Python for AI

Дәуренов ӘД

Дәріс 9. Деректерді талдау және машиналық оқыту үшін дайындау және тазалау

1. Дәрістің мақсаты

Білім алушыларға

- деректерді алдын ала дайындау және
- тазалау ұғымдарын,
- олардың негізгі кезеңдерін,
- деректер сапасына әсер ететін факторларды
- және машиналық оқыту жүйелеріндегі маңызын түсіндіру.

2. Деректерді дайындау дегеніміз не?

Деректерді дайындау (Data Preparation) — бастапқы деректерді талдау, модельдеу немесе машиналық оқыту үшін жарамды күйге келтіру процесі.

Нақты инженерлік жүйелерден алынатын деректер әрдайым толық, дұрыс және біркелкі болмайды.

Мысалы, электрмобильдің батареяны басқару жүйесі (BMS) температураны, кернеуді және токты үздіксіз өлшейді. Алайда кейбір өлшеулер жоғалуы, қате тіркелуі немесе бірнеше рет сақталуы мүмкін.

Осындай деректерді алдын ала өңдемей талдау қате қорытындыларға әкелуі ықтимал.

Деректерді дайындаудың жалпы процесі



Бастапқы деректер

Датчиктер, CSV файлдары, дерекқорлар

Деректер сапасын тексеру

Толықтығы, дұрыстығы, бірізділігі

Деректерді тазалау

Қателер, бос мәндер, қайталанулар

Деректерді түрлендіру

Типтер, өлшем бірліктері, масштабтар

Талдауға дайын деректер

Статистика және машиналық оқыту

Деректерді дайындау деректердің физикалық мағынасын, өлшеу жүйесінің ерекшеліктерін және талдаудың мақсатын ескеруді талап етеді.

3. Деректер сапасы (Data Quality)

Өлшем	Сипаттамасы	Мысал
Дәлдік (Accuracy)	Мәннің нақты жағдайға сәйкестігі	Датчик температураны дұрыс өлшейді
Толықтық (Completeness)	Қажетті деректердің болуы	Барлық уақыттағы өлшеулер сақталған
Бірізділік (Consistency)	Деректердің өзара қайшы келмеуі	Температура бір өлшем бірлігінде берілген
Жарамдылық (Validity)	Мәндердің белгіленген ережелерге сәйкестігі	SOC 0–100% аралығында
Бірегейлік (Uniqueness)	Артық қайталанулардың болмауы	Бір оқиға негізсіз екі рет сақталмаған
Уақтылылық (Timeliness)	Деректердің қажетті уақытта қолжетімді болуы	Мониторинг ақпараты уақытында жаңартылады

Деректер сапасы — ақпараттың белгілі бір мақсат үшін қаншалықты дұрыс, толық, сенімді және қолдануға жарамды екенін сипаттайтын қасиеттер жиынтығы.

Деректер сапасының негізгі өлшемдері кестеде!

Деректердегі қателердің негізгі түрлері

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.

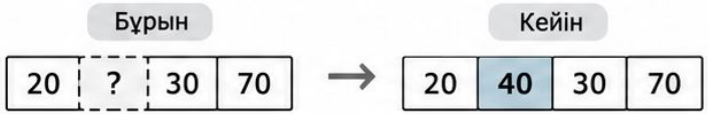
Жетіспейтін деректерді өңдеу әдістері

Жолды алып тастау



Жетіспейтін мәні бар жазбаны жою

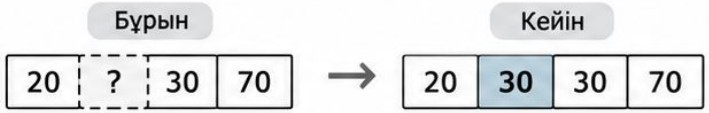
Орташа мәнмен толтыру



Бос мәнді орташа көрсеткішпен алмастыру

(20 + 30 + 70) / 3 = **40**

Медианамен толтыру



Бос мәнді медианамен алмастыру

Медиана = 30

Интерполяция



Көршілес өлшеулер арқылы бағалау

Тең уақыт аралықтары

Арнайы белгімен сақтау



Мәннің жоқ екенін жеке белгімен көрсету

Маңызды: Жетіспейтін мәнді толтыру әрқашан дұрыс шешім емес. Батареяның қызу сәтіндегі жоғалған өлшеуді орташа температурамен алмастыру қауіпті жағдайды жасыруы мүмкін.

Қайталанатын жазбаларды өңдеу

Қайталанатын жазба — деректер жиынында бірдей жазбаның бірнеше рет кездесуі.

Бұрын

ID	Temperature	Voltage
1	30	380
2	35	375
3	40	370
× 3	40	370
× 3	40	370

Бір өлшеу үш рет сақталған

Артық көшірме екені расталса

Кейін

ID	Temperature	Voltage
1	30	380
2	35	375
3	40	370

Бір жазба қалады, екі артық көшірме жойылады

Бірдей мәндер әрдайым қате көшірме емес



Температура датчигі

- 10:00:00 — 40 °C
- 10:00:01 — 40 °C
- 10:00:02 — 40 °C

Әртүрлі уақытта алынған өлшеулер сақталады.

Қайталануды анықтағанда өлшеу уақыты, құрылғы идентификаторы және жазбаның мағынасы ескеріледі.

Диапазоннан тыс мәндер (Out-of-Range Values)

Көптеген физикалық параметрлердің белгілі бір жарамды диапазоны болады.

Мысалы : SOC 0–100%
SOC = −10% немесе SOC = 130% сияқты мәндер әдеттегі заряд деңгейі шкаласына сәйкес келмейді.
Ал температураның жоғары болуы міндетті түрде өлшеу қатесі емес. Ол нақты қызып кетуді білдіруі мүмкін.

Сондықтан екі ұғымды ажырату қажет:

- Физикалық жарамсыз мән — жүйе анықтамасына сәйкес мүмкін емес немесе дұрыс емес мән.
- Қалыпты жұмыс диапазонынан тыс мән — физикалық тұрғыдан мүмкін, бірақ жүйенің қалыпты режиміне сәйкес келмейтін өлшеу.

Екінші жағдайда деректі автоматты түрде жоюға болмайды.

Шеткі мәндер (Outliers)

Шеткі мән (Outlier) — деректер жиынындағы басқа мәндерден айтарлықтай ерекшеленетін өлшеу.

Мысалы: 30, 31, 32, 33, 34, 85
85°C мәні қалған өлшеулерден едәуір жоғары.

Шеткі мәннің пайда болу себептері:

1. Датчиктің өлшеу қатесі.
2. Байланыс немесе деректерді түрлендіру қатесі.
3. Нақты физикалық процестің күрт өзгеруі.
4. Жүйедегі ақаулық немесе төтенше жағдай.

Шеткі мәндерді анықтау әдістеріне стандартты ауытқу, квартильаралық аралық (IQR) және инженерлік шекті мәндер жатады.
Шеткі мәнді анықтау — оны міндетті түрде жою деген сөз емес.

Деректерді түрлендіру (Data Transformation)

Деректерді түрлендіру — бастапқы мәндерді талдау үшін қажетті құрылымға, типке немесе масштабқа келтіру.

Негізгі түрлендірулер:

- Деректер типін өзгерту.
- Өлшем бірлігін түрлендіру.
- Нормализация.
- Стандарттау.
- Санаттық мәндерді кодтау.
- Жаңа есептелетін белгілер құру.

Деректерді түрлендіру (Data Transformation)

Өлшем бірліктерін түрлендіру
(мысалы, температураны Фаренгейттен Цельсийге түрлендіру)

Нормализация (Normalization) – деректерді белгілі бір масштабқа келтіреді.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

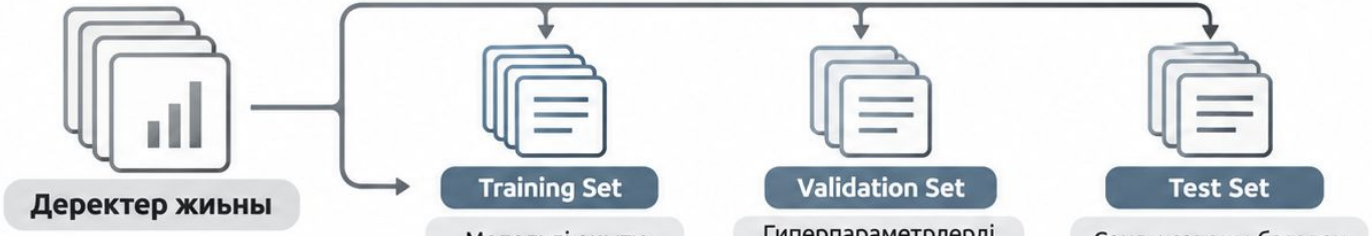
Стандарттау (Standardization) — деректерді орташа мән мен стандартты ауытқу арқылы түрлендіреді:

$$z = \frac{x - \mu}{\sigma}$$

Мұнда:

- x — бастапқы мән.
- μ — орташа мән.
- σ — стандартты ауытқу.

Деректерді бөлу және ақпараттың ағып кетуі



Бөлу мысалы:
70% / 15% / 15%.
Қатынас міндетті емес.

Өңдеудің дұрыс реті



Data Leakage — деректердің ағып кетуі

Бағалау жиынындағы немесе болашақтағы ақпараттың оқытуға еніп кетуі.



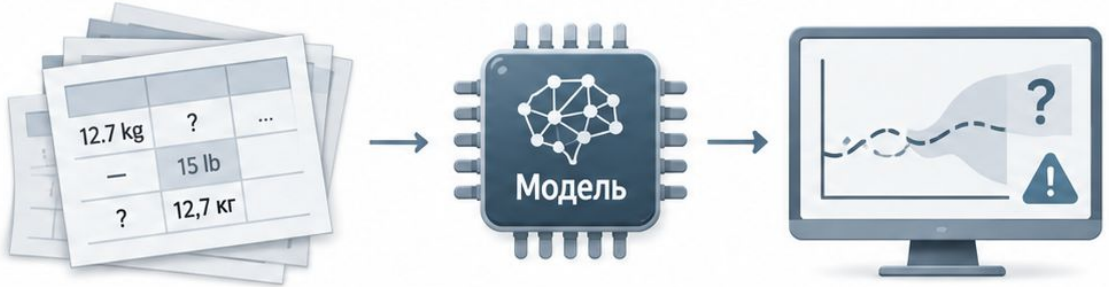
Уақыттық қатарлар хронологиялық ретпен бөлінеді. Болашақ өлшеулер өткенді болжауға қолданылмайды.

Деректер сапасының модель нәтижесіне ықпалы

Модель сапасы тек алгоритмге байланысты емес.

Сапасыз деректер

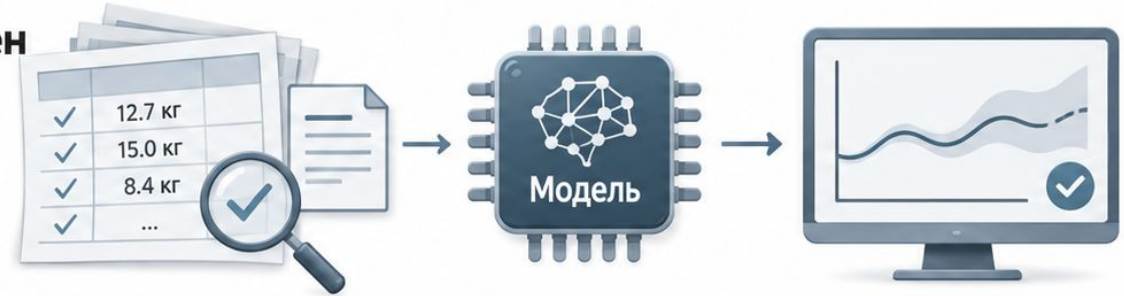
Қате өлшеулер • жетіспейтін мәндер • сәйкессіздіктер



Сенімсіз талдау және болжам қаупі

Сапасы тексерілген деректер

Дұрыс типтер • тексерілген мәндер • құжатталған өңдеу



Сенімді модель құруға жақсы негіз

Маңызды факторлар: дұрыстық • толықтық • көлем • өкілділік • дайындау тәсілі

Деректерді тазалау модель сапасын жақсартуы мүмкін, бірақ жоғары дәлдікке кепілдік бермейді. Модель архитектурасы, оқыту тәсілі және бағалау әдістемесі де маңызды.

Қорытынды

Деректерді дайындау және тазалау — деректерді талдау мен машиналық оқытудың маңызды кезеңдерінің бірі.

Бұл процесс бастапқы ақпараттың сапасын бағалауды, қателерді анықтауды, жетіспейтін мәндерді өңдеуді, деректерді түрлендіруді және модельге қажетті белгілерді дайындауды қамтиды.

Инженерлік жүйелерде деректерді тазалау кезінде өлшеулердің физикалық мағынасын сақтау ерекше маңызды. Себебі кейбір қалыптан тыс мәндер нақты ақаулықтың белгісі болуы мүмкін.

Бақылау сұрақтары

1. Деректерді дайындау дегеніміз не?
2. Деректер сапасының негізгі өлшемдері қандай?
3. Жетіспейтін деректер қандай себептерден пайда болады?
4. Қайталанатын жазба дегеніміз не?
5. Неліктен бірдей мәндерді әрқашан жоюға болмайды?
6. Диапазоннан тыс мән мен шеткі мәннің айырмашылығы қандай?
7. Нормализация және стандарттау не үшін қолданылады?
8. Feature Engineering дегеніміз не?
9. Training, Validation және Test жиындарының міндеттері қандай?
10. Неліктен инженерлік деректерді тазалау кезінде физикалық мағына ескеріледі?
11. Деректер сапасы машиналық оқыту нәтижесіне қалай әсер етеді?