Python for AI

Дәуренов ӘД

Дәріс 8. Pandas: кестелік деректерді өңдеу және талдау

1. Дәрістің мақсаты

Кестелік деректердің құрылымын, Pandas кітапханасының жұмыс принциптерін, деректер сапасын бағалау, алдын ала өңдеу және талдау ұғымдарын түсіндіру.
Дәріс соңында білім алушылар:

- Деректердің негізгі түрлерін ажыратады.
- Кестелік деректердің құрылымын түсінеді.
- Series және DataFrame ұғымдарын сипаттайды.
- Деректерді тазалау, сүзу, сұрыптау және топтау принциптерін түсіндіреді.
- Pandas кітапханасының Data Science және жасанды интеллекттегі рөлін біледі.

2. Деректер және олардың түрлері

Деректер (Data) — белгілі бір объект, процесс немесе құбылыс туралы тіркелген ақпарат.

Автомобиль жүйелерінде деректер электрондық басқару блоктарынан, датчиктерден, диагностикалық жүйелерден және телеметриялық құрылғылардан алынуы мүмкін.

Деректер ұйымдастырылу тәсілі бойынша бірнеше түрге бөлінеді.

Деректердің негізгі түрлері

Құрылымдалған деректер

Structured Data

Алдын ала анықталған құрылымы бар ақпарат: кестелер, бағандар, жазбалар.

ID	Temperature	RPM
1	85	2500
2	92	3100

Мысал: CSV файлындағы көлік өлшемдері

Жартылай құрылымдалған деректер

Semi-structured Data

Ұйымдастыру ережелері бар, бірақ тұрақты кесте түрінде сақталмайтын ақпарат.

```
{
  "ID": 1,
  "Temperature": 85,
  "RPM": 2500
}
```

Мысал: JSON форматындағы телеметриялық хабарлама

Құрылымдалмаған деректер

Unstructured Data

Алдын ала белгіленген кестелік құрылымы жоқ ақпарат.



Мысал: жол камерасының бейнежазбасы, сурет, еркін мәтін

Pandas негізінен құрылымдалған және кесте түріне келтірілген деректерді талдау үшін қолданылады.

Кестелік деректер ұғымы

Кестелік деректер (Tabular Data) — жолдар мен бағандар түрінде ұйымдастырылған ақпарат.
Мысалы, электрмобиль батареясының мониторинг деректері:

ID	Temperature	Voltage	SOC
1	30	380	90
2	35	375	80
3	42	370	70
4	48	365	60

Кестенің негізгі элементтері:
- Жол (Row) — бір объект немесе бір өлшеу туралы жазба.
- Баған (Column) — белгілі бір параметр немесе белгі.
- Жазба (Record) — бір жолдағы байланысты мәндер жиынтығы.
- Белгі (Feature) — талдау немесе модельдеу кезінде пайдаланылатын сипаттама.
- Индекс (Index) — жолды белгілеуге арналған идентификатор.

4. Деректерді сақтау форматтары

Формат	Ерекшелігі	Қолданылуы
CSV	Қарапайым мәтіндік кесте	Өлшеулерді алмасу, экспорт
Excel	Электрондық кесте, форматтау және формулалар	Есептер, қолмен талдау
JSON	Иерархиялық құрылымдағы деректер	API, IoT хабарламалары
Parquet	Бағандық сақтауға негізделген формат	Үлкен деректер жиындарын тиімді сақтау
SQL дерекқоры	Кестелер мен олардың байланыстарын ұйымдастыру	Ақпараттық жүйелер, деректерді басқару

Форматты таңдау деректер көлеміне, құрылымына, өңдеу жылдамдығына және жүйелер арасындағы алмасу талаптарына байланысты.

5. Pandas кітапханасы дегеніміз не?

Pandas — Python тілінде құрылымдалған деректерді өңдеу және талдау үшін қолданылатын ашық бастапқы кодты кітапхана.

Pandas кітапханасының негізгі мүмкіндіктері:

1. Кестелік деректерді ұйымдастыру.
2. Әртүрлі файл форматтарынан деректерді оқу.
3. Деректерді тазалау және түрлендіру.
4. Жолдар мен бағандарды таңдау.
5. Сүзу, сұрыптау және топтау.
6. Статистикалық талдау.
7. Бірнеше кестені біріктіру.
8. Уақыттық қатарлармен жұмыс істеу.

Pandas көптеген Data Science жобаларында деректерді дайындау кезеңінде қолданылады.

6. Pandas кітапханасының негізгі құрылымдары

Series — индекстері бар бірөлшемді деректер құрылымы.

DataFrame — жолдар мен бағандардан тұратын екіөлшемді кестелік деректер құрылымы.

Pandas: Series құрылымы

Series — индекстері бар бірөлшемді деректер құрылымы.

Мысал: бір датчиктің температура өлшеулері



Индекс	Температура, °C
0	30
1	35
2	40
3	45

Индекс
Элементтің белгісі

Мән
Температура өлшемі

Series бір бағанды немесе бірөлшемді көрсеткіштер жиынтығын ұсынуға ыңғайлы.

DataFrame ішінде әр бағанның деректер типі әртүрлі болуы мүмкін.

Мысалы:

- ID — бүтін сан.
- Temperature — нақты сан.
- Status — мәтін.
- Timestamp — уақыт белгісі.

D

ataFrame — Pandas кітапханасының негізгі жұмыс құрылымы.

7. Индекстер және деректер типтері

Pandas кестесіндегі әр жолдың индексі болады.

Индекс деректерді таңдау, сәйкестендіру және біріктіру кезінде маңызды рөл атқарады.

Мысалы, кестеде 1000 өлшеу болса, олардың әрқайсысына индекс арқылы қол жеткізуге болады.

Алайда Pandas индексі мен автомобильдің ID бағаны міндетті түрде бірдей емес.
Индекс – кесте ішіндегі белгілер жүйесі болса,
ID – нақты өлшеудің немесе объектінің идентификаторы болуы мүмкін.

Деректер типтері

Тип	Мысал	Қолданылуы
Integer	1200	RPM, ID
Float	36.5	Температура, кернеу
String	NORMAL	Жүйе күйі
Boolean	True / False	Шарттың орындалуы
Datetime	2026-10-23 10:30	Өлшеу уақыты
Category	LOW, NORMAL, HIGH	Санаттық белгілер

Деректер типтерін дұрыс анықтау есептеулердің сенімді орындалуы үшін қажет.

8. Деректер сапасы (Data Quality)

Нақты инженерлік жүйелерде алынған деректер әрдайым дұрыс және толық бола бермейді.

Мысалы, датчиктің байланысы үзілуі, өлшеу қатесі пайда болуы немесе бір жазба бірнеше рет тіркелуі мүмкін.

Деректер сапасының негізгі мәселелері:

Мәселе	Сипаттамасы	Мысал
Missing Values	Жетіспейтін мәндер	Температура тіркелмеген
Duplicates	Қайталанған жазбалар	Бір өлшеу екі рет сақталған
Outliers	Басқа мәндерден қатты ерекшеленетін өлшеулер	35, 36, 37, 150°C
Incorrect Data Types	Дұрыс емес деректер типі	"45" мәтін ретінде сақталған
Inconsistent Units	Өлшем бірліктерінің сәйкес келмеуі	°C және °F араласуы
Timestamp Errors	Уақыт белгілеріндегі қателер	Өлшеулер ретінің бұзылуы

Маңызды: шеткі мән міндетті түрде қате емес. Мысалы, өте жоғары температура нақты қызып кету жағдайын көрсетуі мүмкін. Сондықтан деректерді тазалау кезінде физикалық мағынасы мен өлшеу жүйесінің ерекшеліктері ескеріледі.

Деректерді алдын ала өңдеу
(Data Preprocessing)

Бастапқы деректерді талдауға немесе машиналық оқытуға жарамды күйге келтіру процесі.

Деректерді алдын ала өңдеу кезеңдері



Деректерді жинау

Датчиктер, файлдар, ақпараттық жүйелер

→



Деректерді тексеру

Құрылымын, типтерін және толықтығын бағалау

→



Деректерді тазалау

Қателерді, қайталануларды және жетіспейтін мәндерді өңдеу

→



Деректерді түрлендіру

Өлшем бірліктерін, форматтарды және белгілерді сәйкестендіру

→



Талдауға дайындау

Қажетті көрсеткіштерді таңдау және есептеу

Деректерді алдын ала өңдеу машиналық оқыту нәтижесінің сапасына айтарлықтай әсер етуі мүмкін.

Сапасыз деректер дұрыс емес қорытындыларға және сенімсіз болжамдарға әкелуі ықтимал.

Деректерді сүзу, сұрыптау және топтау

Бұл үш операция деректерді талдауда жиі қолданылады, бірақ олардың мақсаттары әртүрлі.

Сүзу (Filtering) — белгілі бір шартқа сәйкес келетін жазбаларды таңдау.
Мысалы, температурасы 45°C-тан жоғары өлшеулерді ғана көрсету.

Сұрыптау (Sorting) — деректерді белгілі бір ретпен орналастыру.
Мысалы, температураларды кішіден үлкенге қарай реттеу.

Топтау (Grouping) — ортақ белгісі бар жазбаларды біріктіріп, олардың көрсеткіштерін есептеу.

Мысалы, автомобильдердің әр түрі бойынша орташа энергия тұтынуын анықтау.

Сипаттамалық статистика (Descriptive Statistics)

Сипаттамалық статистика — деректер жиынының негізгі қасиеттерін сандық көрсеткіштер арқылы сипаттау әдістері.
Негізгі көрсеткіштер:

- Өлшеулер саны.
- Орташа мән.
- Минималды және максималды мән.
- Медиана.
- Стандартты ауытқу.
- Квартильдер.

Мысалы, бір айдағы автомобильдердің отын шығынын талдағанда тек орташа шығынды ғана емес, оның өзгеру диапазонын да қарастыру қажет.

Статистикалық талдау деректердің жалпы жағдайын түсінуге көмектеседі, бірақ оның өзі себеп-салдарлық байланысты дәлелдемейді.

Кестелерді біріктіру

Кейбір инженерлік жүйелерде қажетті ақпарат бірнеше бөлек кестеде сақталады.

Автомобильдер туралы ақпарат

VehicleID	Model
101	EV-A
102	EV-B

Өлшеулер

VehicleID	Temperature
101	35
102	42

VehicleID арқылы біріктіру

Нәтиже

VehicleID	Model	Temperature
101	EV-A	35
102	EV-B	42

Pandas кітапханасында мұндай операциялар деректерді біріктіру әдістері арқылы орындалады.

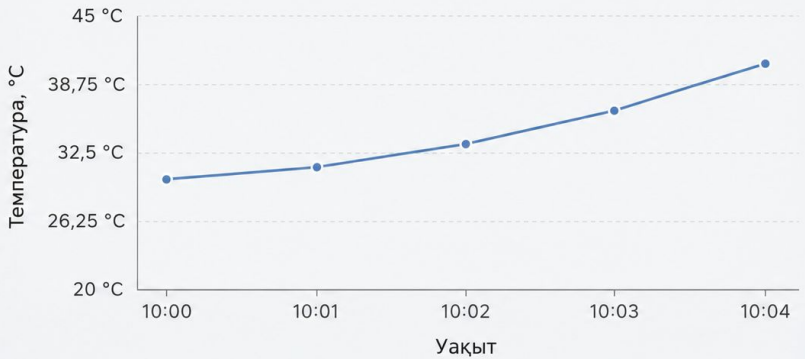
Қолданылуы: автопарк мониторингі, өндірістік ақпараттық жүйелер және диагностикалық деректерді талдау.

Уақыттық қатар (Time Series) — уақыт бойынша реттелген өлшеулер жиынтығы.

Уақыттық қатарлар (Time Series)

Батарея температурасының уақыт бойынша өзгеруі

Оқу мақсатындағы шартты телеметриялық деректер



Уақыттық қатарлар көмегімен:

- Өзгеру бағытын анықтау
- Орташа мәндерді есептеу
- Күрт өзгерістерді байқау
- Болашақ мәндерді болжау

Автомобиль телеметриясында уақыт белгілерінің дұрыстығы ерекше маңызды, себебі әртүрлі датчиктердің өлшеулерін уақыт бойынша сәйкестендіру қажет.

Pandas кітапханасының автомобиль инженериясында қолданылуы

Қолданылу бағыты	Талдау мысалы
Электрмобиль батареясы	Температура, кернеу, SOC тарихын талдау
Қозғалтқыш диагностикасы	RPM және температура деректерін салыстыру
Автопарк мониторингі	Әртүрлі көліктердің пайдалану көрсеткіштерін талдау
Энергия тиімділігі	Энергия тұтыну деректерін салыстыру
Техникалық қызмет көрсету	Диагностикалық жазбалар мен ақаулар тарихын талдау
IoT және телеметрия	Уақыт белгілері бар өлшеулерді өңдеу

Pandas көбінесе талдауға арналған компьютерлік ортада пайдаланылады. Ол автомобильдің нақты уақыттағы қауіпсіздік функцияларын орындайтын басқару контроллерінің орнына қолданылмайды.

Деректерден машиналық оқыту моделіне дейін



Деректерді жинау
Көлік датчиктері

Pandas арқылы өңдеу
Тазалау • біріктіру • сүзу • талдау

Белгілерді дайындау
Қажетті параметрлерді таңдау

Машиналық оқыту
Модельді оқыту және бағалау

Нәтижелерді пайдалану
Болжау • диагностика • талдау

Мысал: батарея температурасының болашақ өзгерісін болжау.

Pandas модельге арналған деректерді дайындауда маңызды рөл атқарады.

NumPy және Pandas кітапханаларының өзара байланысы

NumPy	Pandas
Сандық массивтерге бағытталған	Кестелік деректерге бағытталған
Математикалық операцияларды тиімді орындайды	Деректерді ұйымдастыру мен талдауды жеңілдетеді
Векторизация және broadcasting	Индекстеу, топтау, біріктіру
Матрицалық есептеулер	CSV, Excel және басқа дереккөздермен жұмыс
Сандық модельдер үшін негіз	Деректерді алдын ала дайындау үшін ыңғайлы

Қорытынды

Pandas — кестелік деректерді өңдеу мен талдаудың маңызды құралы.

Оның көмегімен инженерлер мен деректер талдаушылары әртүрлі ақпарат көздерінен алынған өлшеулерді біріктіріп, сапасын тексеріп, статистикалық талдау жүргізе алады.

Pandas кітапханасының негізгі маңызы — бастапқы деректерді түсінікті, құрылымдалған және әрі қарай талдауға жарамды ақпаратқа айналдыру.

Бақылау сұрақтары

1. Құрылымдалған және құрылымдалмаған деректердің айырмашылығы қандай?
2. Кестелік деректер дегеніміз не?
3. CSV, Excel және Parquet форматтарының ерекшеліктері қандай?
4. Pandas кітапханасының негізгі мақсаты қандай?
5. Series және DataFrame құрылымдарының айырмашылығы неде?
6. Деректер сапасы дегеніміз не?
7. Жетіспейтін мәндер мен шеткі мәндердің айырмашылығы қандай?
8. Data Preprocessing қандай кезеңдерден тұрады?
9. Filtering, Sorting және Grouping ұғымдары нені білдіреді?
10. Деректерді біріктіру не үшін қажет?
11. Уақыттық қатар дегеніміз не?
12. Pandas машиналық оқытуда қандай мақсатта қолданылады?