Python for AI

Дәуренов ӘД

Дәріс 11. Машиналық оқытуға кіріспе және ML модельдері үшін деректерді дайындау

Бұл дәріс алдыңғы сабақтарда қарастырылған NumPy, Pandas, деректерді тазалау және визуализациялау тақырыптарының логикалық жалғасы болып табылады.

Негізгі назар машиналық оқытудың теориялық принциптеріне, модельдің қалай үйренетініне, деректер сапасының маңызына және модельдің жаңа жағдайларда дұрыс жұмыс істеуіне аударылады.

Дәрістің мақсаты

Машиналық оқытудың негізгі ұғымдарын, жұмыс істеу принциптерін және модельдің тиімді оқытылуына әсер ететін факторларды түсіндіру.

Білім алушылар:

- Машиналық оқытудың мәнін және дәстүрлі бағдарламалаудан айырмашылығын түсіндіреді.
- Машиналық оқытудың негізгі түрлерін ажыратады.
- Модельдің деректерден қалай үйренетінін сипаттайды.
- Регрессия және классификация міндеттерін ажыратады.
- Модель қатесі мен оны азайту принципін түсінеді.
- Underfitting, Overfitting және Generalization ұғымдарын түсіндіреді.
- Машиналық оқыту үшін сапалы деректердің маңызын анықтайды.

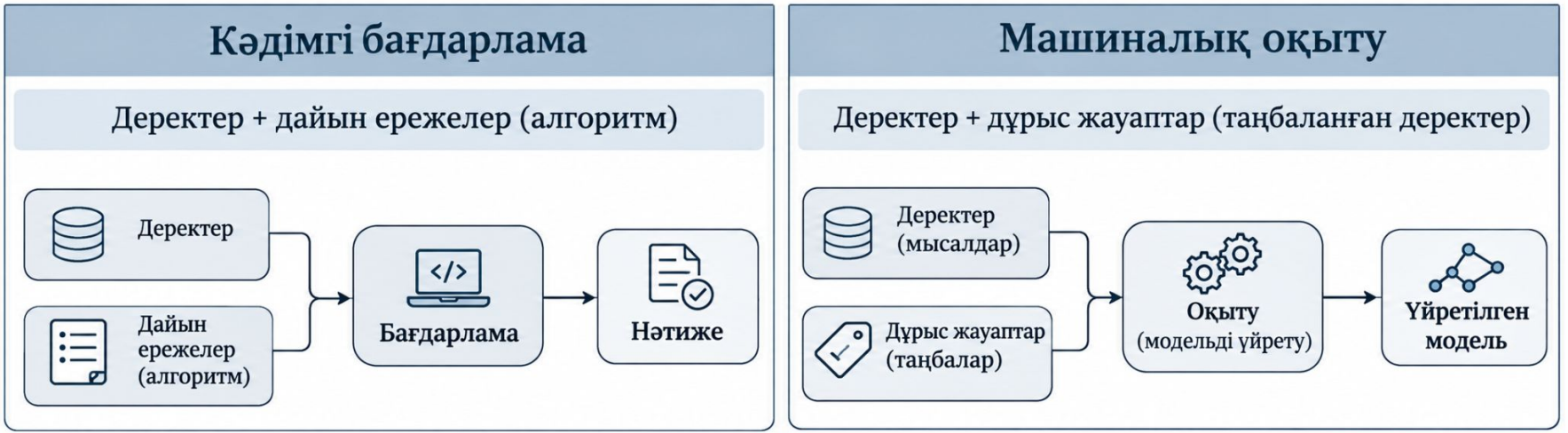
Машиналық оқыту дегеніміз не?

Машиналық оқыту (Machine Learning, ML) — компьютердің деректерден заңдылықтарды анықтап, сол заңдылықтар негізінде жаңа деректерге болжам жасауына мүмкіндік беретін әдістер жиынтығы.

Дәстүрлі бағдарламалауда есепті шешу ережелерін бағдарламашы анықтайды.
Машиналық оқытуда компьютер көптеген мысалдарды талдап, олардың арасындағы байланысты үйренеді.

Мысал: электромобиль батареясының температурасы, ток күші және пайдалану жағдайлары туралы деректер арқылы оның болашақтағы температурасын болжау.

Модель дайын жауапты жаттамай, көрсеткіштер арасындағы байланысты үйренуге тырысады.


Кәдімгі бағдарлама
Деректер + дайын ережелер (алгоритм)
Деректер
Дайын ережелер (алгоритм)
Бағдарлама
Нәтиже
Машиналық оқыту
Деректер + дұрыс жауаптар (таңбаланған деректер)
Деректер (мысалдар)
Дұрыс жауаптар (таңбалар)
Оқыту (модельді үйрету)
Үйретілген модель

```
temperature = 55

if temperature > 45:
    print("WARNING")
else:
    print("NORMAL")
```

Мұнда температураның шекті мәнін бағдарламашы өзі белгілейді.

Машиналық оқытуда жүйе көптеген өлшеулер мен олардың нәтижелерін талдап, температураға байланысты белгілі бір жағдайдың туындау ықтималдығын болжауды үйренуі мүмкін.

Машиналық оқыту барлық жағдайда дәстүрлі бағдарламалауды алмастырмайды. Мысалы, қауіпсіздік жүйесіндегі нақты шекті мәндер мен қорғаныс ережелері әлі де маңызды.

Оқыту процесін қарапайым түрде төрт кезеңге бөлуге болады:

Машиналық оқытудағы негізгі идея

Машиналық оқытудағы негізгі идея — болжау жасау және қателікті азайту.

Оқыту процесінің кезеңдері



- Оқыту алгоритмі осы әрекеттер арқылы модельдің болжау жасау қабілетін жақсартуға тырысады.
- Барлық ML алгоритмдері бірдей тәсілмен оқытылмайды, бірақ болжам сапасын жақсарту — олардың ортақ мақсаттарының бірі.

Машиналық оқытудың негізгі түрлері

Машиналық оқытудың үш кең таралған түрі бар.

Түрі	Негізгі принципі	Мысал
Бақыланатын оқыту (Supervised Learning)	Дұрыс жауаптары белгілі деректер арқылы үйрену	Автомобиль бағасын болжау
Бақыланбайтын оқыту (Unsupervised Learning)	Дұрыс жауаптарсыз деректердегі құрылымды анықтау	Автомобильдерді ұқсастығына қарай топтау
Нығайтпалы оқыту (Reinforcement Learning)	Әрекеттер нәтижесіндегі марапатқа сүйеніп үйрену	Роботтың қозғалысты басқаруды үйренуі

Бақыланатын оқыту (Supervised Learning)

Бақыланатын оқыту — кіріс деректерімен бірге оларға сәйкес дұрыс нәтижелер берілетін машиналық оқыту түрі.

Негізгі міндеттері



Регрессия
Сандық мәнді болжау (мысалы, баға)



Классификация
Объектіні санаттарға бөлу (мысалы, авто: арзан / қымбат)

Мысал: Автомобильдер туралы деректер

Шығарылған жылы	Жүрілген жолы	Бағасы
2015	180 000 км	8 000 000 ₸
2018	110 000 км	11 000 000 ₸
2021	55 000 км	15 000 000 ₸



Оқыту (модельді құру)



Жаңа автомобиль

Шығарылған жылы: 2023
Жүрілген жолы: 40 000 км

Болжам: Бағасы ≈ 16 000 000 ₸

- Модель автомобильдің шығарылған жылы мен жүрілген жолы сияқты сипаттамаларды беггілі бағалармен салыстырып үйренеді.
- Кейін жаңа автомобильдің сипаттамалары берілгенде, оның ықтимал бағасын болжай алады.

Бақыланбайтын оқыту *(Unsupervised Learning)*

- **Бақыланбайтын оқыту** — алдын ала белгіленген дұрыс жауаптарсыз деректердің ішкі құрылымын анықтау тәсілі.
- Модель ұқсас объектілерді топтастыруы немесе деректердегі ерекше заңдылықтарды анықтауы мүмкін.

Негізгі міндеті



Деректерді ұқсас топтарға бөлу (топтастыру)



Деректердегі ерекше заңдылықтарды анықтау

Мысал

Электромобильдер туралы деректерді талдау



Энергия тұтынуы (кВт·сағ/100 км)

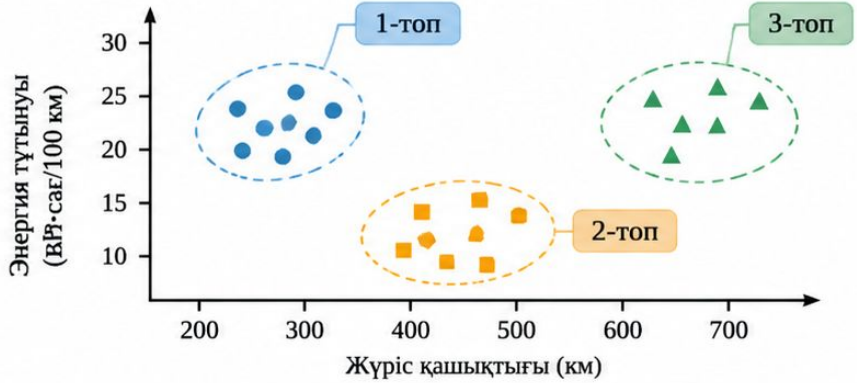


Жүріс қашықтығы (км)



Батарея сыйымдылығы (кВт·сағ)

Топтастыру нәтижесінің мысалы

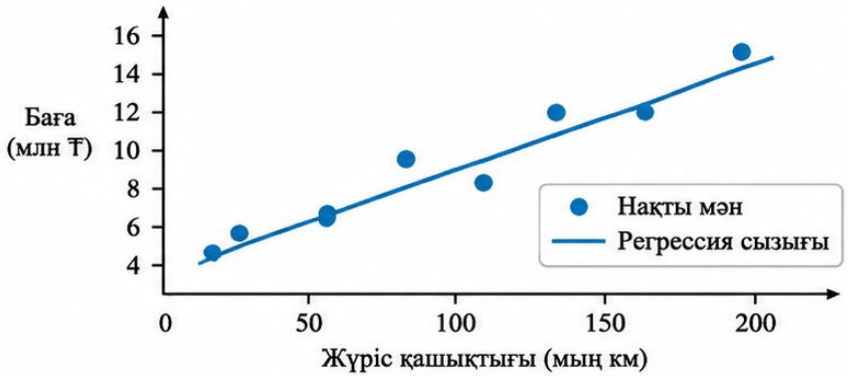


- Алгоритм автомобильдерді ұқсас сипаттамаларына карай бірнеше топқа бөлуі мүмкін. Бұл топтар алдын ала берілмейді, олар деректерді талдау нәтижесінде анықталады.

Бақыланатын оқытудың негізгі міндеттері

■ Бақыланатын оқытудың негізгі міндеттері — регрессия және классификация.

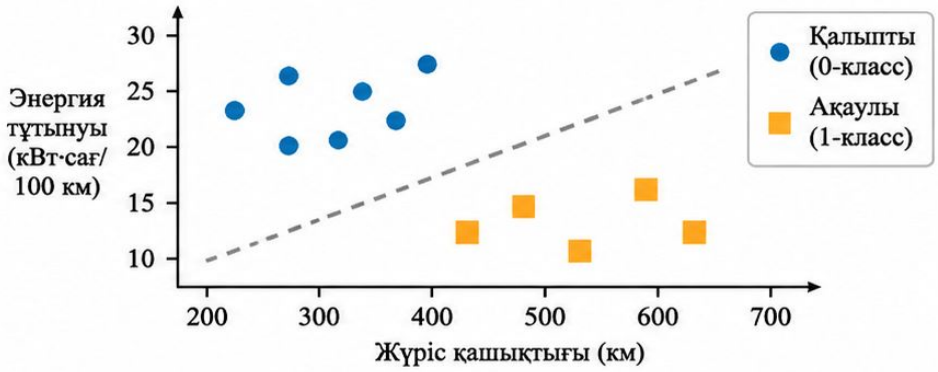
Регрессия (Regression)



Нәтижесі: сандық мән (үздіксіз шама).

Мысал: автомобиль бағасын немесе энергия тұтынуын болжау.

Классификация (Classification)



Нәтижесі: санат немесе класс (дискретті белгі).

Мысал: техникалық жағдайды «қалыпты» немесе «ақаулы» деп анықтау.

Регрессия сандық нәтижені болжайды, ал классификация объектінің қай санатқа жататынын анықтайды.

Машиналық оқыту моделі дегеніміз не?

Модель (Model) — деректерден үйренілген заңдылықтар негізінде нәтиже есептейтін математикалық немесе алгоритмдік жүйе.

Оқыту алгоритмі (Learning Algorithm) — модельді деректерге сәйкес қалыптастыру немесе оның параметрлерін анықтау тәсілі.

Параметрлер (Parameters) — модельдің оқыту барысында анықталатын ішкі мәндері.

Қарапайым сызықтық модель *(Simple Linear Model)*

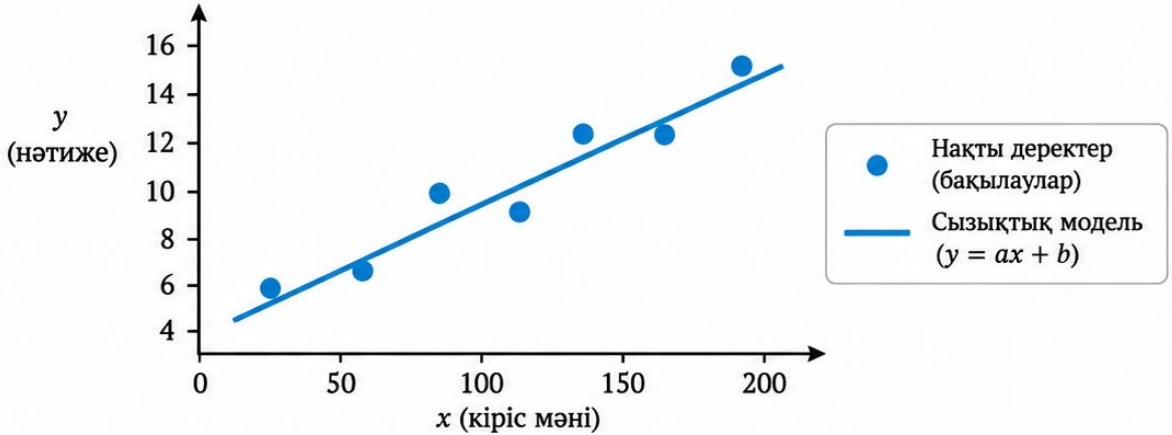
Модель формуласы

$$y = ax + b$$

Параметрлердің мәні

x	— кіріс мәні (тәуелсіз айнымалы);
y	— болжанатын нәтиже (тәуелді айнымалы);
a	— көлбеу коэффициенті (байланыс күшін сипаттайды);
b	— бос мүше (бастапқы ығысу).

Графикалық көрінісі



Модельді оқыту барысында алгоритм a және b мәндерін деректерге сәйкес анықтайды.

Бұл — машиналық оқытудың ең **қарапайым** мысалдарының бірі. Күрделі модельдерде параметрлер саны әлдеқайда көп болуы мүмкін.

Негізгі идея: кіріс пен нәтиже арасындағы байланысты қарапайым сызықтық теңдеу арқылы сипаттау.

Модельді үйрету процесі және қателікті азайту

Мақсаты

Модельдің үйрену принципін қарапайым мысал арқылы қарастыру.

Зерттеу мысалы

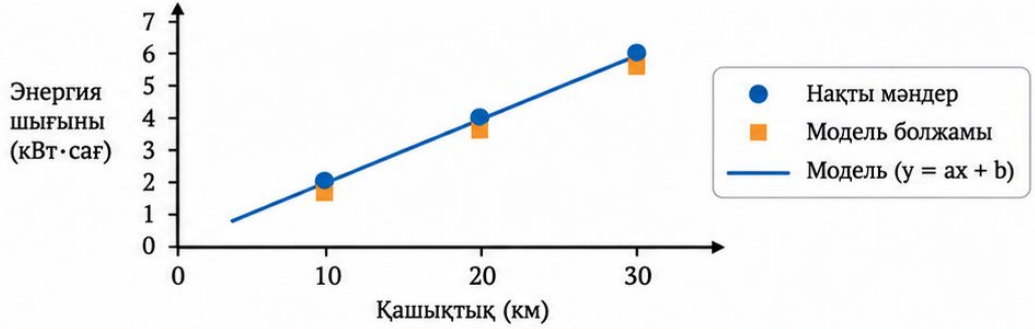


Электромобильдің жүрген қашықтығы мен жұмсалған энергиясы арасындағы байланысты зерттеу.

Деректер және модель болжамы

Қашықтық (км)	Нақты энергия шығыны (кВт·сағ)	Модель болжамы (кВт·сағ)
10	2,0	1,8
20	4,0	3,7
30	6,0	5,8

Нәтиже: нақты және болжамды мәндер



Негізгі қорытынды

- Модель болжамдары нақты мәндерден аздап ерекшеленеді.
- Оқыту алгоритмі осы айырмашылықтарды ескеріп, модель параметрлерін өзгерту арқылы қателікті азайтуға тырысады.

Негізгі идея: модельдің болжамы нақты нәтижеге мүмкіндігінше жақын болуы керек.

Жалпылау қабілеті (Generalization)

Жалпылау қабілеті (Generalization) — модельдің оқыту кезінде көрмеген жаңа деректерге дұрыс болжам жасай алу қасиеті.

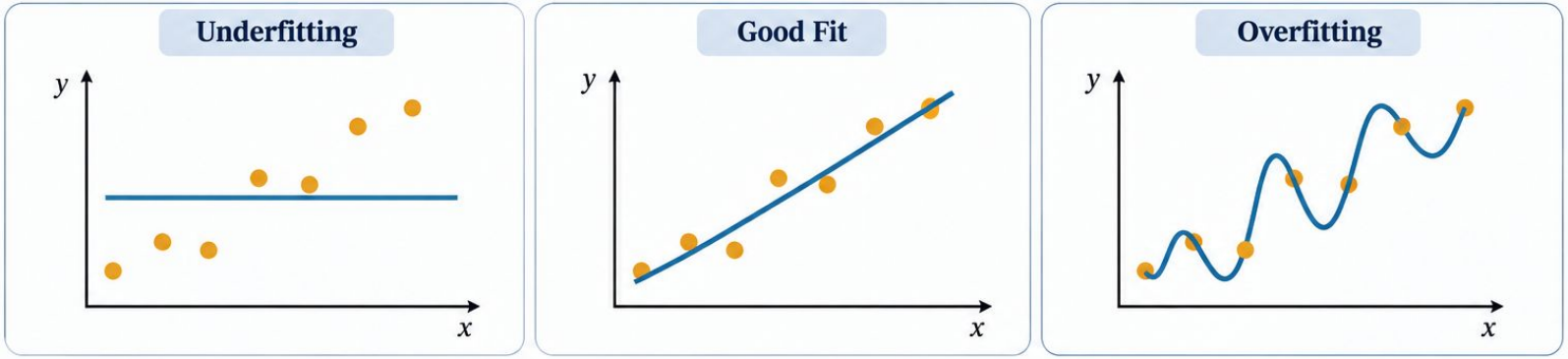
Мысалы, модель белгілі бір автомобильдердің бағалары бойынша оқытылды.
Кейін сол модельге бұрын кездеспеген автомобильдің сипаттамалары беріледі.

Егер модель оның бағасын жеткілікті дәл болжай алса, бұл модельдің жаңа деректерге жалпылау қабілеті бар екенін көрсетеді.

Машиналық оқытудың мақсаты — тек оқыту мысалдарын есте сақтау емес, жаңа жағдайларда да пайдалы нәтиже беру.

Модельдің жалпы үйренуі: жеткіліксіз және артық үйрену

- **Underfitting (жеткіліксіз үйрену)** — модель деректердегі негізгі заңдылықтарды жеткілікті меңгере алмауы.
- **Overfitting (артық үйрену)** — модель оқыту деректерінің кездейсоқ ерекшеліктеріне шамадан тыс бейімделуі.



Үш жағдайдың шартты салыстырылуы: нүктелер — деректер, сызықтар — модель.

Жағдай	Оқыту деректеріндегі нәтиже	Жаңа деректердегі нәтиже
Underfitting	Нашар	Әдетте нашар
Good Fit	Жақсы	Жақсы
Overfitting	Өте жақсы болуы мүмкін	Нашар болуы мүмкін

Қорытынды: тиімді модель негізгі заңдылықтарды үйреніп, кездейсоқ ауытқуларға шамадан тыс бейімделмеуі тиіс.

ML үшін деректердің сапасы неге маңызды?

Модельдің сапасы тек алгоритмге ғана емес, оқыту деректерінің сапасына да байланысты.

Машиналық оқыту үшін маңызды талаптар:

- **Дұрыстық:** мәндер нақты объектіні дұрыс сипаттауы қажет.
- **Толықтық:** маңызды сипаттамалар мүмкіндігінше қамтылуы тиіс.
- **Репрезентативтілік:** деректер модель қолданылатын жағдайларды көрсетуі қажет.
- **Әртүрлілік:** мысалдар тек бір жағдаймен шектелмеуі керек.
- **Сәйкестік:** деректер шешілетін міндетке қатысты болуы тиіс.

Мысалы, электромобиль батареясының температурасын болжау үшін тек қалыпты ауа райында жиналған өлшеулер жеткіліксіз болуы мүмкін.

Себебі модель суық немесе ыстық жағдайлардағы өзгерістерді үйренбеген болады.

Деректерді бөлу және валидация

Деректердің бөлінуі

Машиналық оқытуда деректер көбінесе үш мақсатқа бөлінеді.

Деректер жиыны	Мақсаты
Training Set	Модельдің параметрлерін оқыту
Validation Set	Модель нұсқасын және баптауларын таңдау
Test Set	Соңғы модельдің жаңа деректердегі сапасын бағалау

Деректерді бөлу мысалы

Мысал: **300 бақылау**



60% / 20% / 20% — түсіндіруге арналған ықтимал бөлу мысалы, міндетті қатынас емес.

Кросс-валидация (*Cross-validation*) — деректер көлемі аз болғанда валидация үшін қолданылатын тәсіл. Онда деректер бірнеше бөлікке бөлініп, модель кезекпен әр бөлікке валидацияланады.

Негізгі принцип — тестілеу деректерін модельді таңдау немесе баптау үшін пайдаланбау.

Деректердің ағып кетуі (Data Leakage)

Data Leakage — модельді оқыту немесе таңдау барысында нақты болжау сәтінде қолжетімсіз ақпараттың пайдаланылуы.
Бұл модель сапасын шынайы деңгейінен жоғары етіп көрсетуі мүмкін.

Мысалдар:

- Тестілеу деректерінің статистикасын оқыту кезіндегі түрлендіруге пайдалану.
- Болашақта ғана белгілі болатын ақпаратты кіріс белгісі ретінде қолдану.
- Бір объектіге қатысты бірдей немесе өте ұқсас жазбалардың оқыту және тестілеу жиындарына қатар түсуі.

Мысалы, бір автомобиль туралы қайталанған хабарландырулар екі жиында кездессе, модель тестілеу кезінде бұрын көрген автомобильді қайта кездестіруі мүмкін.
Сондықтан деректерді дайындауда олардың тәуелсіздігі мен ақпараттың қолжетімді уақытын ескеру қажет.

Машиналық оқытудың шектеулері

Машиналық оқыту көптеген міндеттерді шешуге көмектескенімен, оның мүмкіндіктері шексіз емес.

Негізгі шектеулер:

- Деректер жеткіліксіз немесе сапасыз болуы мүмкін.
- Модель бұрын кездеспеген жағдайларда қателесуі мүмкін.
- Деректердегі біржақтылық модель нәтижесіне әсер етуі мүмкін.
- Кейбір күрделі модельдердің шешімін түсіндіру қиын.
- Модельдің болжамы әрқашан нақты нәтижеге тең болмайды.

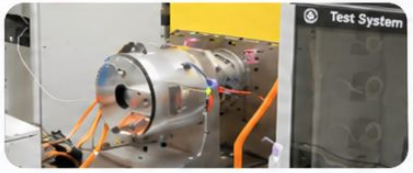
Мысал: автомобиль бағасын болжайтын модель оның шығарылған жылы мен жүрілген жолын ескеруі мүмкін, бірақ жол-көлік оқиғасынан кейінгі жасырын зақымдарын білмеуі ықтимал. Сондықтан модель нәтижесін абсолютті дұрыс шешім емес, белгілі бір белгісіздігі бар болжам ретінде қарастыру керек.

Машиналық оқытудың қолдану мысалдары



Батарея күйін болжау

Батареяның температурасы, кернеуі, ток күші және пайдалану тарихы негізінде оның күйін бағалау.



Техникалық ақауларды анықтау

Қозғалтқыш пен басқа құрылғылардың өлшеу деректерін талдап, ақаудың ықтималдығын бағалау.



Көлік ағынын болжау

Бұрынғы қозғалыс деректері негізінде белгілі бір уақыттағы көлік тығыздығын болжау.

Қорытынды: Бұл міндеттердің әрқайсысында машиналық оқыту әдісі деректердің түріне, мақсатты нәтижеге және жүйенің талаптарына қарай таңдалады.

Қорытынды

Машиналық оқыту — компьютердің деректерден заңдылықтарды үйренуіне негізделген әдістер жиынтығы.

Машиналық оқытуда модель мен оқыту алгоритмі әртүрлі ұғымдар: алгоритм модельді қалыптастырады, ал модель жаңа деректерге болжам жасайды.

Модельді оқытуда қателікті азайту маңызды, бірақ тек оқыту деректеріндегі жақсы нәтиже жеткіліксіз.

Тиімді машиналық оқытудың негізгі мақсаты — жаңа деректерге жалпылау қабілеті жоғары модель құру.
Бұл үшін сапалы және репрезентативті деректер, дұрыс оқыту және бағалау тәртібі қажет.

Бақылау сұрақтары

1. Машиналық оқыту дегеніміз не?
2. Дәстүрлі бағдарламалау мен машиналық оқытудың айырмашылығы қандай?
3. Машиналық оқытудың қандай негізгі түрлері бар?
4. Бақыланатын және бақыланбайтын оқыту қалай ерекшеленеді?
5. Регрессия мен классификацияның айырмашылығы қандай?
6. Модель және оқыту алгоритмі дегеніміз не?
7. Қате функциясы қандай мақсатта қолданылады?
8. Модель параметрлерін өзгерту не үшін қажет?
9. Generalization дегеніміз не?
10. Underfitting және Overfitting қалай ерекшеленеді?
11. Репрезентативті деректер неліктен маңызды?
12. Training, Validation және Test жиындарының міндеттері қандай?
13. Data Leakage дегеніміз не?
14. Машиналық оқытудың негізгі шектеулері қандай?