Prednosti i slabosti za One Hot Encoder :

✅ Предности на One Hot Encoding:

Нема рангирање или редослед:
Категориите се трансформираат во бинарни вектори, со што се избегнува вештачко рангирање кај номинални (неуредени) вредности.

Поддржан од повеќето ML алгоритми:
Работи добро со алгоритми кои не се базираат на редослед или растојание, како што се: логистичка регресија, decision trees, random forest и др.

Јасна репрезентација:
Секоја категорија добива сопствена колона, што ја прави интерпретацијата полесна.

❌ Слабости на One Hot Encoding:

Проклетството на димензионалноста (Curse of Dimensionality):
За променливи со многу уникатни категории (нпр. земји, пошт. кодови), бројот на колони експлодира → што може да доведе до мемориски и перформансни проблеми.

Склоност кон ретки категории:
Ретки категории добиваат сопствена колона, што може да доведе до overfitting.

Неефикасно користење на простор:
Вектори се полни со нули → sparse matrix. Иако некои алгоритми и библиотеки го поддржуваат sparse форматот, тоа не е секогаш случај.

Недостиг од информации за сличност:
Слични категории (нпр. „црвено“, „темноцрвено“, „бордо“) ќе бидат третирани како сосема различни, без никаква врска меѓу нив.

////////

Seq2Seq pozitivni negativni strani i objasni detalno

☑ Позитивни страни на Seq2Seq:

1. Може да обработува секвенци со променлива должина – идеално за преводи, резимирање и чатботи.
2. Го памети контекстот преку енкодер-декодер архитектура.
3. Флексибилен е и се комбинира со attention за подобри резултати.

✖ Негативни страни:

1. Без attention, губи контекст кај долги секвенци.
2. Бавен за тренирање и бара многу податоци.
3. Тешко е да се дебагира – „црна кутија“.

    - Наместо да се потпира само на еден контекстен вектор, attention овозможува моделот да „гледа“ кон сите енкодирани зборови.
    - Помага кај долги реченици и драматично ја подобрува точноста.
    - Se sostoi od encoder i decoder

Razlika megju boosting i bagging, koj podobro se spravuva so overfit I sum, so kakvi podatoci se spravuvaat i dali moze da kombiniraat povekje modeli

Bagging (Bootstrap Aggregating) гради повеќе модели паралелно на различни подмножества од податоците и ги комбинира нивните резултати (на пр. Random Forest), додека **Boosting** гради модели серијално, каде секој нов модел учи од грешките на претходниот (на пр. XGBoost).
Bagging подобро се справува со **overfitting** и **шумовити податоци**, бидејќи го намалува варијансот преку агрегирање.
И двата метода се **ensemble техники** и можат да комбинираат повеќе модели, иако најчесто користат decision trees.

1. Koj algoritam kreira prazni klasteri i objasni zosto i kako se slucuva toa?

K-Means алгоритмот може да креира **празни кластери**.
Ова се случува кога ниту една точка од податоците не е најблиска до некој центроид при некоја итерација, па тој центроид останува без доделени точки.
Причини може да бидат лоша иницијализација, нерамномерна распределба на податоци или избор на премногу кластери (K > реалниот број групи).

2. Dali ednonasocna Rnn i dvonasocna se razlikuvaat i kako? Moze li da ima dlaboka rnn mreza i objasni kako bi rabotela

Да, **еднонасочна RNN** чита податоци само од **лево кон десно**, додека **двонасочна RNN (BiRNN)** чита и **од лево и од десно**, што овозможува подобар контекст.
Длабока RNN постои и се гради со **повеќе RNN слоеви еден врз друг**, каде излезот од еден слој е влез за следниот, што овозможува учење на покомплексни временски зависимости.
Во длабока BiRNN, секој слој може да биде и двонасочен, со што уште повеќе се збогатува контекстот од сите страни.

prednosti i slabosti na label encoder

Предности:

- Едноставен и брз за имплементација.
- Ја претвора секоја категорија во бројка, што е погодна за алгоритми што бараат нумерички вредности.

Слабости:

- Воведува **вештачки редослед** меѓу категориите, што може да ја **збуни моделот** кај номинални податоци.

- Не е соодветен за непоредливи категории (нпр. „црвено“, „зелено“, „сино“).

Koj algoritam za klasteriranje ima problem so lokalen neshto optimum Za gini index i information neshto da se objasnat I koj e podobar zsh e podobar?

K-Means алгоритмот има проблем со локален минимум, бидејќи итеративно го оптимизира внатрешниот просек и може да остане заглавен во локален оптимум.

Gini Index и **Information Gain** се критериуми за мерење на квалитетот на поделеноста во дрвјата на одлука:

- **Gini Index** мери веројатност од погрешна класификација (пониска вредност = подобра чистота).
- **Information Gain** мери намалување на ентропијата (повеќе информација по поделба).

Information Gain често се смета за подобар бидејќи е базиран на ентропија и подобро ја мери нееднаквоста во распоредот на класи.

1. **Koj algoritam za klasteriranje generira prazni klasteri, zoshto se sluchuva toa**
   K-Means може да генерира празни кластери затоа што некој центроид може да нема најблиски точки во текот на итерација. Тоа се случува поради лоша иницијализација или ако има премногу кластери за дадените податоци.
2. **Objasni koi slichnosti i razliki gi imaat Seq2Seq i transformeri**

Сличности: И двата модели користат енкодер-декодер структура за трансформација на една секвенца во друга (на пр. превод).**Разлики:** Seq2Seq обично користи

RNN/GRU/LSTM кои обработуваат податоци последователно, додека Transformers користат attention механизам што овозможува паралелна обработка и подобро доловување на далечни зависности. Transformers се побрзи и поефикасни кај долги секвенци, додека Seq2Seq без attention има проблеми со долги влезови.

razlika pomegu aglomerativ i knn slika so clustering za dbscan pravni tekstovi da napravis model esejsko za sto sluzi FFNN kaj transformeri hyerarchical clasterint za losi performansi zosto kako shto i da dades primer

1. Разлика помеѓу agglomerative clustering и k-NN:

- **Agglomerative clustering** е алгоритам за кластерирање кој почнува со секој примерок како свој кластер и постепено ги спојува најблиските кластери, создавајќи хиерархија.
- **k-NN (k-nearest neighbors)** е класификациски алгоритам кој ги класифицира новите примери според најблиските k точки во тренирачкиот сет, не создава кластери.
- Значи, agglomerative clustering групира податоци, додека k-NN предвидува класа.

2. DBSCAN clustering со слика (опис):

DBSCAN ја групира точките во кластери базирани на густината, каде точките во густите региони формираат кластери, а ретките точки се третираат како шум (noise). На сликата се гледаат јасни групи точки (кластери) и неколку точки одвоени како шум. DBSCAN е добар за кластеризација на нерамномерни и комплексни форми.

3. Модел за правни текстови (есејски стил):

Моделот за обработка на правни текстови треба да разбира специфичен речник и структура. Често се користат NLP техники како токенизација, векторизација (Word2Vec, BERT), и класификациски модели за автоматско категоризирање или извлекување информации. Важно е да се обучи на голема количина правни документи за подобри резултати.

4. За што служи FFNN кај трансформери?

FFNN (Feed-Forward Neural Network) во трансформерите е слој што ја обработува секоја позиција од влезот независно по attention слојот. Тој ја додава нелинеарноста и ја зголемува моќта за моделирање на комплексни односи помеѓу елементите во секвенцата.

5. Зошто hierarchical clustering може да има лоши перформанси и пример:

Hierarchical clustering е пресметковно скап за големи сетови бидејќи треба да пресметува растојанија меѓу сите точки и кластери ($O(n^2)$). Лошите перформанси се забележуваат при големи податоци или шумни податоци, што доведува до непотполни или несоодветни кластери. Пример: ако имаш 10,000 правни документи, hierarchical clustering ќе биде многу бавен и тежок за интерпретација.

///////

1. **Spored Universal Approximation Theorem dovolno e eden skrien sloj vo nevronskata mreza za da se pretstavi aproksimacija na bilo koja funkcija do odredena mera na tocnost. Sepak se pretpocita koristenjeto na podlaboki nevronski mrezi so poveke skrieni sloevi. Objasni zosto,**

Иако еден скриен слој теоретски може да апроксимира било која функција, во практика тоа бара **екстремно голем број неврони**, што ја прави мрежата тешка за тренирање и неефикасна. Подлабоките мрежи со повеќе слоеви може да учат **поедноставни, хиерархиски претстави**, што ја подобрува ефикасноста, генералзацијата и конвергенцијата на моделот. Затоа, подлабоките мрежи се пофлексибилни и се полесни за тренирање во реални апликации.

2. **Kako DBSCAN se spravuva so outliers a kako K-means? Navedi situacii za primer koga DBSCAN bi dovel do podobri klasteri od K-means**

DBSCAN ги третира outliers како *шум* (noise) и не ги вклучува во кластери, што го прави отпорен на нив.
K-means ги вклучува сите точки во некој кластер, па outliers можат да го поместат центроидот и да ги нарушат кластери.

Пример: Кога имаш податоци со кластери различна форма и густина, и со шум, како групи од точки во форма на круг и ленти, DBSCAN ќе ги идентификува правилно кластери и ќе ги исклучи шумот, додека K-means ќе создаде „кружни“ кластери и ќе ги меша шумот со вистинските кластери.

Зошто е потребна нормализација на податоците неопходна во хиерархиските алгоритми за кластерирање. Наведете пример како не-нормализираните податоци можат да ги деградираат перформансите на хиерархискиот алгоритам за кластерирање

Нормализацијата е потребна бидејќи хиерархиските алгоритми користат мерки на растојание (на пр. Евклидово), кои се чувствителни на скалата на податоците. Ако една карактеристика има поголем опсег, ќе доминира во пресметката на растојанието и ќе влијае непропорционално врз кластерирањето.

Пример: Ако имаш податоци со две карактеристики — една во опсег 0-1 (на пр. висина во метри) и друга во 0-1000 (на пр. годишен приход),

ненормализираните податоци ќе направат алгоритмот да ги групира според приходот, игнорирајќи ја висината, што може да доведе до непрактични кластери.

Која е улогата на feed forward neural networks во transformer архитектурата

Во Transformer архитектурата, FFNN слојот обработува секоја позиција од влезната секвенца **независно**, по вниманието (attention) слојот. Тој додава **нелинеарност** и ја зголемува способноста на моделот да учи комплексни трансформации и релашни помеѓу елементите во секвенцата. FFNN е исто така одговорен за трансформација на карактеристиките во повисоко-димензионален простор, што ја подобрува изразноста на моделот.

Zosto vo praksa se koristi deep network namesto shallow network, da se objasne.,

Длабоките мрежи можат да научат **покомплексни и хиерархиски претстави** на податоците, што овозможува подобро справување со сложени задачи. Тие се поефикасни во моделирање на високо-нивоу функции со помалку параметри отколку плитките мрежи со слична моќност. Исто така, длабоките мрежи подобро генералзираат и постигнуваат повисока точност во многу апликации.

1. **So kakvi predizvici moze da se sooci k-means algoritamot i da se dade realen primer kade takov predizvik bi napravil problem**

K-means се соочува со предизвици како:

- Потреба од претходно дефиниран број на кластери (K).
- Претпоставка дека кластери се кружни и со слична големина.

- Чувствителност на outliers и шум.
- Можност за заглавување во локални оптимуми.

Реален пример: Ако имаш податоци со несиметрични, различно густински кластери, како групи од точки во форма на прстени или со различна густина, K-means може да ги сечи или меша кластери неправилно, што го прави непогоден за задачи како групирање на клиенти со комплексно однесување.

\/// kvalifikaciski ?

Za prv kol kval Matrica so tp fp tn fn Recall precsion da se izracuna Ima banka 90k normalnj transk 70k nekakvi 10k maliciosni Modelot predvediva 90%accuracy - modelot dobro predviduva za 3te klasi -modelot dobro gi predibudva pozitivnire

- modelot e dobro isteniran,

- Ako imame drvo so 10 koloni i 300 redici kolku e najdobar i najlos broj na listovi so treba Sistem na finki koj treba da predvidi dali studentot ke prodolzi na studii ili ne

- spored ocenite vo iknow,
- spored logovite kolku e aktiven na kurses,

-uste ponudeni Neso za supervised i unsupervised learning na drag and drop Slikata so bias i variance od baza Grafik so pirsonova korelacija i dali e slaba silna i uste 2 ponudeni

1. Razlika od rnn i lstm (4 ponudeni ne gi pamtam),
2. Sto e tocno za feed forward nn? Deka podatocite se dvizat vo edna nasoka bez loop (ova zaokruziv),
3. Slika i koe clustering e tocno,
4. Imas da napravis model za bolnica kaj so imas 1000 statii i nekoi medicinski zborovi so e najdobro :a) sopstven bert b) fine tune clinical bert,
5. Sto e tocno za lagovi?tocen odg deka mora da zimaat vrednosti od minatoto,

6. Nesto lstm so keras sto znaci naredbata return sequences ???