**№1 Дәріс тақырыбы: Үлкен көлемді деректер және машиналық оқыту негіздері**

**Кіріспе:**

Үлкен көлемді деректер (Big Data) мен машиналық оқыту (Machine Learning) – қазіргі замандағы ақпараттық технологиялардың негізгі бағыттарының бірі. Бүгінгі таңда ақпарат көлемінің қарқынды өсуі, жаңа есептеу технологиялары мен әдістерінің пайда болуы үлкен деректерді талдау мен өңдеудің маңызды мәселелерін туындатады. Осы дәрісте біз үлкен деректердің және машиналық оқытудың негізгі ұғымдарына тоқталамыз.

**1. Үлкен көлемді деректер (Big Data)**

**1.1 Үлкен деректер дегеніміз не?**

Үлкен көлемді деректер – көлемі өте үлкен және күрделі болғандықтан, дәстүрлі дерекқорлар мен бағдарламалық құралдар арқылы өңделмейтін деректер. Үлкен деректердің анықтамасы әдетте **3V** арқылы беріледі:

* **Көлем (Volume)**: Деректердің шексіз көлемі. Деректер көлемі терабайт, петабайт және эксабайттарға жетеді.
* **Жылдамдық (Velocity)**: Деректердің жылдамдықпен жиналуы мен өңделуі. Мәліметтерді нақты уақыт режимінде (real-time) өңдеу қажеттілігі туындайды.
* **Түрлілік (Variety)**: Деректердің әртүрлі форматтарда (текст, бейне, аудио, логтар, құрылғылардан алынған деректер) болуы.

Үлкен деректерден ақпарат алу үшін заманауи технологиялар мен платформаларды қолдану қажет, мысалы: Hadoop, Spark, Cassandra, және т.б.

**1.2 Үлкен деректердің түрлері:**

Үлкен деректерді үш негізгі түрге бөлуге болады:

1. **Құрылымдалған деректер (Structured Data)**: Бұл алдын ала белгілі бір форматқа түсірілген деректер. Мысалы, дерекқорлардағы жазбалар, кестелер.
2. **Құрылымдалмаған деректер (Unstructured Data)**: Бұл деректер қандай да бір форматсыз беріледі. Мысалы, мәтіндер, бейнелер, аудио файлдар.
3. **Жартылай құрылымдалған деректер (Semi-structured Data)**: Кейбір құрылымы бар, бірақ толық емес. Мысалы, JSON немесе XML форматындағы деректер.

**1.3 Үлкен деректердің маңыздылығы:**

Үлкен деректер бизнес, ғылым, медицина, қаржы салаларында маңызды шешімдер қабылдауға мүмкіндік береді. Оның көмегімен:

* Бизнес процестерді оңтайландыруға болады.
* Клиенттердің қажеттіліктерін жақсырақ түсініп, қызмет көрсету сапасын арттыруға болады.
* Денсаулық сақтау саласында диагноз қою және емдеу әдістерін жақсарту мүмкіндігі туады.

**2. Машиналық оқыту (Machine Learning)**

**2.1 Машиналық оқыту дегеніміз не?**

Машиналық оқыту – компьютерлерге нақты бағдарламалаусыз, деректер арқылы үйренуге мүмкіндік беретін жасанды интеллекттің бір бағыты. Басқаша айтқанда, машиналық оқыту алгоритмдері өздеріне берілген деректер негізінде шешім қабылдауды үйренеді.

**2.2 Машиналық оқытудың негізгі түрлері:**

Машиналық оқытуда негізгі үш әдіс бар:

1. **Қадағаланатын оқыту (Supervised Learning)**:
   * Бұл әдісте модельге дайын деректер жиыны беріледі, яғни деректермен бірге олардың дұрыс жауаптары (target labels) да беріледі.
   * Модель үйрену барысында дұрыс нәтижелерді болжауға тырысады.
   * Қолданылатын негізгі алгоритмдер: сызықтық регрессия (Linear Regression), логистикалық регрессия (Logistic Regression), шешім ағаштары (Decision Trees), нейрондық желілер (Neural Networks).
   * Мысал: Үйдің бағасын болжау, клиенттердің сатып алу ықтималдығын болжау.
2. **Қадағаланбайтын оқыту (Unsupervised Learning)**:
   * Бұл әдісте деректердің дұрыс жауаптары жоқ. Модель деректердің ішкі құрылымын анықтауға тырысады.
   * Қолданылатын негізгі алгоритмдер: кластерлеу (Clustering), басты компоненттер анализі (PCA).
   * Мысал: Клиенттерді сегментациялау, деректерді қысу.
3. **Нықтыландырып оқыту (Reinforcement Learning)**:
   * Бұл әдісте агент (модель) қоршаған ортада әрекет жасайды және әрбір әрекет үшін марапат (reward) немесе жаза (penalty) алады.
   * Агенттің мақсаты – қоршаған ортаны зерттеп, ең көп марапат алатын стратегияны үйрену.
   * Мысал: Роботтың кедергілерден өтуі, ойындарда стратегия құру.

**2.3 Машиналық оқытудың негізгі қадамдары:**

1. **Деректерді жинау**: Алдымен, модельді үйрету үшін деректер жиналады.
2. **Деректерді өңдеу және тазалау**: Жетпейтін мәндер, қайталанатын жазбалар және артық деректер жойылып, деректер дайындалады.
3. **Модель құру және таңдау**: Деректер негізінде машиналық оқыту алгоритмі таңдалып, модель құрылады.
4. **Модельді оқыту**: Модельге деректер беріліп, ол үйренеді.
5. **Модельді тексеру және бағалау**: Модельдің тиімділігін тексеру үшін тест деректері қолданылады.
6. **Модельді жетілдіру**: Модельді жақсарту үшін параметрлерді өзгерту, деректер көлемін ұлғайту сияқты әдістер қолданылады.

**2.4 Машиналық оқытуда жиі қолданылатын кітапханалар:**

Python тілі машиналық оқытуда кеңінен қолданылады, себебі оның көптеген кітапханалары бар:

* **Scikit-learn**: Қарапайым алгоритмдер мен модельдер үшін қолданылатын кітапхана.
* **TensorFlow** және **Keras**: Нейрондық желілер және терең оқыту үшін қолданылады.
* **PyTorch**: Тензорлармен жұмыс істейтін терең оқыту кітапханасы.
* **Pandas**, **NumPy**: Деректерді өңдеуге арналған кітапханалар.

**3. Үлкен деректер мен машиналық оқытудың байланысы**

**3.1 Үлкен деректер машиналық оқыту үшін неге маңызды?**

Машиналық оқыту алгоритмдері жақсы жұмыс істеу үшін көп деректерді қажет етеді. Үлкен деректер машиналық оқыту модельдерін дәлірек және нәтижелі етіп жасауға мүмкіндік береді. Көбірек деректер алгоритмдерге үлгілерді жақсырақ тануға, шешімдерді дәлірек қабылдауға жағдай жасайды.

**3.2 Машиналық оқыту үлкен деректерді қалай қолданады?**

* **Деректерді талдау:** Машиналық оқыту алгоритмдері үлкен деректерді талдап, олардың ішкі заңдылықтарын анықтауға мүмкіндік береді.
* **Автоматтандырылған шешім қабылдау:** Мысалы, интернет-дүкендерде қолданушылардың мінез-құлқын талдап, ұсыныстар беру.
* **Үлкен деректерді кластерлеу:** Қадағаланбайтын оқыту әдістері арқылы деректерді автоматты түрде кластерлерге бөлу.

**Қорытынды:**

Үлкен деректер мен машиналық оқыту қазіргі заманғы ақпараттық технологиялар мен бизнес процестердің ажырамас бөлігі болып табылады. Үлкен деректерді тиімді талдау және өңдеу машиналық оқыту модельдерінің көмегімен жүзеге асып, бизнеске, ғылымға және өнеркәсіпке жаңа мүмкіндіктер ашады. Сондықтан осы екі бағыттың негізгі түсініктерін меңгеру өте маңызды.