

№15 Дәріс. Data Science жобасын іске асырудағы негізгі кезеңдер мен этикалық мәселелер

Дәрістің мақсаты:

Студенттерді Data Science жобасының өмірлік циклымен, жобаны жүзеге асырудың кезеңдерімен және деректерді өңдеуде кездесетін негізгі этикалық талаптармен таныстыру. Әділеттілік, құпиялық, транспаренттілік, жауапкершілік және қауіпсіздік қағидаларын түсіндіру.

Кілттік сөздер:

Деректер ғылымының өмірлік циклі, CRISP-DM, деректерді жинау, алдын ала өңдеу, модельдеу, орналастыру, бақылау, жасанды интеллекттегі этика, бейтараптық, құпиялылық, ашықтық, әділдік, GDPR, жауапты жасанды интеллект.

Кіріспе

Қазіргі заманда Data Science (деректер ғылымы) ірі корпорациялардан бастап мемлекеттік секторға дейін ең маңызды салалардың біріне айналды. Деректердің қарқынды өсуі мен есептеу қуатының артуы аналитикалық жүйелер мен жасанды интеллектке негізделген шешімдердің маңызын айқындай түсті.

Дегенмен, Data Science тек алгоритмдер жазу ғана емес — ол нақты бизнес мәселесін шешуге бағытталған кешенді процесс. Сонымен қатар, деректерді пайдалану барысында этикалық нормалар мен заңнамалық талаптарды сақтау өте маңызды.

Data Science жобасының дұрыс ұйымдастырылуы нәтижеге жетудің басты факторы болып табылады. Ал этикалық нормаларды елемей қоғамға, ұйым беделіне және шешім сапасына үлкен қауіп төндіруі мүмкін. Сондықтан жобаның өмірлік циклімен қатар этикалық аспектілерді де жақсы білу қажет.

1. Data Science жобасының негізгі кезеңдері

Data Science-те ең жиі қолданылатын модель — CRISP-DM (Cross-Industry Standard Process for Data Mining).

CRISP-DM кезеңдері:

✓1. Мәселені анықтау (Business Understanding)

- Жобаның мақсатын түсіну
- Мәселе қойылымын нақтылау
- Табыстылық критерийлерін анықтау

Мысалы:

Банк клиенттің дефолтқа ұшырауын болжағысы келеді.

2. Деректерді түсіну (Data Understanding)

- Деректер көздерін анықтау
- Деректер құрылымын зерттеу
- Алғашқы визуализация

- Аномалияларды анықтау
3. Деректерді дайындау (Data Preparation)
- Тазалау
 - Фичер инженериясы
 - Нормализация / масштабтау
 - Тренинг/тест бөлу

Бұл кезең жобаның 80% уақытын алады.

4. Модель құру (Modeling)
- Алгоритм таңдау
 - Параметрлерді баптау
 - Кросс-валидация
5. Бағалау (Evaluation)
- Модель сапасын тексеру
 - Метрикалар: Accuracy, Recall, ROC-AUC
 - Бизнес мақсатпен сәйкестік
6. Енгізу (Deployment)
- Модельді өндірістік ортаға енгізу
 - API, веб-сервис, мобильді қосымша
 - Қолданушы нұсқаулығын дайындау
7. Мониторинг және қолдау (Monitoring)
- Модель жұмысын бақылау
 - Деректер өзгерісін бақылау (data drift)
 - Модельді қайта үйрету

2. Data Science-тің қосымша әдістемелері

Әдістеме	Сипаттамасы
CRISP-DM	Ең кең қолданылатын стандарт
KDD	Knowledge Discovery in Databases
TDSP	Microsoft-тың Data Science процесі
Agile Data Science	Итеративті әдіс

3. Data Science жобаларындағы этикалық мәселелер

Деректердің құпиялылығы (Privacy)

- GDPR, HIPAA заңдары
- Жеке деректерді қорғау
- Анонимизация және шифрлау

Әділеттілік және дискриминация (Bias & Fairness)

Модель келесі факторларға әділетсіз болмауы керек:

- Жыныс
- Жасы
- Ұлты
- Дін
- Мүгедектік жағдайы

Мысал:

Жұмысқа қабылдау алгоритмі әйелді “төмен бағалады”.

Транспаренттілік (Transparency):

Пайдаланушы модель шешімінің қалай пайда болғанын түсінуі керек.

Жауапкершілік (Accountability):

Модель қатесі үшін жауапкершілік — компанияда.

Қауіпсіздік (Security):

- Хакерлік шабуылдардан қорғау
- Adversarial attacks

Этикалық принциптер:

- Fairness — Әділдік
- Transparency — Ашықтық
- Explainability — Түсіндіру мүмкіндігі
- Privacy — Деректер құпиясы
- Safety — Қауіпсіздік
- Human-centered AI — Адам мүддесі басты орында

Қорытынды

Data Science — техникалық біліммен қатар басқарушылық, заңдық және этикалық жауапкершілікті талап ететін сала. Жобаны дұрыс жоспарлау, деректерді сауатты өңдеу және этикалық нормаларды сақтау — сапалы әрі сенімді нәтижеге жеткізетін басты фактор.

Жасанды интеллект пен үлкен деректер жүйелері дамып жатқан қазіргі заманда жауапты аналитика мәдениетін қалыптастыру өте маңызды. Этикалық талаптарды сақтай отырып жүзеге асқан жоба ғана қоғамдық сенімге ие бола алады.

Бақылау сұрақтары

1. Data Science жобасының негізгі кезеңдерін атаңыз.
2. CRISP-DM моделі қандай мақсатта қолданылады?
3. Деректерді алдын ала өңдеу не үшін қажет?
4. Этикалық мәселелер қандай қауіп туғызады?
5. Bias дегеніміз не және оны қалай азайтуға болады?
6. AI жүйелеріндегі транспаренттілік неліктен маңызды?

Пайдаланылған әдебиеттер

1. Provost & Fawcett — *Data Science for Business*, 2013
2. Goodfellow, Bengio, Courville — *Deep Learning*, MIT Press
3. Foster Provost — *Data Mining and Analytics*
4. Microsoft TDSP documentation
5. CRISP-DM стандарттық нұсқаулығы
6. GDPR directives
7. Andrew Ng — *AI Ethics Lecture Notes*