MASHINALI O’QITISH JARAYONIDAGI XISOBLASHLARNI APACHE SPARK YORDAMIDA KAMAYTIRISH

Abdullayeva Umida Farhadovna
E-mail: umidafarhadovna11_07g@mail.ru

Received: 19.09.2025
Revised: 20.09.2025
Accepted: 21.09.2025

Kalit so‘zlar:
mashinali o‘qitish, Apache Spark, katta ma’lumotlar, taqsimlangan hisoblash, MLlib, hisoblash samaradorligi, ma’lumotlar tahlili

Annotatsiya. Ushbu maqolada mashinali o‘qitish (machine learning) jarayonida yuzaga keladigan yuqori hisoblash yuklamalarini kamaytirish masalasi ko‘rib chiqiladi. Ayniqsa, katta hajmdagi ma’lumotlar bilan ishlashda samarali hisoblash muhim ahamiyat kasb etadi. Tadqiqotda Apache Spark texnologiyasidan foydalanish orqali mashinali o‘qitish algoritmlarining samaradorligini oshirish va ishlov berish vaqtini qisqartirish yo‘llari tahlil qilinadi. Amaliy tajribalar Spark’ning taqsimlangan hisoblash imkoniyatlari katta ma’lumotlar ustida ishlashda sezilarli ustunlik berishini ko‘rsatadi. Shuningdek, maqolada Spark MLlib kutubxonasi yordamida turli modellarni qurish va ularning bajarilish tezligini baholash natijalari keltirilgan

Kirish. So‘nggi yillarda mashinali o‘qitish texnologiyalari juda keng qo‘llanila boshladi va ular turli sohalarda — tibbiyot, moliya, sanoat va boshqalarda muhim ahamiyat kasb etdi. Biroq mashinali o‘qitish jarayoni katta hajmdagi ma’lumotlar bilan ishlashni talab qilgani sababli, hisoblash resurslari va vaqt jihatidan katta yuklama hosil qiladi. Ayniqsa, katta ma’lumotlar (big data) asosida ishlaydigan modellarni yaratish va ularni optimallashtirish zamonaviy hisoblash vositalarining samaradorligini oshirishga ehtiyoj tug‘diradi. Apache Spark ochiq manbali katta ma’lumotlarni qayta ishlash mexanizmi bo‘lib, u mashinani o‘qitish vazifalari uchun ishlatilishi mumkin. Spark bir necha usulda mashinani o‘qitish hisoblashlarini kamaytirishga yordam beradi. Bu usullarga quyidagilar kiradi:

1.Taqsimlangan hisoblash: Spark mashinalar klasterida parallel ravishda hisob-kitoblarni amalga oshirishi mumkin. Bu bitta mashinada xisoblash uchun kunlar yoki haftalar ketishi mumkin bo‘lgan yirik mashina o‘rganish modellarini o‘qitishni sezilarli darajada tezlashtiradi.

2. Xotira ichidagi saqlash: Spark ma'lumotlarni xotirada saqlashi mumkin, bu esa diskdan ma'lumotlarni o'qish vaqtini qisqartiradi. Bu katta hajmdagi ma’lumotlar to‘plamini o‘z ichiga olgan mashina o‘rganish hisoblarini tezlashtirishga yordam beradi.

3. Quvurlarni o'tkazish operatsiyalari: Spark quvur liniyasi operatsiyalarini amalga oshirishi mumkin, ya'ni u ma'lumotlar mavjud bo'lishi bilanoq hisob-kitoblarni amalga oshirishi mumkin. Bu, ayniqsa, ma’lumotlar bir necha bosqichda qayta ishlanadigan mashina o‘rganish vazifalari uchun foydali bo‘lishi mumkin.

4. MLlib: Spark MLlib bilan birga keladi, mashina o‘rganish algoritmlari kutubxonasi, undan tashqarida foydalanish mumkin. Bu algoritmlar taqsimlangan hisoblashlar uchun optimallashtirilgan, shuning uchun ular katta ma’lumotlar to‘plamiga o‘tkazilishi va klasterda samarali ishlashi mumkin. Sparkdan foydalanish orqali biz mashinani o‘qitish xisoblashlarini qisqartirishingiz va kattaroq ma’lumotlar to‘plami va murakkabroq modellarni boshqarish uchun mashina o‘rganish vazifalaringizni kengaytirishingiz mumkin.



1- rasm. Apache Spark yordamida mashinali o’qitish jarayoni.

Biz ushbu dissertatsiya ishimizda Spark Mllib kutubxonasidan foydalandik. Spark MLlib kuchli va moslashuvchan mashinani o‘rganish kutubxonasi bo‘lib, u ma’lumotshunoslar va mashinani o‘qitish muhandislariga oddiy tasniflash va regressiya vazifalaridan tortib murakkab klasterlash va tavsiyalar tizimlarigacha bo‘lgan keng ko‘lamli muammolarni hal qilishda yordam beradi. Spark MLlib tasniflash, regressiya, klasterlash va hamkorlikda filtrlash uchun turli xil algoritmlarni hamda xususiyatlarni ajratib olish, o‘zgartirish va tanlash vositalarini taqdim etadi. MLlib kengaytirilishi mumkin va Spark klasteri bo‘ylab taqsimlangan katta ma’lumotlar to‘plamini boshqara oladi. Spark Mllib yoramida o’qitish uchun ma’lumotlar to’plamimizni taqsimlangan saqlash tizimlarida saqlab olamiz. Taqsimlangan saqlash tizimi sifatida HDFS yoki Spark RDD dan foydalanish mumkin. Biz Spark RDD dan foydalanib ma’lumotlarni saqlab oldik.

do = spark.sql("Select DO from df_sql")

do = do.rdd.map(lambda row : row.DO).collect()

ph = spark.sql("Select pH from df_sql")

ph = ph.rdd.map(lambda row : row.pH).collect()

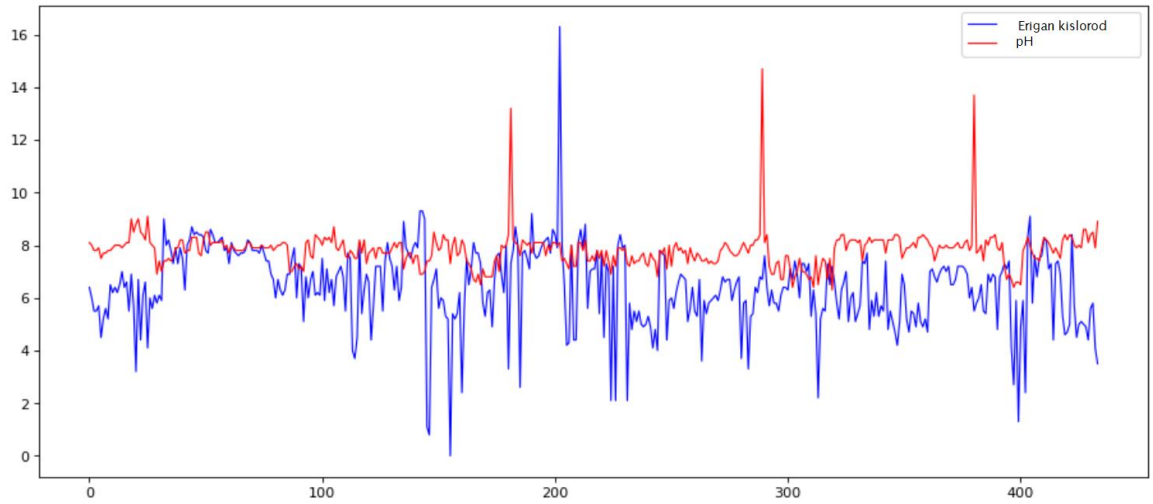
bod = spark.sql("Select BOD from df_sql")

bod = bod.rdd.map(lambda row : row.BOD).collect()

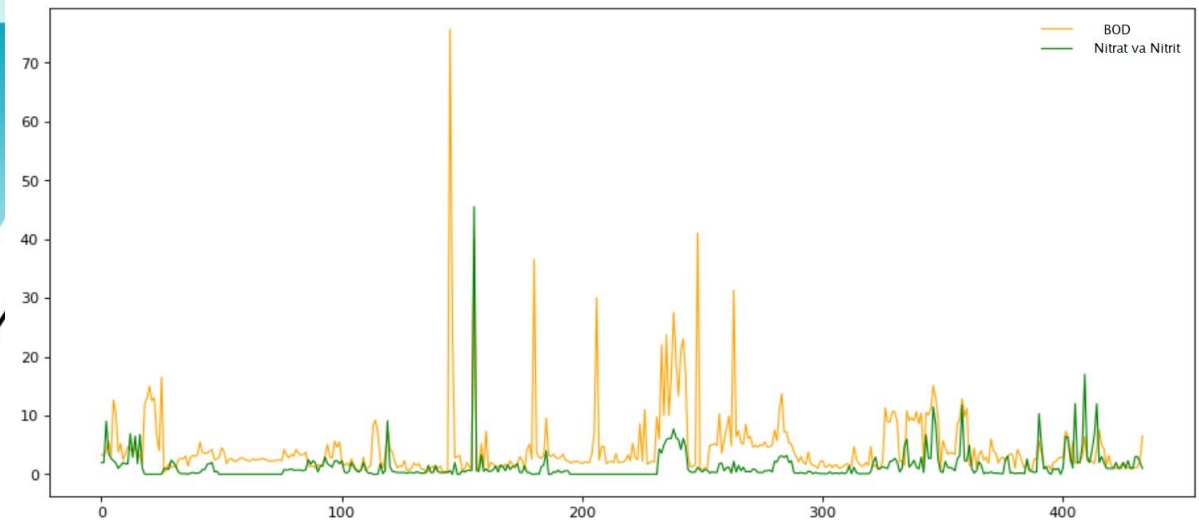
nn = spark.sql("Select NITRATE_N_NITRITE_N from df_sql")

nn = nn.rdd.map(lambda row : row.NITRATE_N_NITRITE_N).collect()

Saqlab olingan ma’lumotlarni tahlil qilish orqali quyidagi natijalarga ega bo’lishimiz mumkin. Quyidagi rasmda suvda erigan kislorod va pH miqdori orasidagi bog’lanishni ko’rish mumkin.



2 - rasm. Suvda erigan kislorod va pH miqdori orasidagi bog’lanish



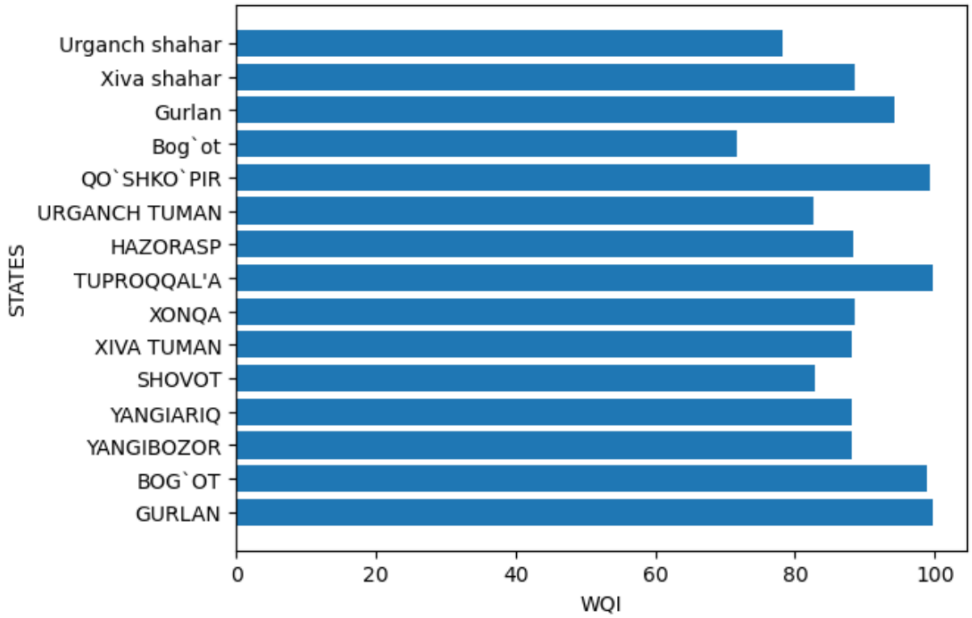
3- rasm. BOD va suvdagi Nitrat va Nitrit orasidagi bog’lanish

Suv sifatini baxolovchi asosiy mezon WQI (Water quality index) xisoblanadi. WQI Jahon sog'liqni saqlash tashkiloti (JSST) va O’zbekiston texnik jihatdan tartibga solish agentligi (O’zST) tomonidan tavsiya etilgan ichimlik suvi sifati standartlaridan foydalangan holda hisoblab chiqilgan. Suv sifati indeksi sifat reytingini chiziqli og'irlik bilan yig'ish yo'li bilan hisoblab chiqilgan,

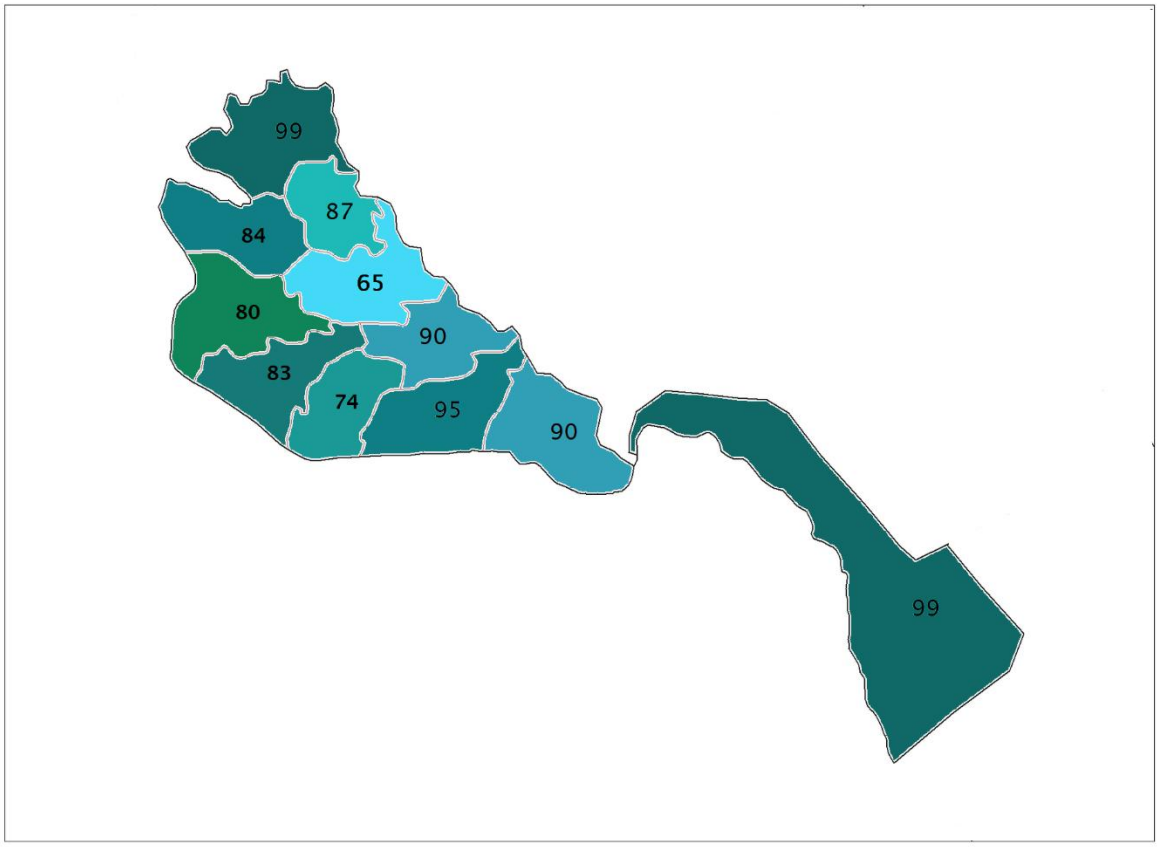
$$\text{WQI}=\sum(q_n \cdot W_n),$$

bunda q_n Suv sifatining n-parametri uchun sifat ko'rsatkichi, W_n n-chi parametr uchun birlik og'irligi. Q_n hisoblash uchun bizda standart formula mavjud bo'lsa-da, biz har bir parametr uchun sifat reytingini hisoblashning standart usulini qo'lladik. Holbuki W_n standart qiymat bo'lib, bu parametr WQI ga qanchalik ta'sir qilishini va belgilangan qiymatga ega ekanligini ko'rsatadi. Ushbu formula yordamida xisoblangan WQI ni Xorazm viloyatidagi har bir tuman kesimida

hisoblab chiqamiz.



4- rasm. Xorazm viloyati tumanlari kesimida xisoblangan WQI darajasi.



5- rasm. Xorazm viloyati tumanlari kesimida xisoblangan WQI ning xaritadagi ko’rinishi.

Xulosa. Ushbu tadqiqotda mashinali o‘qitish jarayonida yuzaga keladigan hisoblash yuklamalarini kamaytirish maqsadida Apache Spark texnologiyasidan foydalanish imkoniyatlari o‘rganildi. Taqsimlangan hisoblash platformasi sifatida Spark katta hajmdagi ma’lumotlarni samarali qayta ishlash va mashinali o‘qitish modellari ustida ishlash vaqtini sezilarli darajada qisqartirish imkonini beradi. Tajribalar ko‘rsatdiki, Spark MLlib kutubxonasi yordamida mashinali o‘qitish algoritmlarini taqsimlangan muhitda amalga oshirish hisoblash samaradorligini oshirishga yordam beradi. Shu bilan birga, Apache Spark texnologiyasi katta ma’lumotlar muhitida raqamli transformatsiyani tezlashtirish uchun qulay vosita ekanligi tasdiqlandi. Kelajakda ushbu platformaning imkoniyatlari yanada chuqurroq o‘rganilib, murakkab modellarda qo‘llanilishi va optimallashtirilishi istiqbollari mavjud.

FOYDALANILGAN ADABIYOTLAR

1. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets. *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, 10-10.
2. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Zaharia, M. (2016). MLlib: Machine Learning in Apache Spark. *Journal of Machine Learning Research*, 17(34), 1-7.
3. Gulli, A., & Pal, S. (2017). Deep Learning with Keras. *Packt Publishing*.
4. Russel, S., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach. *Pearson Education*.
5. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1), 107-113.