлошијег резултата.

Табела 3. *Просечни резултати у целокупном експерименту*

Метода	**Просечан резултат**
Target encoding	0.9109
Weight of Evidence encoding	0.8951
One-Hot encoding	0.8549
CatBoost encoding	0.8535
Backward difference encding	0.8526
Count encoding	0.8492

7. ЗАКЉУЧАК И ПРАВЦИ БУДУЋЕГ ИСТРАЖИВАЊА

У овом истраживању по резултатима истакли су се *Target* и *Wegiht of evidence* методе, међутим то не значи да се остале методе могу одбацити и сматрати нужно лошијим. Чињеница је да не постоји једна најбоља метода која ће се увек истицати, већ је генерално најбоља пракса да се за податке над којима се врши експеримент испроба неколико метода и на тај начин одабере она која највише одговара у тој конкретној ситуацији.

Уколико је време битан фактор у експерименту, методе које увећавају скуп података, попут *One-Hot* и *Backward difference* метода, можда нису оптималан избор, поготову у комбинацији са комплексним *boosting* алгоритмима који се извршавају секвенцијално, попут *Gradient Boosting, AdaBoost, LightGBM* итд.

Метода *binning* за смањење кардиналности катеогирчких атрибута се у овом истраживању показала као добра, јер је умањила време извршавања експеримента, а дала је сличне резултате као и у првој инстанци експеримента.

Будуће истраживање могло би се више усмерити на избор методе квантификације у зависности од типа категоричког атрибута (номинални, ординални). За разлику од приступа у овом раду, где је иста метода примењена на цео скуп података, у даљем истраживању би се методе прилагођавале специфичним атрибутима према њиховом типу. На тај начин би експеримент могао да покаже колико додатне информације, сачуване кроз пажљивији избор методе за сваки атрибут, могу да унапреде квалитет и тачност модела машинског учења.

ЛИТЕРАТУРА

[1] Wohlwend, B. (2023, July 16). Converting categorical data into numerical form: A practical guide for data science. датум приступа 01.05.2023., преузето са: https://medium.com/@brandon93.w/converting-categorical-data-into-numerical-form-a-practical-guide-for-data-science-99fdf42d0e10

[2] De La Bourdonnaye, F., & Daniel, F. (2021). Evaluating categorical encoding methods on a real credit card fraud detection database. *arXiv preprint arXiv:2112.12024*.

[3] Valdez-Valenzuela, E., Kuri-Morales, A., & Gomez-Adorno, H. (2021). Measuring the effect of categorical encoders in machine learning tasks using synthetic data. In *Advances in Computational Intelligence: 20th Mexican International Conference on Artificial Intelligence, MICAI 2021, Mexico City, Mexico, October 25–30, 2021, Proceedings, Part I 20* (pp. 92-107). Springer International Publishing.

[4] Breskuvienė, D., & Dzemyda, G. (2023). Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions. *International Journal of Computers Communications & Control*, *18*(3).

[5] Matteucci, F., Arzamasov, V., & Böhm, K. (2024). A benchmark of categorical encoders for binary classification. *Advances in Neural Information Processing Systems*, *36*.

[6] Zhu, W., Qiu, R., & Fu, Y. (2024). Comparative Study on the Performance of Categorical Variable Encoders in Classification and Regression Tasks. *arXiv preprint arXiv:2401.09682*.

[7] Cerda, P., & Varoquaux, G. (2020). Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering*, *34*(3), 1164-1176.

[8] Hosni, M. (2023). Encoding Techniques for Handling Categorical Data in Machine Learning-Based Software Development Effort Estimation.

[9] Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, *175*(4), 7-9.

[10] Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, *175*(4), 7-9.

[11] Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. Machine Learning, 107(8), 1477-1494..

[12] Pyle, D. (1999). Data preparation for data mining. morgan kaufmann.

[13] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

[14] Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD explorations newsletter*, *3*(1), 27-32.
[15] Mougan, C., Masip, D., Nin, J., & Pujol, O. (2021). Quantile encoder: tackling high cardinality categorical features in regression problems. In *International Conference on Modeling Decisions for Artificial Intelligence* (pp. 168-180). Springer, Cham.
[16] Alkharusi, H. (2012). Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education*, *4*(2), 202.
[17] Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision support systems*, *72*, 72-81.
[18] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.
[19] Seca, D., & Mendes-Moreira, J. (2021, March). Benchmark of encoders of nominal features for regression. In *World Conference on Information Systems and Technologies* (pp. 146-155). Cham: Springer International Publishing.
[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.
[21] Justin M. Johnson and Taghi M. Khoshgoftaar. "Encoding Techniques for High-Cardinality Features and Ensemble Learners". In: IRI. IEEE, 2021, pp. 355–361
[22] Gnat, S. (2021). Impact of categorical variables encoding on property mass valuation. Procedia Computer Science, 192, 3542-3550.
[23] Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics, 37(5), 2671-2692.
[24] Wright, M. N., & König, I. R. (2019). Splitting on categorical predictors in random forests. PeerJ, 7, e6339.
[25] Carneiro, E. M., Forster, C. H. Q., Mialaret, L. F. S., Dias, L. A. V., & da Cunha, A. M. (2022). High-cardinality categorical attributes and credit card fraud detection. Mathematics, 10(20), 3808.
[26] Slakey, A., Salas, D., & Schamroth, Y. (2019). Encoding categorical variables with conjugate bayesian models for wework lead scoring engine. arXiv preprint arXiv:1904.13001.
[27] Uyar, A., Bener, A., Ciray, H. N., & Bahceci, M. (2009, September). A frequency based encoding technique for transformation of categorical variables in mixed IVF dataset. In 2009 annual international conference of the Ieee engineering in medicine and biology society (pp. 6214-6217). IEEE.
[28] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and information systems, 34, 483-519.
[29] McGinnis, W. D., Siu, C., Andre, S., & Huang, H. (2018). Category encoders: a scikit-learn-contrib package of transformers for encoding categorical data. Journal of Open Source Software, 3(21), 501.
[30] Oracle Developers (2021, Jun 14). ML Concepts - Encoding of Categorical Attributes: OneHot vs Mean vs WoE and when to use them [Video]. датум притупа 21.07.2024, преузето са: https://www.youtube.com/watch?v=IvZfw5lRedY
[31] UCLA: Statistical Consulting Group. R Library Contrast Coding Systems for categorical variables, датум приступа 21.07.2024, преузето са: https://stats.oarc.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/
[32] GeeksforGeeks (2021, Mar 09). Categorical Encoding with CatBoost Encoder датум притупа 21.07.2024, преузето са: https://www.geeksforgeeks.org/categorical-encoding-with-catboost-encoder/

Милан Вукичевић, редовни професор, Факултет организационих наука, Универзитет у Београду
Контакт: milan.vukicevic@fon.bg.ac.rs
Области интересовања: машинско учење, наука о подацима

Алекса Милосављевић, дипл. инг. организационих наука,
Контакт: am20233066@student.fon.bg.ac.rs
Области интересовања: машинско учење, анализа и визуелизација података, пословна интелигенција и базе података