ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ

Дәріс 6

Деректерді тазалау

Оспан Ә.Ғ., PhD, доцент-зерттеуші

ҚазҰУ, Жасанды интеллект және Big Data кафедрасы

2026-2027 оқу жылы

Дәрістің мақсаты мен мазмұны

Мақсат және оқу нәтижелері

- Схема, тип және мән ақауларын ажырату.
- Жетіспейтін мәндер мен қайталанатын жазбаларды өңдеуді негіздеп таңдау.
- Тазалау нәтижесін санауыштар және кейінгі шарттар арқылы тексеру.

Дәріс мазмұны

1. **Деректер жиынының паспорты**
2. **Жетіспейтін мәндер мен типтер**
3. **Мәтіндік жолдар мен сүзгілер**
4. **Қайталанатын жазбалар**
5. **Нәтижені бақылау**

Тазалау: ережеден тексеруге

1

Ережені анықтау

Рұқсат етілген типтерді, мәндерді және өрістер арасындағы байланыстарды атау.

2

Бұзушылықтарды табу

Маскалар құрып, әсер етілетін жазбаларды санау.

3

Әрекетті таңдау

Түзету, белгісіз күйде қалдыру немесе карантинге жіберу.

4

Нәтижені тексеру

Санауыштарды, үлестірімдерді және бастапқы нұсқаның сақталуын салыстыру.

- Ережелер белгі мен есептің мағынасына тәуелді.
- Карантин күмәнді жолдарды себептерімен бірге сақтайды.

Қайта орындалатын тазалау = бастапқы нұсқа + ережелер + код + журнал.

Сапаны тексерудің үш деңгейі

Құрылым

- Бағандар саны мен атаулары.
- Файл бөлгіші, кодтауы және тақырыбы.

Бейнелеу

- Сан мәтін түрінде немесе бос орындармен жазылған.
- Шартты маркер белгісіз мәнді жасырады.

Мағына

- Сан пәндік ауқымнан шығады.
- Бір жазбаның өрістері бір-біріне қайшы келеді.

«Нашар жолдардың» бір ортақ санауышы әртүрлі себептерді жасырады; әр деңгейге жеке тексеру қажет.

football.csv паспорты және шектеулер

Өріс немесе қасиет	Не белгілі	Салдары
Көлем	12897 жол, 42 баған	Оқығаннан кейін пішінін тексереміз
Жетіспейтін мәндер	Бастапқы файлда жоқ	«Табылған» NaN мәндерін ойдан қоспаймыз
Unnamed: 0	Жолдардың ескі индексі	Аналитикалық белгі емес
Name	Бірегейлікке кепілдік жоқ	Аттары сәйкес жазбаларды жоймаймыз
Penalties	Пенальти дағдысының рейтингі	Гол саны емес
Wage / Value	Өлшем бірліктері мен кезең белгісіз	Жылдық еуро деп атамаймыз

Тарихи кескін қазіргі шындықты сипаттамайды.

Жүктеу және бастапқы аудит

Python мысалы

```
from pathlib import Path
import pandas as pd

raw = pd.read_csv(Path('data') / 'football.csv')
assert raw.shape == (12897, 42)
print(raw.isna().sum().sum())
raw.info()
```

Нәтиже: Күтілетін пішін: (12897, 42); жетіспейтін мәндер саны: 0.

Тексереміз

- Кесте пішінін және жетіспейтін мәндер санын.
- Автоматты танудан кейінгі типтерді.

Түсіндіреміз

- NaN санының нөл болуы осы файлға қатысты.
- Дерек типі мәннің мағынасын растамайды.

Тану қатесі және мән қатесі

Тану

- «қате» санға айналмайды.
- errors="coerce" қызметтік жетіспейтін мән жасайды.
- Бастапқы жетіспейтін мән бөлек есептеледі.

Пәндік ауқым

- «-4» жолы -4 санына айналады.
- Жас ережесі мағына бұзылуын анықтайды.
- Шектер есеппен негізделуі тиіс.

Типті түрлендіргеннен кейін мәндерді жеке тексеру және әр бұзушылық себебіне түсінікті санауыш қажет.

Қателердің екі себебін бөлек санаймыз

Python мысалы

```
demo = pd.DataFrame({
    'age_text': ['21', ' 25 ', 'қате', '-4']
})
demo['age'] = pd.to_numeric(
    demo['age_text'], errors='coerce'
)
parse = demo['age_text'].notna() & demo['age'].isna()
domain = demo['age'].notna() & ~demo['age'].between(0, 120)
print(parse.sum(), domain.sum())
```

Нәтиже: 1 1

parse маскасы

- Бастапқы мәні белгілі сандық емес мәтін.
- Алғашқы жетіспейтін мәнмен араластырмаймыз.

domain маскасы

- Сан танылды, бірақ шектерді бұзады.
- Мұндағы 0-120 — тек оқу ережесі.

Белгісіз мәнмен не істеу керек

Шешім	Пайдасы	Шектеуі
NA күйінде қалдыру	Белгісіздік сақталады	Барлық есептеулер анықталмайды
Карантин	Мәселе себебі сақталады	Негізгі жиын кішірейеді
Есептеуден шығару	Қарапайым жергілікті талдау	Құрам өзгеруі мүмкін
Медианамен толтыру	Жол саны сақталады	Шашырау жасанды тарылады
Дереккөзден сұрау	Фактіні қалпына келтіруге болады	Бірден қолжетімді бола бермейді

Белгісіз мән нөлге тең емес.

Мәтінді мағынасын жоғалтпай қалыпқа келтіру

Әдетте түсінікті әрекеттер

- strip сыртқы бос орындарды жояды.
- casefold іздеу кілтін жасайды.
- Бастапқы жазылуы бөлек сақталады.

Пәндік тексеру қажет

- Ішкі бос орындарды жою.
- Қысқа ішкі жолдар бойынша алмастыру.
- Белгісіз санаттарды біріктіру.

«Бастапқы мән - қалыпқа келтірілген кілт» сәйкестіктерін сақтап, ереже қандай санаттарды біріктіргенін тексеріңіз.

Іздеу кілті және тура мәтінді іздеу

Python мысалы

```
clubs = pd.Series([' Alpha ', 'alpha', None],
                  dtype='string')
key = clubs.str.strip().str.casefold()
has_al = key.str.contains('al', regex=False, na=False)
known_without = key.notna() & ~has_al
print(has_al.tolist())
print(known_without.tolist())
```

Нәтиже: has_al: [True, True, False]; known_without: [False, False, False]

Тура мәтінді іздеу

- regex=False мәтінді үлгі ретінде түсіндірмейді.
- na=False масканың мәндерін анық етеді.

Үш топ

- Сәйкестік, белгілі сәйкессіздік, NA.
- Масканы жай терістеу NA мәндерін де қамтиды.

Сүзгілеу: .loc және query

Python мысалы

```
work = raw.drop(columns=['Unnamed: 0']).copy()
limit = 30
mask = (work['Age'] < limit) & (work['Stamina'] >= 70)
selected = work.loc[mask].copy()
same = work.query('Age < @limit and Stamina >= 70')
assert selected.index.equals(same.index)
print(len(selected))
```

Айқын маска

- Әр салыстыру жақшаға алынады.
- Скалярлық and орнына & қолданылады.

Тексеру

- Таңдалған индекстерді салыстырамыз.
- copy тәуелсіз жұмыс кестесін белгілейді.

Жол санының бірдейлігі іріктеудің бірдейлігіне кепілдік бермейді.

Санат құру: алдымен мәндерді тексеру

age_group ережесі

- 23-тен кіші: «23-ке дейін».
- 23-35 қоса алғанда: «23-35».
- 35-тен үлкен: «35-тен жоғары»; NA бөлек.

Операциялар реті

- age мәнін тану және тексеру.
- Қате мәндерге қатысты шешім қабылдау.
- Функцияны 22, 23, 35, 36, NA үшін тексеру.

Санаттау жастың қате бастапқы сандарын жасырмауы тиіс.

Жолдардың және нысандардың қайталануы

Жазбалардың сәйкес келуі

- duplicated таңдалған өрістерді салыстырады.
- Қызметтік индекс қайталануды жасыруы мүмкін.
- Жою үшін шығу тегі түсінікті болуы қажет.

Нақты нысанның сәйкес келуі

- Name сенімді ID емес.
- Жазба нұсқалары ерекшеленуі мүмкін.
- Кілт және нұсқа таңдау ережесі қажет.

Кіріс = сақталған жолдар + карантин + расталған қайталанатын жазбалар.

Тазалау қателері және бақылау тәсілдері

Қате	Салдары	Бақылау
Бастапқы нұсқаның үстіне жазу	Шешімді қайта қарау мүмкін емес	Жеке жұмыс көшірмесі
Барлық NA мәнін нөлмен алмастыру	Мағына мен орташа мәндер бұрмаланады	Әр бағанға арналған ереже
Аттары сәйкес жазбаларды жою	Әртүрлі нысандар жоғалады	Сенімді кілт
Меншіктеулер тізбегі	Уақытша нысанның өзгеруі	Бір айқын .loc
Барлық сирек мәндерді жою	Құрамның ығысуы	Топтарды дейін және кейін тексеру
Ескертулерді жасыру	API мәселесі байқалмайды	Оқып, себебін жою

Дәрістің негізгі нәтижелері

1. Тазалау айқын пәндік ережеге сүйенеді.
2. Санды тану мен ауқымды тексеру — әртүрлі кезеңдер.
3. Жетіспейтін мән нөлге тең емес; шешім есепке тәуелді.
4. Ат нысанның сенімді идентификаторын алмастырмайды.
5. Журнал мен жолдар теңгерімі өзгерістерді тексеруге мүмкіндік береді.

Бақылау сұрақтары және өзіндік жұмыс

1. Тану қатесі рұқсат етілмеген саннан қалай ерекшеленеді?
2. Жетіспейтін мәндер санының нөл болуы неге бүкіл жиынның сапасын дәлелдемейді?
3. na=False берілген contains маскасын терістегенде не болады?
4. Name сәйкестігін қашан нысанның қайталануы деуге болады?
5. Жол санының өзгерісін қандай санауыштар түсіндіруі тиіс?

Өзіндік жұмыс

Алты оқу жолында NA мәндерін жою мен медианамен толтыруды салыстырыңыз. Әр шешім үшін көлемді, орташа мәнді, шашырауды және шектеулерді көрсетіңіз.

Дереккөздер және құжаттама

Құжаттама

1. pandas: жетіспейтін мәндер

https://pandas.pydata.org/docs/user_guide/missing_data.html

2. pandas: мәтіндік деректер

https://pandas.pydata.org/docs/user_guide/text.html

3. pandas: таңдау және меншіктеу

https://pandas.pydata.org/docs/user_guide/indexing.html

Дәрістің негізі

- Берілген 6-дәрістің презентациясы
- Берілген 6-зертханалық жұмыс
- Берілген 6-зертханалық жұмыстың жауаптары