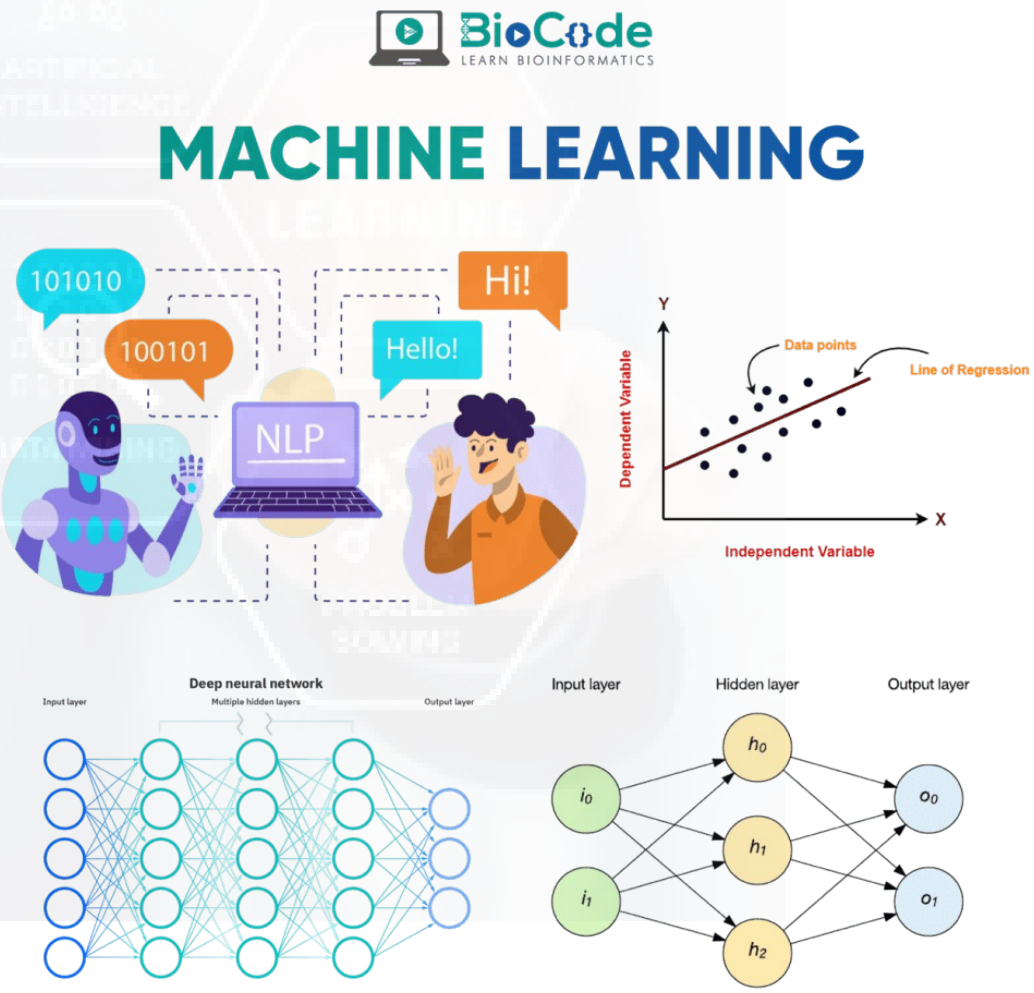
Машиналық оқыту әдістерін биоинформатикада қолдану

6-Дәріс

Кіріспе

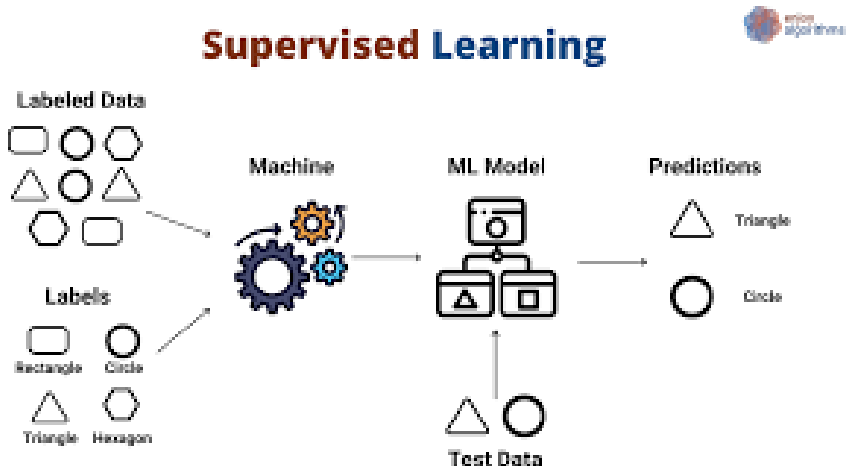
Машиналық оқыту әдістері соңғы жылдары биоинформатика саласында айтарлықтай назар аударып келе жатыр. Бұл әдістер үлкен биомедициналық деректермен жұмыс істеуді жеңілдетеді, жаңа биологиялық заңдылықтарды анықтауға мүмкіндік береді және зерттеушілерге түрлі биомедициналық мәселелерді шешуге көмектеседі. Биоинформатикадағы машиналық оқытудың қолдану салалары генетикалық деректерді өңдеу, ақуыздар құрылымын болжау, геномдық реттіліктерді талдау және аурулардың диагностикалық модельдерін құру сияқты кең ауқымды қамтиды.

Бүгінгі таңда, машиналық оқыту әдістерін қолдану арқылы биомолекулалардың арасындағы өзара әрекеттестікті болжау, гендер мен микрорНҚ молекулаларының байланысын зерттеу, сондай-ақ биомедициналық деректерді жүйелік талдау жүргізу мүмкіндігі бар. Бұл әдістер биоинформатикада дәлдікті арттырып қана қоймай, деректердің үлкен көлемдерін қысқа мерзімде талдауға мүмкіндік береді.



- **Машиналық оқыту негіздері**

Машиналық оқыту — бұл алгоритмдерге деректерді пайдалана отырып, автоматты түрде үйренуге және тәжірибені қолдану арқылы өз көрсеткіштерін жақсартуға мүмкіндік беретін әдістер жиынтығы. Биоинформатика саласында машиналық оқыту деректердің үлкен көлемін тиімді өңдеуге және жаңа биологиялық заңдылықтарды анықтауға көмектеседі. Машиналық оқытудың негізгі түрлері бар, олардың әрқайсысы өзіне тән ерекшеліктерге ие және биоинформатикадағы әртүрлі мәселелерді шешуде қолданылады.

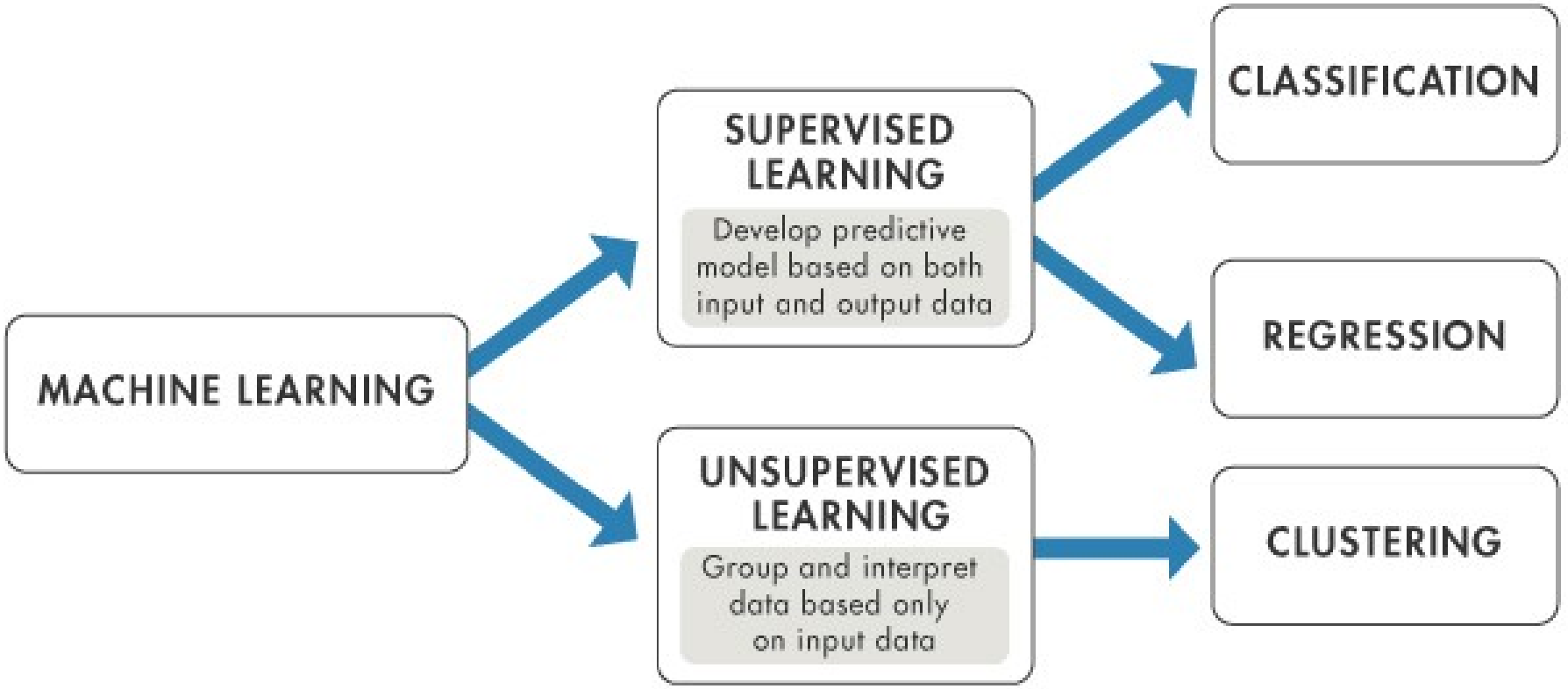


- **Машиналық оқытудың негізгі түрлері:**

1. **Бақылаулы оқыту (Supervised Learning)**
   Бұл әдісте модельге алдын ала анықталған кіріс және шығыс деректер жиыны беріледі, содан кейін модель осы деректерге қарап, жаңа, белгісіз деректерді болжауды үйренеді. Мысалы, биоинформатикада бұл әдіс аурудың даму ықтималдығын болжау үшін генетикалық деректерді талдауда қолданылады. Негізгі бақылаулы оқыту алгоритмдеріне сызықтық регрессия, логистикалық регрессия және шешім ағаштары жатады.

Бақылаусыз оқыту (Unsupervised Learning)

Бұл әдісте модельге тек кіріс деректер беріледі, ал шығыс белгілерсіз болып келеді. Модель өздігінен деректерден заңдылықтарды анықтайды. Бұл әдіс деректерді кластерлеу, ерекшеліктерді анықтау үшін пайдаланылады. Биоинформатикада бақылаусыз оқыту геномдық деректерді топтастыру және кластерлеу сияқты мәселелерді шешуде кеңінен қолданылады.



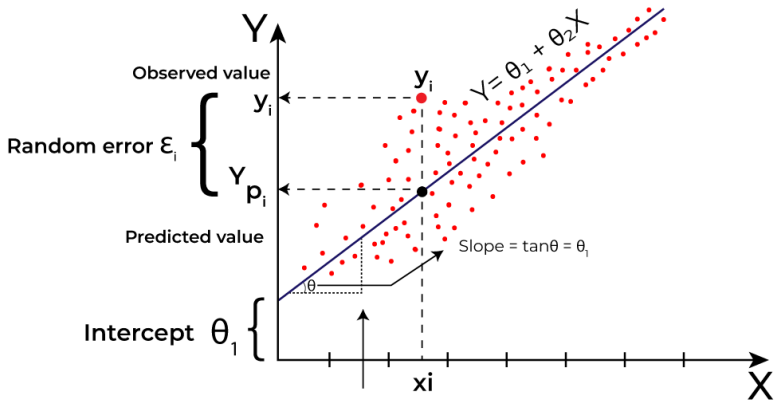
- **Жартылай бақылаулы оқыту (Semi-supervised Learning)**
- Бұл әдіс бақылаулы және бақылаусыз оқытудың комбинациясы болып табылады. Модельге шағын көлемдегі белгілері бар және үлкен көлемдегі белгісіз деректер беріледі. Жартылай бақылаулы оқыту биоинформатикада белгісіз генетикалық деректерді топтастыру және жаңа биомаркерлерді анықтауда тиімді.
- **Нығайту арқылы оқыту (Reinforcement Learning)**Бұл әдісте модель әрекеттерді орындап, ортадан кері байланыс алады, сол кері байланыс негізінде өз стратегиясын жақсартады. Бұл әдіс көбінесе күрделі жүйелерді модельдеуде және биоинформатикадағы динамикалық процестерді басқаруда қолданылады.

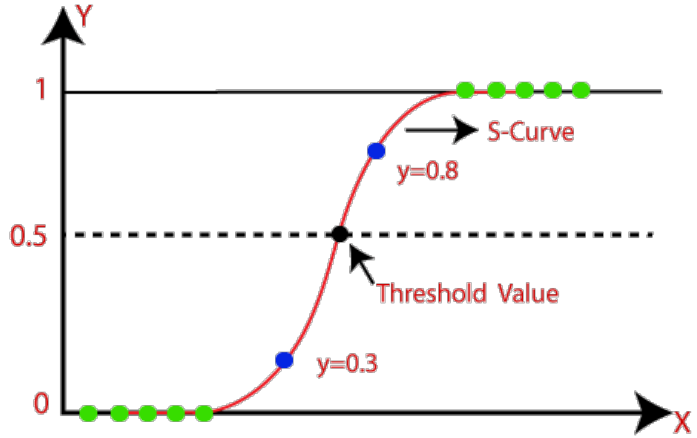
Негізгі модельдер:

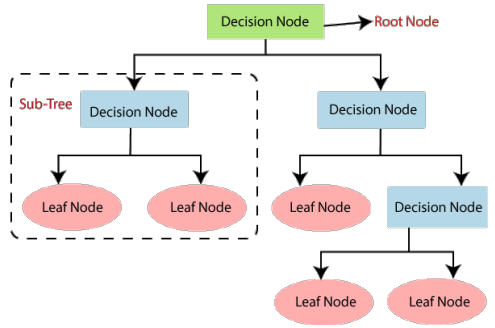
1.Сызықтық регрессия (Linear Regression)
Сызықтық регрессия — кіріс деректері мен шығыс арасындағы сызықтық тәуелділікті анықтайтын модель. Бұл әдіс биомедициналық көрсеткіштер мен ауру қаупінің арасындағы байланыстарды бағалау үшін қолданылады.

2.Логистикалық регрессия (Logistic Regression)
Логистикалық регрессия — екі немесе бірнеше кластарды болжау үшін қолданылатын әдіс. Биологиялық деректерде бұл модель аурудың бар-жоғын немесе гендердің экспрессиялық деңгейін болжау үшін тиімді.

3.Шешім ағаштары (Decision Trees)
Шешім ағаштары — деректерді шешімдер қабылдау ағашына ұйымдастыратын модель. Бұл әдіс генетикалық мутациялар мен ауру арасындағы байланыстарды анықтауға мүмкіндік береді.

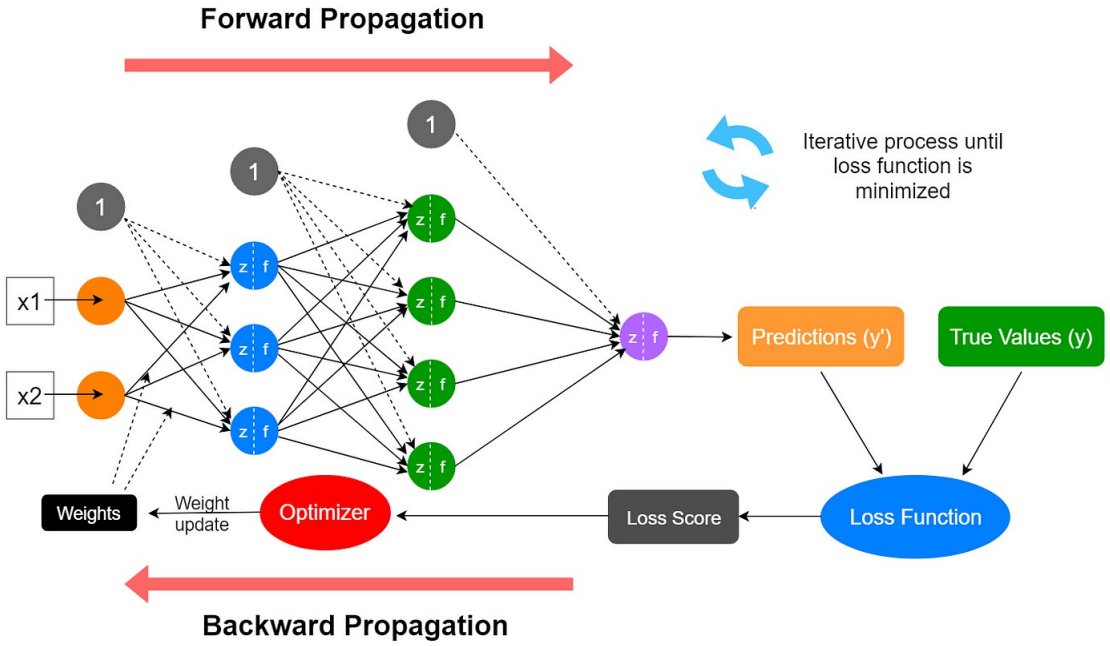






4. Нейрондық желілер (Neural Networks)
Нейрондық желілер — биоинформатикада кеңінен қолданысқа ие модельдер, себебі олар үлкен көлемдегі күрделі деректерді өңдеуге және жаңа заңдылықтарды анықтауға қабілетті. Нейрондық желілер генетикалық деректерді талдауда, ақуыздардың құрылымын болжауда және басқа да биологиялық процестерді модельдеуде тиімді.

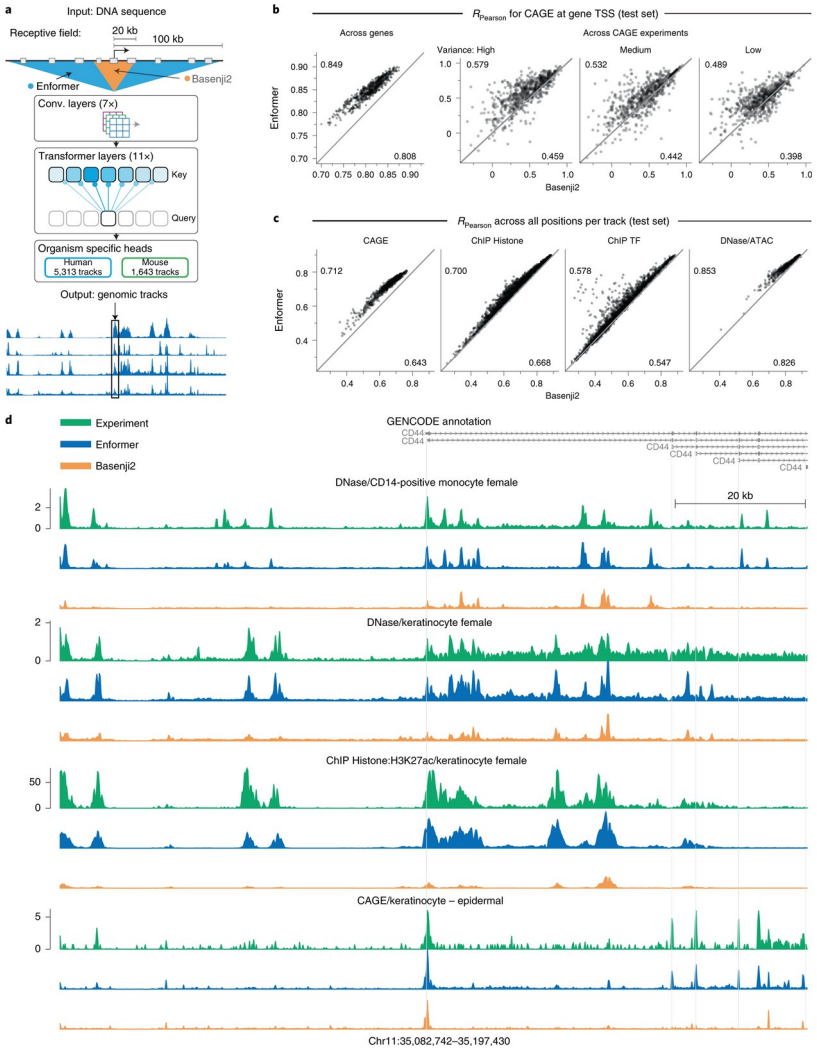
5. Терең оқыту модельдері (Deep Learning Models)
Терең оқыту — нейрондық желілердің көп қабатты нұсқасы, ол деректердің күрделі және көпқабатты құрылымдарын үйренуге қабілетті. Бұл модельдер биомедициналық суреттерді тану, геномдық деректерді талдау және ауруларды болжауда кеңінен қолданылады.



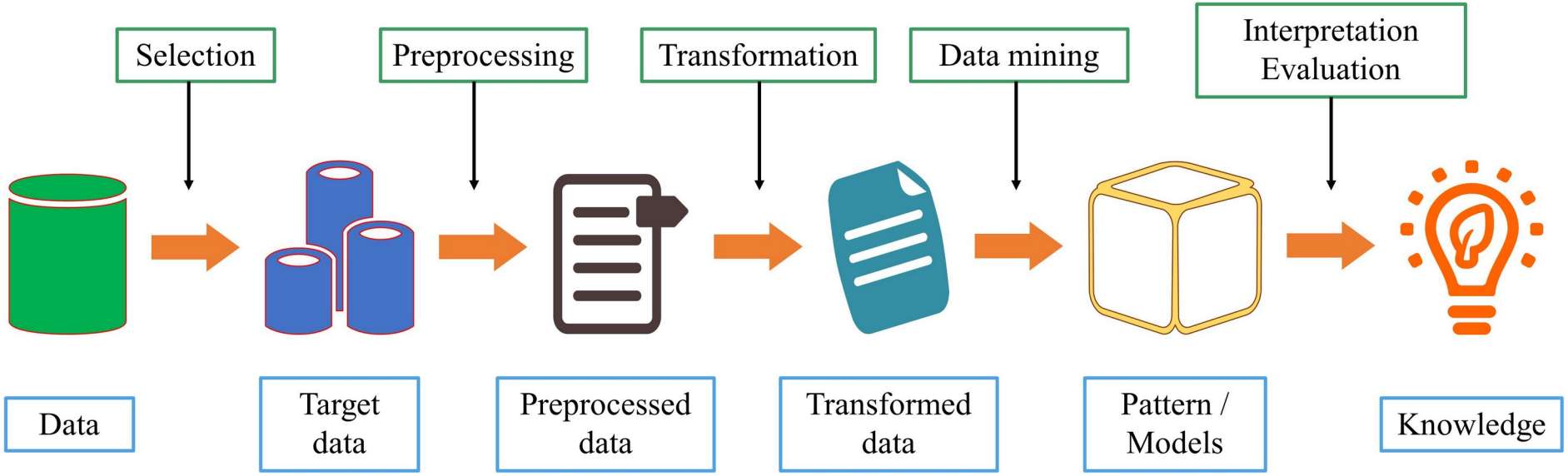
3. Машиналық оқыту әдістерінің биоинформатикада қолданылуы
3.1. Геномдық деректерді талдау

- Геномдық деректерді талдау биоинформатиканың негізгі бағыттарының бірі болып табылады. Геном секвенциясы бойынша алынған деректерді талдау көптеген биомедициналық зерттеулердің негізін құрайды. Бұл талдау жасушалардың құрылымын, гендердің қызметін және олардың аурулармен байланысын түсінуге мүмкіндік береді. Машиналық оқыту әдістері геномдық деректерді талдаудың тиімді құралы ретінде кеңінен қолданылады.
- **Машиналық оқыту әдістерінің геномдық деректерде қолданылу жолдары:**

1. **Гендер экспрессиясын болжау**
   Бақылаулы оқыту әдістері, мысалы, логистикалық регрессия және нейрондық желілер, гендердің экспрессиялық деңгейін болжамдау үшін пайдаланылады. Бұл әдістер биологиялық эксперименттер нәтижелеріне негізделіп, генетикалық деректер мен олардың белгілі бір фенотиптік белгілермен байланысын анықтауға мүмкіндік береді.



•**Геномдық секвенцияларды кластерлеу және классификациялау**
Бақылаусыз оқыту әдістері, мысалы, кластерлеу алгоритмдері, геномдық секвенциялардың үлкен көлемін тиімді топтастыруға мүмкіндік береді. Кластерлеу арқылы әртүрлі биологиялық түрлер арасындағы генетикалық ұқсастықтар мен айырмашылықтарды анықтауға болады.
•**CNPs (бір нуклеотидті полиморфизмдер) және генетикалық мутацияларды анықтау**
Машиналық оқыту әдістері биоинформатикада бір нуклеотидті полиморфизмдерді (SNPs) және генетикалық мутацияларды анықтауда қолданылады. Бұл мутациялар көптеген аурулармен байланысты болғандықтан, оларды уақытылы анықтау маңызды. Кездейсоқ ормандар және шешім ағаштары сияқты бақылаулы оқыту әдістері бұл мәселеде жоғары тиімділік көрсетеді.



- **Аурулармен байланысты гендерді анықтау**

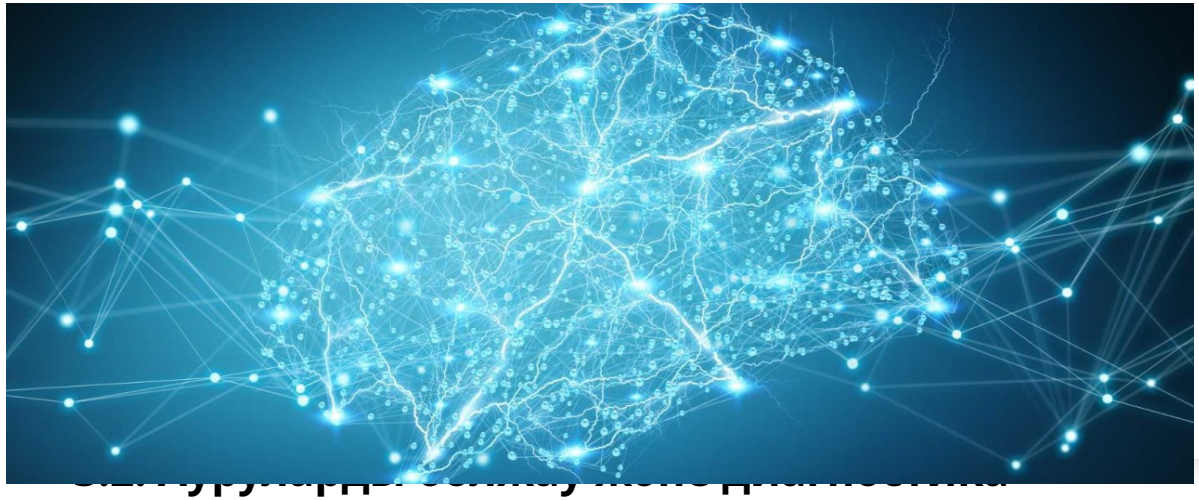
Машиналық оқыту әдістері геномдық деректердің аурулармен байланысын анықтауға көмектеседі. Нейрондық желілер және терең оқыту әдістері генетикалық мәліметтер негізінде белгілі бір аурулармен байланысты гендерді анықтау үшін қолданылады. Мысалы, бұл әдістер қатерлі ісік немесе нейродегенеративті аурулармен байланысты биомаркерлерді табуда тиімді.

- **Геномдық деректерді қысу және визуализациялау**

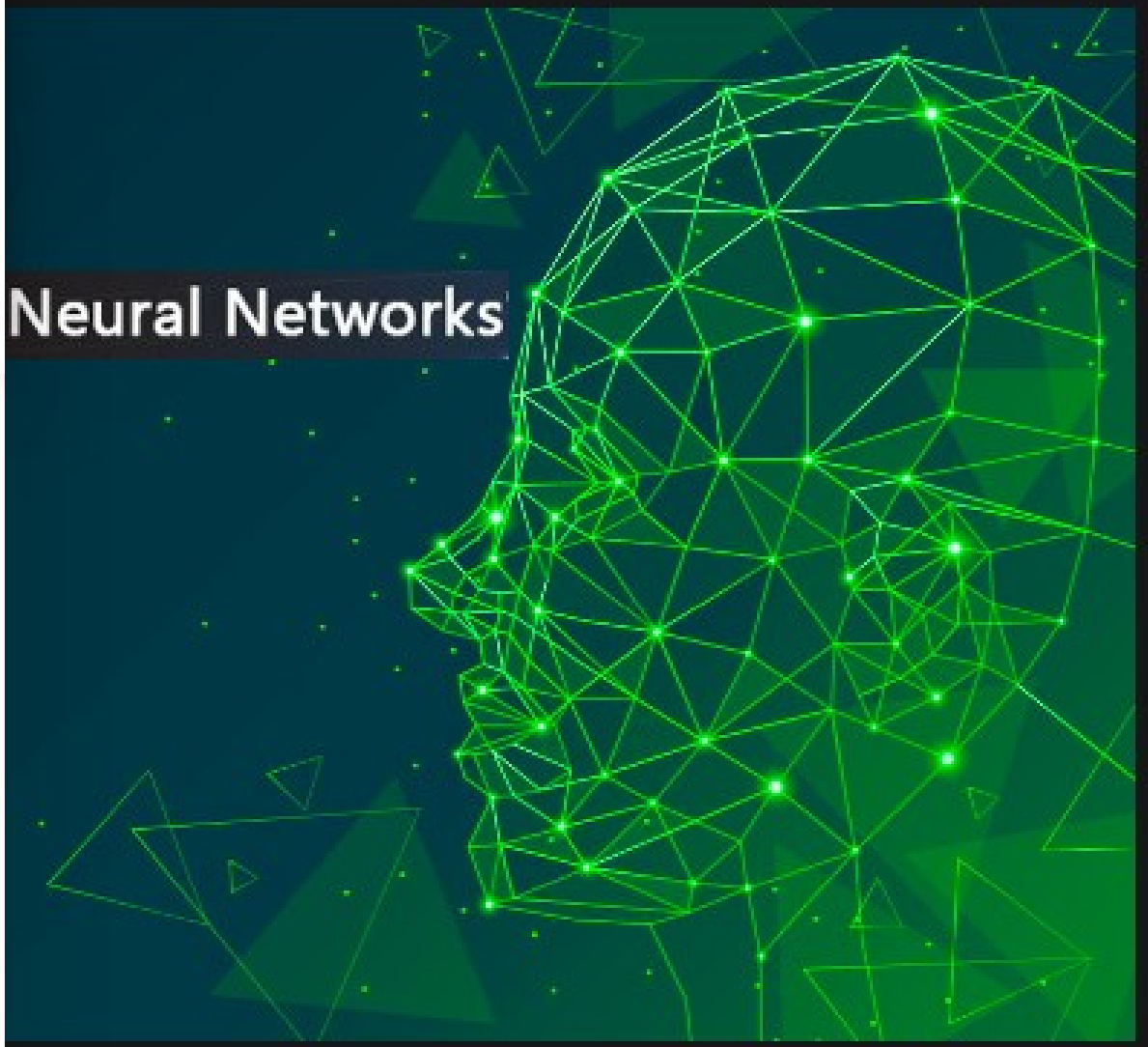
Машиналық оқытудың негізгі мақсаттарының бірі — үлкен көлемдегі деректерді қысу және олардың маңызды аспектілерін визуализациялау. Бұл әсіресе бақылаусыз оқыту әдістерінде, мысалы, бас компоненттер талдауында (PCA) немесе t-SNE алгоритмдерінде қолданылады. Бұл әдістер геномдық деректерді визуализациялау арқылы маңызды паттерндерді анықтауға мүмкіндік береді.

- **Эволюциялық талдау**

Эволюциялық биологияда машиналық оқыту әдістері эволюциялық ағаштар құрастыру және түрлер арасындағы генетикалық ұқсастықтар мен айырмашылықтарды анықтау үшін қолданылады. Бұл әдістер эволюциялық қатынастарды зерттеуде үлкен маңызға ие.



- Машиналық оқыту әдістері қазіргі таңда ауруларды болжау және диагностикалау саласында үлкен маңызға ие. Олардың көмегімен ауруларға бейімділікті бағалауға, диагноз қоюға және емдеу әдістерін жеке адамға бейімдеуге мүмкіндік бар. Әсіресе, генетикалық деректерді талдау арқылы аурулардың алдын алуға және емдеу стратегияларын жақсартуға ерекше назар аударылады.
- **Жасанды нейрондық желілер**
- Жасанды нейрондық желілер (Artificial Neural Networks, ANN) ауруларды болжау және диагностикалау саласында кеңінен қолданылады. Олар күрделі деректер арасындағы байланыстарды анықтауға, көптеген факторларды ескере отырып, аурулардың даму ықтималдығын болжауға қабілетті.



- **ДНҚ секвенциясын талдау**: Нейрондық желілер ДНҚ секвенциясынан алынған деректерді талдау үшін өте тиімді. Бұл әдістер генетикалық мутацияларды, гендердің экспрессиялық деңгейлерін және ауруларға бейімділікті анықтауға мүмкіндік береді. Мысалы, қатерлі ісік, Альцгеймер ауруы және жүрек-қан тамырлары аурулары сияқты аурулардың даму қаупін нейрондық желілер арқылы болжауға болады.

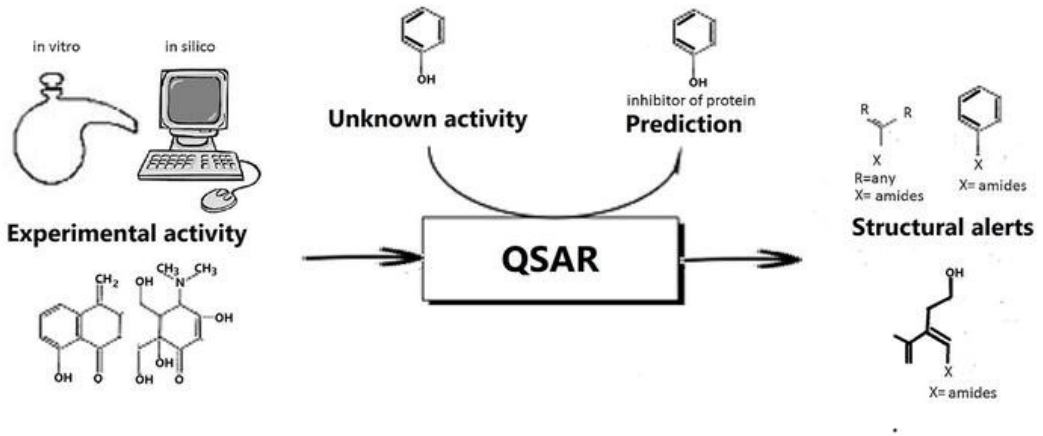
- **Генетикалық ақпарат негізіндегі болжау**: Нейрондық желілер ауруларды болжау кезінде көпқабатты құрылымды қолданады, бұл деректердің күрделі байланыстарын зерттеуге мүмкіндік береді. Оларды қолдану арқылы әртүрлі генетикалық факторлардың арасындағы өзара байланысты түсініп, жеке тұлғалардың ауруларға бейімділігін анықтауға болады.

SVM (Support Vector Machine)

- SVM (қолдау векторлары әдісі) — бұл сызықты немесе сызықты емес шекаралар арқылы деректерді екі немесе одан да көп кластарға бөлуге мүмкіндік беретін машиналық оқытудың бақылаулы әдісі. Биомедициналық деректерде бұл әдіс ауруды немесе денсаулықты классификациялауға жиі қолданылады.
- **Ауру мен денсаулықты классификациялау**: SVM әдісі ауру және сау деректер арасындағы айырмашылықтарды анықтау үшін қолданылады. Мысалы, бұл әдіс қатерлі ісік жасушалары мен сау жасушалар арасындағы айырмашылықтарды табуда тиімді. Ол әртүрлі аурулар мен денсаулық белгілері арасындағы шекараларды нақты анықтап, диагнозды автоматты түрде қоюға көмектеседі.
- **Көпөлшемді деректерді талдау**: SVM аурулардың диагностикасында көпөлшемді деректерді талдау үшін тиімді. Генетикалық деректер, медициналық бейнелер және басқа да биологиялық мәліметтерді талдау кезінде SVM денсаулық пен аурудың арасындағы шекараны анықтауға көмектеседі.

- **3.3. Ақуыздардың құрылымын болжау**
- Ақуыздардың үшөлшемді құрылымын болжау — биоинформатикадағы және молекулалық биологиядағы маңызды мәселелердің бірі. Ақуыздардың құрылымы олардың биологиялық функциясын анықтайды, сондықтан олардың үшөлшемді құрылымын дұрыс болжау ауруларды зерттеуде және жаңа дәрілерді әзірлеуде маңызды рөл атқарады. Машиналық оқыту, әсіресе нейрондық желілер мен терең оқыту (deep learning) әдістері, бұл мәселеде маңызды құралдар ретінде кеңінен қолданылады.
- **Ақуыздардың үшөлшемді құрылымын болжау әдістері:**

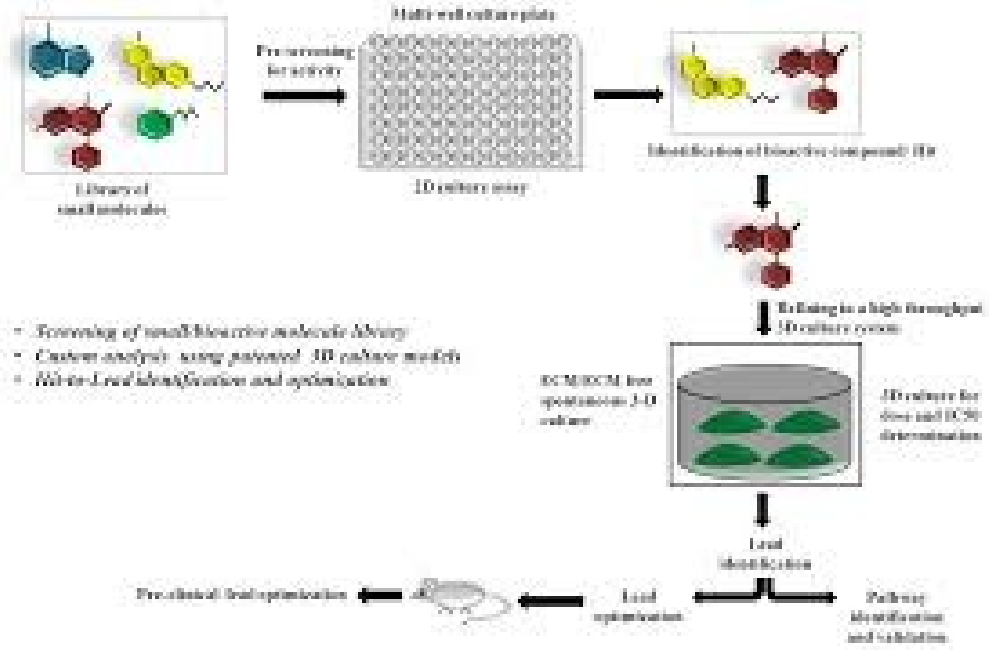
1. **Нейрондық желілер (Artificial Neural Networks, ANN)**
   Ақуыздардың құрылымын болжау кезінде нейрондық желілер ақуыз реттіліктерін және олардың кеңістіктік орналасуын зерттеуге көмектеседі. Ақуыздың аминқышқылдық тізбегін талдай отырып, нейрондық желілер ақуыздың қайталама, үшінші және төртінші деңгейдегі құрылымдарын болжай алады. Бұл әдіс ақуыздардың бүктелуін, байланыстарын және олардың функциясымен тікелей байланысты құрылымдық элементтерін анықтауға мүмкіндік береді.



Құрылым мен белсенділік арасындағы байланыс: QSAR моделін қолдана отырып, химиялық молекулалардың құрылымдық ерекшеліктері мен олардың биологиялық белсенділігі арасындағы корреляцияны анықтауға болады. Мысалы, белгілі бір молекулалық қасиеттер (гидрофобтылық, молекулалық масса, полярлы топтар және т.б.) дәрілік заттардың тиімділігіне қалай әсер ететінін түсінуге болады.

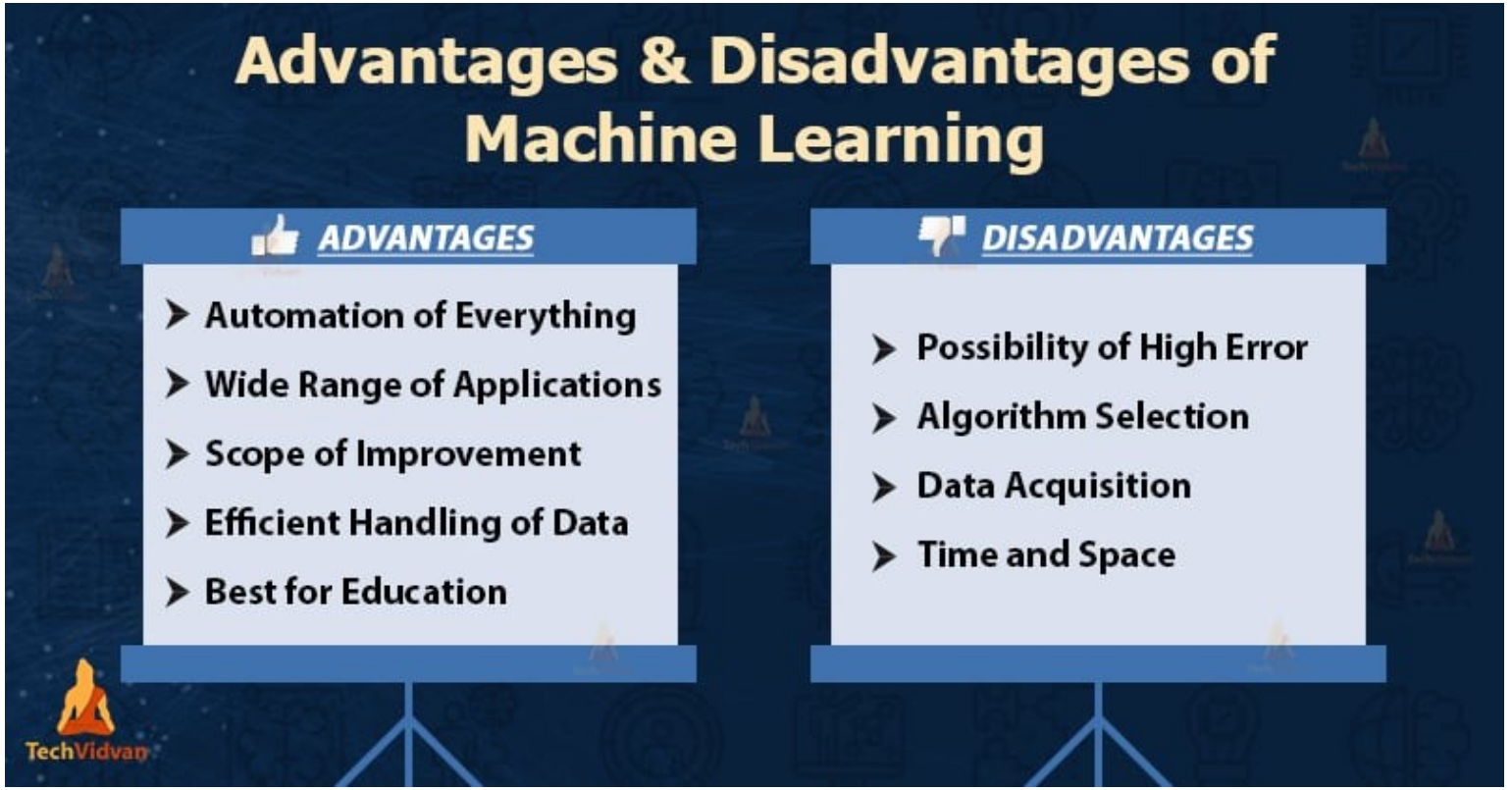
- **3.4. Дәрілік заттардың әсерін болжау**
- Дәрілік заттардың әсерін болжау — дәрі-дәрмектерді әзірлеу процесінің маңызды бөлігі. Бұл кезеңде молекулалардың биологиялық белсенділігін және олардың мақсатты жасушалармен немесе молекулалармен өзара әрекеттесуін болжау арқылы дәрілік заттардың тиімділігі мен қауіпсіздігін бағалау жүзеге асырылады. Осы мақсатта машиналық оқыту әдістері, соның ішінде QSAR (Quantitative Structure-Activity Relationship) моделі, кеңінен қолданылады.
- **QSAR (Quantitative Structure-Activity Relationship)**
- QSAR — дәрілік молекулалардың химиялық құрылымы мен олардың биологиялық белсенділігі арасындағы сандық қатынасты анықтайтын әдіс. Бұл әдіс арқылы белгілі бір химиялық қосылыстың құрылымына қарап, оның қандай биологиялық әсері болуы мүмкін екенін болжауға болады. QSAR моделін қолдану дәрілік молекулалардың жаңа түрлерін жасау

- **Молекулаларды скринингтен өткізу**: QSAR моделі жаңа дәрілік молекулаларды автоматты түрде скринингтен өткізуге мүмкіндік береді. Бұл әдіс арқылы биологиялық белсенділігі жоғары молекулаларды тез анықтауға және олардың әсерін болжауға болады.
- **Жаңа дәрілерді әзірлеу**: QSAR дәрілік заттардың ықтимал мақсатты белсенділігін болжауға көмектеседі, бұл жаңа молекулалар әзірлеу және олардың әсерін алдын ала бағалау кезінде маңызды рөл атқарады. Бұл модельдер молекулалық сипаттамалар негізінде жаңа дәрілердің тиімділігі мен уыттылығын бағалау үшін қолданылады.



- **QSAR моделінің негізгі артықшылықтары:**

1. **Дәрілік молекулалардың қасиеттерін алдын ала болжау**: QSAR дәрілік заттардың химиялық құрылымы мен олардың әсері арасындағы байланыстарды анықтауға көмектеседі, бұл молекулалардың тиімділігін алдын ала болжауға мүмкіндік береді.
2. **Эксперименттік шығындарды қысқарту**: QSAR моделін қолдану арқылы молекулаларды эксперименттік түрде сынамас бұрын, олардың әсерін болжауға болады, бұл эксперименттік зерттеулердің шығындарын азайтады.
3. **Молекулалардың қасиеттерін оңтайландыру**: QSAR арқылы химиялық қосылыстардың белсенділік деңгейін бағалай отырып, молекулалардың құрылымын өзгерту немесе оңтайландыру арқылы олардың тиімділігін арттыруға болады.



Машиналық оқытудың жетістіктері мен шектеулері

- **Жетістіктер**: Үлкен деректерді тез және тиімді өңдеу, жаңа биомаркерлер мен аурулар диагностикасын жақсарту.
- **Шектеулері**: Деректердің сапасы, түсіндірудің қиындығы, биологиялық мәнді нәтижелерді алу қиындықтары.

Қорытынды

Машиналық оқыту әдістері биоинформатика саласында жаңа мүмкіндіктерге жол ашып, үлкен деректерді тиімді талдауға, биологиялық процестерді болжауға және олардың арасындағы күрделі байланыстарды түсінуге көмектеседі.

Геномдық деректерді талдау, ауруларды болжау және диагностика, ақуыздардың үшөлшемді құрылымын болжау, дәрілік заттардың әсерін болжау сияқты маңызды биологиялық мәселелерді шешуде машиналық оқытудың рөлі артып келеді.
Жасанды нейрондық желілер, терең оқыту, SVM және QSAR сияқты әдістер биомедициналық деректермен жұмыс істегенде өте тиімді құралдар болып табылады. Олар ауруларға бейімділікті бағалау, генетикалық мутацияларды анықтау, ақуыз құрылымын болжау және дәрілік молекулалардың белсенділігін болжау сияқты міндеттерді дәл әрі жылдам орындауға мүмкіндік береді.

Машиналық оқытудың биоинформатикаға енуі ғылым мен медицинаның дамуына айтарлықтай үлес қосып, геномика, протеомика және дәрі-дәрмек жасау салаларында жаңа жетістіктерге жетуге жол ашты. Болашақта бұл әдістер биоинформатика саласындағы зерттеулердің негізгі құралдарының біріне айналып, адам денсаулығын жақсарту және ауруларды емдеу жолдарын жетілдіруге елеулі үлес қосатын болады.