BIG DATADA DATA SCIENCE

Umarov Bekzod Azizovich
Farg‘ona davlat universiteti amaliy matematika va informatika kafedrasi o‘qituvchisi, ubaumarov@mail.ru
Omonaliyeva Elinur Umidbek qizi
Farg‘ona davlat unversiteti talabasi,
Gmail: omonaliyevaelinur@gmail.com

Annotatsiya: *Data science tashkilot ma’lumotlarida yashiringan nozik tushunchalarni ochish uchun matematika va statistika, ixtisoslashgan dasturlash, ilg‘or tahlil, sun’iy intellekt (SI) va mashinani o‘rganishni muayyan mavzu tajribasi bilan birlashtiradi. Ushbu tushunchalar qaror qabul qilish va strategik rejalashtirishni boshqarish uchun ishlatilishi mumkin.*
Kalit so‘zlar: *Data science, deduplikatatsiya, Data Scientists, biznes tahlilchi, NumPy, Pandas, Matplotlib, GitHub, Jupyter.*

Anotation: *Data science combines mathematics and statistics, specialized programming, advanced analytics, artificial intelligence (SI), and machine learning with specific subject expertise to uncover compelling insights hidden in an organization's data. These insights can be used to guide decision-making and strategic planning.*
Keywords: *Data science, dedduplicatation, Data Scientists, Business Analyst, NumPy, Pandas, Matplotlib, GitHub, Jupyter.*

Аннотация: *Наука о данных сочетает в себе математику и статистику, специализированное программирование, расширенную аналитику, искусственный интеллект (SI) и машинное обучение с экспертизой по конкретным предметам, чтобы раскрыть убедительные идеи, скрытые в данных организации. Эти аналитические данные могут быть использованы для принятия решений и стратегического планирования.*
Ключевые слова: *Data science, dedduplication, Data Science, Business Analyst, NumPy, Pandas, Matplotlib, GitHub, Jupyter.*

Ma’lumotlar manbalarining tezlashib borayotgan hajmi va keyinchalik ma’lumotlar, Data scienceni har bir sohada eng tez o‘sib borayotgan sohalardan biriga aylantirdi. Tashkilotlar ma’lumotlarni sharhlash va biznes natijalarini yaxshilash uchun amaliy tavsiyalar berish uchun ularga tobora ko‘proq ishonishadi.

Data sciencening hayot sikli turli ro‘llar, vositalar va jarayonlarni o‘z ichiga oladi, bu tahlilchilarga ta’sirchan tushunchalarni olish imkonini beradi. Odatda, Data science loyihasi quyidagi bosqichlardan o‘tadi:

Ma’lumotlarni qabul qilish:hayot sikli ma’lumotlarni to‘plash bilan boshlanadi - turli xil usullardan foydalangan holda barcha tegishli manbalardan xom tuzilgan va tuzilmagan ma’lumotlar. Ushbu usullar tizimlar va qurilmalardan qo‘lda kirish, veb-sayt va real vaqtda oqimli ma’lumotlarni o‘z ichiga olishi mumkin. Ma’lumot manbalari mijozlar ma’lumotlari kabi tuzilgan ma’lumotlarni va jurnal fayllari, video, audio, rasmlar, Internet (IoT), ijtimoiy media va boshqalar kabi tuzilmagan ma’lumotlarni o‘z ichiga olishi mumkin.

Ma’lumotlarni saqlash va ma’lumotlarni qayta ishlash:ma’lumotlar turli formatlar va tuzilmalarga ega bo‘lishi mumkinligi sababli, kompaniyalar qo‘lga olinishi kerak bo‘lgan ma’lumotlar turiga qarab turli xil saqlash tizimlarini ko‘rib chiqishlari kerak. Ma’lumotlarni boshqarish guruhlari tahlil, mashinani o‘rganish va chuqur o‘rganish modellari atrofida ish oqimlarini osonlashtiradigan ma’lumotlarni saqlash va tuzish bo‘yicha standartlarni belgilashga yordam beradi. Ushbu bosqich ma’lumotlarni tozalash, deduplikatatsiya qilish, o‘zgartirish va ETL(Extract(Olish), Transform(O‘zgartirish), Load(Yuklash)) ishlari yoki boshqa ma’lumotlarni integratsiya qilish texnologiyalaridan foydalangan holda ma’lumotlarni birlashtirishni o‘z ichiga oladi. Ushbu ma’lumotlarni tayyorlash ma’lumotlar omboriga, ma’lumotlar ko‘liga yoki boshqa omborga yuklashdan oldin ma’lumotlar sifatini oshirish uchun juda muhimdir.

Ma’lumotlarni tahlil qilish:bu yerda Data Scientists ma’lumotlar ichidagi qiymatlarning noto‘g‘ri qarashlari, naqshlari, diapazonlari va taqsimlanishini o‘rganish uchun kashfiyot ma’lumotlarini tahlil qilishadi. Ushbu ma’lumotlar tahlilini o‘rganish a / b sinovlari uchun gipoteza ishlab chiqarishga olib keladi. Bundan tashqari, tahlilchilarga ma’lumotlarning bashoratli tahlil, mashinani o‘rganish va / yoki chuqur o‘rganish uchun modellashtirish harakatlarida foydalanish uchun dolzarbligini aniqlashga imkon beradi. Modelning aniqligiga qarab, tashkilotlar biznes qarorlarini qabul qilish uchun ushbu tushunchalarga ishonishlari mumkin, bu esa ularga yanada miqyoslilikni oshirishga imkon beradi.

Muloqot qilish: Nihoyat, tushunchalar hisobotlar va boshqa ma’lumotlarni vizuallashtirish sifatida taqdim etiladi, bu esa biznes tahlilchilari va boshqa qaror qabul qiluvchilar uchun tushunchalarni va ularning biznesga ta’sirini osonlashtiradi. R yoki Python kabi ma’lumotlar fanlari dasturlash tili vizualizatsiyalarni yaratish uchun komponentlarni o‘z ichiga oladi. Shu bilan bir qatorda, Data Scientists maxsus vizuallashtirish vositalaridan foydalanishlari mumkin.

Ma’lumotlar fanlari intizom hisoblanadi, Data Scientists esa bu sohada amaliyotchilardir. Ma’lumotlar olimlari ma’lumotlar fanining hayot aylanishida ishtirok etadigan barcha jarayonlar uchun bevosita javobgar emaslar. Masalan, data pipelines odatda ma’lumot muhandislari tomonidan ko‘rib chiqiladi, ammo ma’lumotlar olimi qanday ma’lumotlar foydali yoki zarur ekanligi to‘g‘risida tavsiyalar berishi mumkin. Data Scientists mashinani o‘rganish modellarini yaratishi mumkin bo‘lsa-da, ushbu sa’y-harakatlarni yanada katta darajada kengaytirish dasturni tezroq ishlashini optimallashtirish uchun ko‘proq dasturiy ta’minot muhandisligi ko‘nikmalarini talab qiladi. Natijada, ma’lumotlar bo‘yicha mutaxassis mashinani o‘rganish modellarini kengaytirish uchun mashinani o‘rganish muhandislari bilan hamkorlik qilishi odatiy holdir.

Ma’lumotlar tadqiqotchisining mas’uliyati odatda ma’lumotlar tahlilchisi, xususan, kashfiyot ma’lumotlarini tahlil qilish va ma’lumotlarni vizuallashtirish bilan bog‘liq bo‘lishi mumkin. Biroq, ma’lumotlar bo‘yicha mutaxassisning mahorati odatda o‘rtacha ma’lumotlar tahlilchisidan kengroqdir. Nisbatan ma’lumot olimlari ko‘proq statistik xulosa chiqarish va ma’lumotlarni vizuallashtirish uchun R va Python kabi keng tarqalgan dasturlash tillaridan foydalanadilar.

Ushbu vazifalarni bajarish uchun Data Scientists odatiy biznes tahlilchisi yoki ma’lumotlar tahlilchisidan tashqari kompyuter fanlari va sof ilmiy ko‘nikmalarni talab qiladi. Ma’lumotlar bo‘yicha mutaxassis, shuningdek, avtomobil ishlab chiqarish, elektron tijorat yoki sog‘liqni saqlash kabi biznesning o‘ziga xos xususiyatlarini tushunishi kerak.

Muxtasar qilib aytganda, ma’lumotlar bo‘yicha mutaxassis quyidagilarga ega bo‘lishi kerak:

Tegishli savollar berish va biznesning og‘riqli nuqtalarini aniqlash uchun biznes haqida yetarli ma’lumotga ega bo'ling.

Ma’lumotlarni tahlil qilishda ishbilarmonlik bilan bir qatorda statistika va informatikani qo‘llang.

Ma’lumotlarni tayyorlash va chiqarish uchun keng ko‘lamli vositalar va texnikalardan foydalaning - ma’lumotlar bazalari va SQL dan ma’lumotlarni qazib olishgacha ma’lumotlarni integratsiya usullariga qadar.

Mashinani o‘rganish modellari, tabiiy tillarni qayta ishlash va chuqur o‘rganishni o‘z ichiga olgan bashoratli tahlil va sun’iy intellekt (SI) yordamida katta ma’lumotlardan tushunchalar oling.

Ma’lumotlarni qayta ishlash va hisoblashni avtomatlashtiradigan dasturlar yozing.

Natijalarning ma'nosini texnik tushunishning har bir darajasida qaror qabul qiluvchilar va manfaatdor tomonlarga aniq etkazadigan hikoyalarni aytib bering va tasvirlang.

Natijalarni biznes muammolarini hal qilishda qanday ishlatish mumkinligini tushuntiring.

Ma’lumotlar va biznes tahlilchilari, IT arxitektorlari, ma’lumotlar muhandislari va dasturlarni ishlab chiquvchilar kabi ma’lumotlar fanlari guruhining boshqa a’zolari bilan hamkorlik qiling.

Ushbu ko‘nikmalar yuqori talabga ega va natijada, ma’lumotlar fanlari karerasiga kirayotgan ko‘plab shaxslar, sertifikatlash dasturlari, ma’lumotlar fanlari kurslari va ta’lim muassasalari tomonidan taqdim etiladigan daraja dasturlari kabi turli xil ma’lumotlar fanlari dasturlarini o‘rganadilar.

"Data science" va "biznes intellekti" (BI) atamalarini aralashtirish oson bo‘lishi mumkin, chunki ularning ikkalasi ham tashkilotning ma’lumotlari va ushbu ma’lumotlarni tahlil qilish bilan bog‘liq, ammo ular yo‘nalishda farq qiladi.

Biznes intellekti (BI) odatda ma’lumotlarni tayyorlash, ma’lumotlarni yig‘ish, ma’lumotlarni boshqarish va ma’lumotlarni vizuallashtirishni ta’minlaydigan texnologiya uchun soyabon atamasidir. Biznes-razvedka vositalari va jarayonlari oxirgi foydalanuvchilarga xom ma’lumotlardan ta’sirchan ma’lumotlarni aniqlashga imkon beradi, bu esa turli sohalardagi tashkilotlarda ma’lumotlarga asoslangan qarorlar qabul qilishni

osonlashtiradi. Ma’lumotlar fanlari vositalari bu borada bir-biriga bog‘liq bo‘lsa-da, biznes intellekti ko‘proq o‘tmishdagi ma’lumotlarga e’tibor qaratadi va BI vositalaridan olingan tushunchalar tabiatda ko‘proq tavsiflovchidir. U harakat yo‘nalishini xabardor qilish uchun oldin nima bo‘lganini tushunish uchun ma’lumotlardan foydalanadi. BI odatda tuzilgan statik (o‘zgarmas) ma’lumotlarga qaratilgan. Ma’lumotlar fanlari tavsiflovchi ma’lumotlardan foydalansa-da, odatda uni bashoratli o‘zgaruvchilarni aniqlash uchun foydalanadi, keyinchalik ma’lumotlarni tasniflash yoki prognozlarni qilish uchun ishlatiladi.

Ma’lumotlar fanlari va BI o‘zaro istisno emas - raqamli jihatdan bilimdon tashkilotlar o‘z ma’lumotlarini to‘liq tushunish va ulardan qiymat olish uchun foydalanadilar.

Data Scientists kashfiyot ma’lumotlarini tahlil qilish va statistik regressiyani o‘tkazish uchun mashhur dasturlash tillariga tayanadi. Ushbu ochiq manbali vositalar oldindan tayyorlangan statistik modellashtirish, mashinani o‘rganish va grafik qobiliyatlarini qo‘llab-quvvatlaydi. Ushbu tillar quyidagilarni o‘z ichiga oladi:

R Studio: Statistik hisoblash va grafikalarni ishlab chiqish uchun ochiq kodli dasturlash tili va muhit.

Python: Bu dinamik va moslashuvchan dasturlash tili. Python ma’lumotlarni tezda tahlil qilish uchun NumPy, Pandas, Matplotlib kabi ko‘plab kutubxonalarni o‘z ichiga oladi.

Kod va boshqa ma’lumotlarni almashishni osonlashtirish uchun Data Scientists GitHub va Jupyter daftarlaridan foydalanishlari mumkin.

Ba’zi ma’lumotlar olimlari foydalanuvchi interfeysini afzal ko‘rishlari mumkin va statistik tahlil uchun ikkita keng tarqalgan korporativ vosita quyidagilarni o‘z ichiga oladi:

SAS:tahlil qilish, hisobot berish, ma’lumotlarni yig‘ish va prognozli modellashtirish uchun vizuallashtirish va interaktiv panellarni o‘z ichiga olgan keng qamrovli vositalar to‘plami.

IBM SPSS:ilg‘or statistik tahlil, mashinani o‘rganish algoritmlarining katta kutubxonasi, matn tahlili, ochiq manbali kengaytirilish, katta ma’lumotlar bilan integratsiya va ilovalarga uzluksiz joylashtirishni taklif etadi.

Data Scientists, shuningdek, Apache Spark, ochiq manbali Apache Hadoop ramkasi va NoSQL ma’lumotlar bazalari kabi katta ma’lumotlarni qayta ishlash platformalaridan foydalanishda malakaga ega bo'lishadi. Ular, shuningdek, biznes taqdimoti va elektron jadval ilovalari (Microsoft Excel kabi) bilan o‘rnatilgan oddiy grafik vositalari, Tableau va IBM Cognos kabi maqsadli tijorat vizuallashtirish vositalari va D3.js (interaktiv ma’lumotlarni vizuallashtirish uchun JavaScript kutubxonasi) va RAW Graphs kabi ochiq manbali vositalarni o‘z ichiga olgan ko‘plab ma’lumotlarni vizuallashtirish vositalariga ega. Mashinani o‘rganish modellarini yaratish uchun Data Scientists tez-tez PyTorch, TensorFlow, MXNet va Spark MLib kabi bir nechta ramkalarga murojaat qilishadi.

Data Scienceda keskin o‘rganish egri chizig‘ini hisobga olgan holda, ko‘plab kompaniyalar AI loyihalari uchun investitsiyalarning daromadini tezlashtirishga intilmoqda. Ular ko‘pincha Data Science loyihasining to‘liq potentsialini amalga oshirish uchun zarur bo‘lgan iste’dodni yollash uchun kurashadilar. Ushbu bo‘shliqni bartaraf etish uchun ular

ko‘p kishilik Data Science va mashinani o‘rganish (DSML) platformalariga murojaat qilmoqdalar va "fuqarolar ma’lumotlari bo‘yicha mutaxassis" ro‘lini keltirib chiqaradilar.

Multipersona DSML platformalari avtomatlashtirish, o‘z-o‘ziga xizmat ko‘rsatish portallari va past kodli / kodsiz foydalanuvchi interfeyslaridan foydalanadi, shunda raqamli texnologiyalar yoki ekspert ma’lumotlar fanlari bo‘yicha kam yoki hech qanday ma’lumotga ega bo‘lmagan odamlar Data Science va mashinani o‘rganishdan foydalanib, biznes qiymatini yaratishlari mumkin. Ushbu platformalar shuningdek, ko‘proq texnik interfeysni taklif qilish orqali ekspert Data Scientists qo‘llab-quvvatlaydi. Multipersona DSML platformasidan foydalanish korxona bo‘ylab hamkorlikni rag‘batlantiradi.

Data Science tez-tez katta ma’lumotlar to‘plamlaridan foydalanganligi sababli, ma’lumotlarning hajmiga mos keladigan vositalar, ayniqsa vaqtga sezgir loyihalar uchun juda muhimdir. Ma’lumotlar ko‘llari kabi bulutli saqlash yechimlari katta hajmdagi ma’lumotlarni osonlik bilan qabul qilish va qayta ishlash imkoniyatiga ega bo‘lgan saqlash infratuzilmasiga kirishni ta’minlaydi. Ushbu saqlash tizimlari oxirgi foydalanuvchilarga moslashuvchanlikni ta’minlaydi va kerak bo‘lganda katta klasterlarni aylantirishga imkon beradi. Shuningdek, ular ma’lumotlarni qayta ishlash ishlarini tezlashtirish uchun qo‘shimcha hisoblash tugunlarini qo‘shishlari mumkin, bu esa biznesga uzoq muddatli natija uchun qisqa muddatli savdolashuvlarni amalga oshirishga imkon beradi. Bulutli platformalar odatda oxirgi foydalanuvchilarning ehtiyojlarini qondirish uchun turli xil narxlash modellariga ega, masalan, har bir foydalanish yoki obuna, ular yirik korxona yoki kichik boshlang‘ich bo‘ladimi, saqlash va Data Science loyihalari uchun zarur bo‘lgan boshqa vositalarga kirishni ta’minlaydi.

Ochiq manba texnologiyalari Data Science asboblar to‘plamida keng qo‘llaniladi. Ular bulutda joylashtirilganda, jamoalar ularni mahalliy ravishda o‘rnatish, sozlash, texnik xizmat ko‘rsatish yoki yangilashlari shart emas. Bir nechta bulut provayderlari, shu jumladan IBM Cloud, shuningdek, Data Scientistsga kodlamasdan modellarni yaratishga imkon beradigan oldindan paketlangan asboblar to‘plamlarini taklif etadi, bu esa texnologiya yangiliklari va ma’lumotlarni tushunishga kirishni yanada demokratlashtiradi.

Korxonalar ma’lumotlar fanidan ko‘plab foyda olishlari mumkin. Umumiy foydalanish holatlari orasida aqlli avtomatlashtirish orqali jarayonlarni optimallashtirish va mijozlar tajribasini (CX) yaxshilash uchun kengaytirilgan maqsadlashtirish va shaxsiylashtirish kiradi. Biroq, yanada aniq misollar quyidagilarni o‘z ichiga oladi:

Bu yerda Data Science va sun’iy intellekt uchun bir nechta vakillik holatlari mavjud:

Xalqaro bank mashinani o'rganishga asoslangan kredit xavfi modellari va kuchli va xavfsiz gibrid bulutli hisoblash arxitekturasidan foydalangan holda mobil ilova yordamida tezroq kredit xizmatlarini taqdim etadi.

Elektronika firmasi ertangi haydovchisiz transport vositalarini boshqarish uchun ultra kuchli 3D bosilgan sensorlarni ishlab chiqmoqda. Yechim real vaqtda ob'ektlarni aniqlash qobiliyatini oshirish uchun ma'lumotlar fanlari va tahlil vositalariga tayanadi.

Robotik jarayonlarni avtomatlashtirish (RPA) echim provayderi o'z mijoz kompaniyalari uchun hodisalarni boshqarish vaqtini 15% dan 95% gacha kamaytiradigan kognitiv biznes

jarayonlarini qazib olish echimini ishlab chiqdi. Yechim mijozlarning elektron pochta xabarlarining mazmuni va kayfiyatini tushunishga o'rgatilgan bo'lib, xizmat ko'rsatish guruhlarini eng dolzarb va shoshilinch bo'lganlarni birinchi o'ringa qo'yishga yo'naltiradi.

Raqamli media texnologiyalari kompaniyasi o‘z mijozlariga tobora ko‘payib borayotgan raqamli kanallar taklif etilayotganda televizion tomoshabinlarni jalb qiladigan narsalarni ko‘rishga imkon beradigan auditoriya tahlil platformasini yaratdi. Ushbu echim tomoshabinlarning xatti-harakatlari to'g'risida real vaqtda tushuncha to'plash uchun chuqur tahlil va mashinani o'rganishdan foydalanadi.

Shahar politsiyasi bo‘limi zobitlarga jinoyatchilikning oldini olish uchun resurslarni qachon va qayerga joylashtirishni tushunishga yordam berish uchun statistik hodisalarni tahlil qilish vositalarini yaratdi (havola ibm.com tashqarida joylashgan). Ma'lumotlarga asoslangan yechim dala xodimlari uchun vaziyatli xabardorlikni oshirish uchun hisobotlar va asboblar panellarini yaratadi.

Shanxay Changjiang Fan va Texnologiya Rivojlanishi IBM, Watson texnologiyasidan foydalanib, bemorlarni insultni boshdan kechirish xavfiga qarab tasniflash uchun mavjud tibbiy yozuvlarni tahlil qila oladigan va turli xil davolash rejalarining muvaffaqiyat darajasini bashorat qila oladigan AI asosidagi tibbiy baholash platformasini yaratish.

Natija: Ma’lumotlar fanlari (Data Science) sohasining ahamiyati va o‘sishi tashkilotlarga ma’lumotlar orqali qaror qabul qilish jarayonini o‘zgartirish imkonini yaratdi. Data Scientists, statistik tahlil va mashinani o‘rganish yordamida biznes muammolarini yechish uchun kuchli vositalarga ega. Ma’lumotlarni to‘plash, saqlash, qayta ishlash va tahlil qilish jarayonlarida zamonaviy texnologiyalar va platformalar, jumladan, bulutli hisoblash va katta ma’lumotlar tizimlari qo‘llaniladi. Ular yuqori samarali va tezkor natijalar olishga yordam beradi. Shuningdek, Data Science va biznes intellekti (BI) o‘rtasidagi farqni tushunish va ularni to‘g‘ri qo‘llash tashkilotlarga o‘z biznes jarayonlarini optimallashtirishda va ma’lumotlardan strategik qarorlar qabul qilishda muhim ro‘l o‘ynaydi. Tashkilotlar uchun bu sohada malakali mutaxassislarni tayyorlash va mos keluvchi platformalar va vositalarni qo‘llash, raqobatbardoshlikni oshirishda muhim ahamiyatga ega.

Xulosa: Ma’lumotlar fanlari (Data Science) sohasidagi yuksalish, ayniqsa, ma’lumotlar hajmining tezlashib borayotganligi va ularning biznesdagi muhim ahamiyatini tushunishdan kelib chiqadi. Tashkilotlar ma’lumotlarni tahlil qilish va biznes natijalarini yaxshilashda Data Science ga murojaat qilmoqda. Bu jarayon ma’lumotlarni to‘plash, saqlash, qayta ishlash, tahlil qilish va vizualizatsiya qilish kabi bosqichlarni o‘z ichiga oladi. Data Scientistlar, statistik tahlil, bashoratli modellar va mashinani o‘rganish yordamida tashkilotlarga qiymat yaratishda yordam beradi. Ular texnologik vositalardan keng foydalanadilar, masalan, R, Python, SQL, Apache Spark, TensorFlow va boshqalar. Bundan tashqari, bulutli saqlash va katta ma’lumotlar platformalari Data Science ishlarini tezlashtirish va samarali amalga oshirishda muhim ro‘l o‘ynaydi. Ma’lumotlar fanlari va biznes intellekti (BI) o‘rtasidagi farqni tushunish ham zarur, chunki BI o‘tmishdagi

ma’lumotlarni tahlil qilishga qaratilgan bo‘lsa, Data Science ko‘proq bashoratli tahlil va mashinani o‘rganish asosida ishlaydi.

FOYDALANILGAN ADABIYOTLAR:

1. "Python for Data Analysis" by Wes McKinney
2. "Data Science for Business" by Foster Provost and Tom Fawcett
3. "Big Data: A Revolution That Will Transform How We Live, Work, and Think" by Viktor Mayer-Schönberger and Kenneth Cukier
4. "Data Science from Scratch: First Principles with Python" by Joel Grus
5. "The Big Data-Driven Business: How to Use Big Data to Win Customers, Beat Competitors, and Boost Profits" by Russell Glass and Sean Callahan
6. "Artificial Intelligence: A Modern Approach" by Stuart Russell and Peter Norvig

FOYDALANILGAN SAYTLAR

https://www.ibm.com/topics/data-science
https://www.kaggle.com
https://www.datacamp.com