№4 Дәріс. Деректерді тазалау және алдын ала өңдеу

Дәрістің мақсаты:

- Деректерді алдын ала өңдеудің маңызын түсіндіру;
- Деректерді тазалау әдістерін меңгерту;
- Жетіспейтін мәндермен, шуыл және аномалиялармен жұмыс істеуді үйрету;
- Деректерді нормализациялау, масштабтау және кодтау тәсілдерін көрсету.

Кілттік сөздер: Деректерді тазалау, алдын ала өңдеу, жетіспейтін деректер, деректер сапасы, дубликаттарды жою, шулы деректер, аномалиялар, импутация, нормализация, масштабтау, стандарттау, one-hot кодтау, label кодтау, деректерді түрлендіру, фильтрация, деректерді теңестіру, SMOTE, preprocessing pipeline, feature engineering, machine learning preprocessing.

Дәріс жоспары:

1. Деректерді тазалау түсінігі
2. Ластанған деректер түрлері
3. Жетіспейтін деректермен жұмыс (Missing data handling)
4. Шуыл және аномалиялар
5. Дубликаттарды анықтау және жою
6. Мәліметтерді нормализациялау және масштабтау
7. Категориялық деректерді өңдеу (кодтау әдістері)
8. Деректерді баланстау (imbalanced data handling)
9. Автоматтандырылған алдын ала өңдеу құралдары
10. Қорытынды

Кіріспе

Қазіргі заманда деректер ғылымы, жасанды интеллект және машиналық оқыту салаларында шешім қабылдау деректер сапасына тікелей байланысты. Егер алынған деректер қате, толық емес немесе дұрыстап құрылымдалмаған болса, ең қуатты алгоритмдер де дұрыс нәтиже бере алмайды. Осы себептен, деректерді тазалау және алдын ала өңдеу – аналитикалық жобалардың ең маңызды және ең көп уақыт алатын кезеңдерінің бірі болып табылады.
Зерттеулер бойынша, деректер ғалымдарының уақытының **70–80%** деректерді дайындау процесіне жұмсалады. Бұл деректердің күрделілігі, әркелкілігі және квадраты тәрізді қателіктерге бейімділігімен түсіндіріледі. Алдын ала өңдеу деректерді толық, түсінікті форматқа келтіріп, модельдер дәлдігін арттырады, жүйенің сенімділігін қамтамасыз етеді және болашақта аналитикалық процестердің дұрыстығын қамтамасыз етеді.

1. Деректерді тазалау түсінігі

Деректерді тазалау — бұл деректердегі қателіктерді табу, жөндеу және құрылымға келтіру процесі.

Негізгі мақсаттар:

- Қателерді түзету
- Бос мәндерді толтыру немесе жою
- Форматтарды біріздендіру
- Қажетсіз айнымалыларды түсіру

2. Ластанған деректер түрлері

Ластану түрі	Мысал
Жетіспейтін деректер	NaN, бос өріс
Қате формат	"5O" орнына "50"
Дубликаттар	Бірдей жазба қайталануы
Шуды деректер	Қате сенсор өлшемдері
Аномалиялар	Адам биіктігі 3 метр
Екі мағыналы/логикалық қате	жасы = 150

3. Жетіспейтін деректермен жұмыс

Әдістер:

Тәсіл	Сипаттамасы
Жою	Null бар жол/бағанды жою
Орташа/медиана/мода	Сандық мәнді есептеп толтыру
KNN imputation	көрші мәндер арқылы толтыру
Interpolation	Уақытша деректерге интерполяция

4. Шуыл және аномалиялар

Шуылдан тазалау:

- Фильтрация (moving average, median filter)
- Статистикалық шектеу (Z-score, IQR)

Аномалияларды анықтау:

- IQR (Interquartile Range)
- Z-score
- Machine Learning (Isolation Forest, LOF)

5. Дубликаттарды анықтау

Тәсілдер:

- Бірдей жолдарды жою
- Hashing әдісі
- Кілт бойынша тексеру (primary key, id)

6. Нормализация және масштабтау

Әдіс	Формула	Қолданылуы
Min–Max Scaling	(x − min)/(max − min)	Нейрондық желілер

Әдіс	Формула	Қолданылуы
Standardization	(x − mean)/std	Логрег, SVM
Robust Scaling	IQR қолдану	Outlier көп

7. Категориялық деректерді кодтау

Әдіс	Мысал
Label Encoding	Male → 0, Female → 1
One-Hot Encoding	Color → [1,0,0]
Target Encoding	Категорияны мақсатқа байланысты ауыстыру

8. Imbalanced Data Handling

Тәсіл	Түсінігі
Oversampling	Аз класс үлгісін көбейту (SMOTE)
Undersampling	Көп класс үлгісін азайту
Ensemble methods	Balanced Random Forest

9. Алдын ала өңдеу құралдары

- Python: pandas, numpy, scikit-learn
- Orange Data Mining
- RapidMiner
- Power BI / Excel
- Apache Spark MLlib

Қорытынды

Деректерді дұрыс өңдеу — деректер ғылымының негізгі және ең стратегиялық кезеңі. Таза және стандартизирленген деректер модельдердің нәтижелілігін арттырады, дұрыс аналитикалық шешім қабылдауға мүмкіндік береді және жүйенің қателіксіз жұмыс істеуін қамтамасыз етеді.

Дұрыс өңделмеген деректер кез келген интеллектуалды жүйені қате шешім қабылдауға итермелейді, сондықтан бұл кезеңге жауапкершілікпен және жүйелі түрде қарау қажет. Қазіргі заманғы автоматтандыратын құралдар бар болғанымен, аналитиктің тәжірибесі мен логикалық ойлауы ең маңызды фактор болып қала береді.

Бақылау сұрақтары

1. Деректерді тазалау не үшін қажет?
2. Жетіспейтін деректердің негізгі түрлері?
3. Outlier дегеніміз не? Қалай анықтаймыз?
4. One-Hot encoding қашан қолданылады?
5. Нормализация мен стандарттаудың айырмашылығы қандай?
6. Imbalanced data деген не және қалай шешіледі?

Пайдаланылған әдебиеттер

1. Han, J., Kamber, M., & Pei, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
2. Géron, A. *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. O'Reilly.
3. Provost, F., & Fawcett, T. *Data Science for Business*. O'Reilly.
4. Aggarwal, C. *Data Mining: The Textbook*. Springer.
5. Scikit-learn Documentation — preprocessing module
6. Orange Data Mining Official Documentation