Projektni zadatak – Veliki podaci

Projektni zadatak radi se u Databricks. Projekat obuhvata pripremu datasetova i smeštanje podataka na MongoDB Atlas, učitavanje podataka sa Atlasa u Databricks, pripremu i obradu podataka statički i kroz streaming kako bi se podaci ukrstili i izdvojili izveštaji, i analizu dobijenih izveštaja. Podatke treba prikazivati tabelarno i kroz prigodne grafike. Notebook treba detaljno dokumentovati i objasniti primenjene operacije i dobijene rezultate. Dodatne korake rađene van Databricks možete opisati u posebnom dokumentu ili u okviru notebook-a. Izbor jezika i biblioteka je na studentu.

Projektni zadatak nosi 25 bodova. Ima dva dela.

Prvi deo – priprema datasetova 7 bodova (2 ako se podaci unose van databricks)

Pripremiti datasetove, učitati u Databricks, analizirati, formirati šemu dokumenta i pripremljene podatke kao dokumente ubaciti u MongoDB na Atlasu. Formirati najmanje dve kolekcije dokumenata unutar jedne baze.

Drugi deo – obrada podataka 18 bodova (definisano u temama)

Učitati pripremljeni dataset iz monga u notebook. Izvesti transformacije podataka i napraviti izveštaje.
Podatke prikazati i grafički, upotrebom odgovarajućih chartova.
Obraditi podatke kroz structured streaming i prikazati rezultate grafički.

Predaje se:

- notebooks za prvi i drugi deo, eksportovano u dbc i html formatu, nakon izvršavanja čvorova
- linkovi ka datasetovima koji su upotrebljeni
- konekcija ka Atlas instanci
- šeme mongo kolekcija (može u okviru prvog notebooka)
- opis postupka rada i propratni kod

Opisi mogu biti unutar notebook-a ili u posebnom fajlu. Za prvi deo notebook nije obavezan, ako podatke pripremate i unosite van databricks. Prateća dokumentacija se predaje u pdf formatu, sa uobičajenim formatiranjima (A4, 12pt, numerisane slike, ime studenta na početku... – formatiranje se ne ocenjuje posebno, ali budite elementarno uredni).

Projekat se može raditi **samostalno** ili u grupama do **3 člana**. Potrebno je prijaviti temu i tim do 20.12. mejlom. Grupni projekat se prijavljuje tako što jedan član tima šalje mejl asistentu i ostalim članovima tima. U mejlu navesti članove i temu, subject = „Projekat VP“, i hipoteze koje nameravate da testirate, ukoliko nisu medju ponudjenima. Projekte treba predati do 2. januara. Odbrana projekata bice online, u poslednjem terminu vezbi, u prvoj nedelji januara.

Dato je tri teme. Tema 1 je za samostalni rad. Teme 2 i 3 su za grupe.

Dozvoljavam izborne teme, gde možete da primenite ML i druge koncepte, ako ste ih ranije radili. Ove teme treba da najavite do 18.12. kako bismo ih potvrdili ili korigovali tako da budu uporedive složenosti kao zadate teme. U mejlu navedite članove tima, subject = „Predlog projekta VP“. Sagledajte koncept ponuđenih tema i u skladu sa tim predložite temu.

Uz teme su navedeni datasetovi koje možete da koristite, ali niste obavezni - možete da koristite i druge datasetove ako nađete pogodnije, bogatije podatke.

Tema 1 – COVID19 (individualno)

Analiziranjem podataka primetiti da li postoji povezanost aktivnosti na twitter-u sa tokom epidemije.

Primeri hipoteza:

1. zastupljenost određenih poruka („feeling bad, could be #covid“ i sl.) je u korelaciji sa brojem potvrđenih slučajeva/smrtnošću. Trendovi imaju otklon od nekoliko dana: skok u zastupljenosti poruka prethodi 2-3 dana istom takvom skoku u broju potvrđenih slučajeva, i 7-10 dana skoku u broju smrtnih slučajeva. Uporednim predstavljanjem podataka potvrditi ili odbaciti hipotezu.
2. brzina skoka zaražavanja je u korelaciji sa stopom smrtnosti. Da li postoji bitno veća stopa smrtnosti u državama nakon perioda gde je broj potvrđenih slučajeva naglo skočio? (jer npr. bolnički sistem ne može da pruži kvalitetan tretman zbog preopterećenosti)

Definisati jednu hipotezu i analizom podataka je potvrditi ili osporiti.

Uvidom u dostupne podatke izvršiti analize kao što je stopa smrtnosti u odnosu na broj potvrđenih slučajeva, u odnosu na opštu populaciju i dr. Rezultate pokazati kroz prigodne chartove na ukupnom i izabranim segmentima podataka. Napraviti najmanje 3 charta.

Izvesti structured streaming dnevnih podataka i prikazati jedan takav chart.

Fokusirati se na 2 države po izboru.

Datasetovi koji se mogu koristiti:

https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_global_deaths.csv
https://www.kaggle.com/gpreda/covid19-tweets
https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april?select=2020-04-30+Coronavirus+Tweets.CSV

Bodovna raspodela:

Hipoteza	6
Analize i chartovi	6
Structured streaming	6

Tema 2 – (ne)Bezbednost u saobraćaju (grupni)

Analiziranjem podataka o saobraćajnim nezgodama identifikovati faktore rizika.

Primeri hipoteza:

1. Saobraćajne nesreće sa materijalnom štetom se češće dešavaju posle kiše.
2. Saobraćajne nesreće sa povređenima se češće dešavaju između 8.30-9.30 ujutru radnim danima.
3. Saobraćajne nesreće sa žrtvama se češće dešavaju petkom i subotom pred zoru.
4. Učestalost saobraćajnih nezgoda sa materijalnom štetom je veća u periodu punog meseca.

Definisati tri hipoteze i analizom podataka ih potvrditi ili osporiti.

Napraviti heatmap ili na drugi prigodan način na mapi prikazati statistiku nezgoda i identifikovati rizične zone uopšte i u izabranim vremenskim periodima (ujutru, vikend, pun mesec...). Možete koristiti eksterne biblioteke za generisanje mapa.

Izvesti structured streaming dnevnih podataka i prikazati jedan takav chart.

Meteo podaci za nekoliko godina unazad sa ovih izvora se moraju scrape-ovati – uz malo javascript-a to nije problem. Slobodni ste da scraping uradite po svom izboru. Skriptove i proces scrape-ovanja treba dokumentovati. Ako se radi ručno, u kratkim crtama opisati kako su podaci pripremljeni. Mesečeve mene se mogu računati i formulom, u tom slučaju u Notebook-u testirajte da li je pun mesec 28.12.2012. u 11:21 (GMT+1).

Možete se bazirati na jednoj ili više godina, na letnjim periodima i sl.

Datasetovi koji se mogu koristiti:

https://data.gov.rs/en/datasets/podatsi-o-saobratshajnim-nezgodama-za-teritoriju-grada-beograda/
https://data.gov.rs/sr/datasets/podatsi-o-saobratshajnim-nezgodama-po-politsijskim-upravama-i-opshtinama/
http://www.meteologos.rs/beograd-godisnji-mesecni-i-dnevni-meteoroloski-podaci-1750-2019-1848-2019/
https://www.timeanddate.com/weather/serbia/belgrade/historic?month=1&year=2020
https://www.accuweather.com/sr/rs/belgrade/298198/february-weather/298198
https://www.accuweather.com/sr/rs/belgrade/298198/daily-weather-forecast/298198
https://tidesandcurrents.noaa.gov/moon_phases.shtml?year=2020&data_type=monFeb

Bodovna raspodela:

3 hipoteze	6
Mapa	6
Structured streaming	6

Tema 3 – Fudbal

„Ko se ne razume u fudbal ulaže u berzu.“ Pokazati da li se klađenju može pristupiti bez poznavanja konteksta igre. Analiziranjem rezultata utakmica, kvota i drugih podataka o utakmicama identifikovati strategije klađenja koje rade u korist kladioničara.

Hipoteza: Klađenje na sistem se ne isplati.
Izabrati tri sistema (algoritma formiranja tiketa), implementirati potrebne funkcije i proveriti da li se i koliko izabrani sistemi isplate. Testirati na ukupnom datasetu i na izabranim ligama. Sistem može biti u dnevnim, nedeljnim, dvonedeljnim ... vremenskim slotovima.

Analizirati predvidivost liga, po meri entropije rezultata utakmica timova, odstupanja od predikcije kvote (manja kvota = veća verovatnoća ishoda), ili dr. Prikazati predvidivost liga i druge međurezultate grafički, gde je primenjivo.
Proveriti da li se predvidivije lige više isplate.

Izvesti structured streaming simulacije primene sistema kroz vreme, tokom sezone, i pokazati kretanje balansa računa (koliko novca se u tom trenutku ima dobijeno, ili u minusu) na chartu, trenutke klađenja...

Izdvojiti ukupnu zaradu na kraju sezone, maksimalnu postignutu sumu novca, minimalnu postignutu sumu novca (koliko se najviše otišlo u minus), uporedi prikazati za izabrane sisteme. Komentarisati proces obrade i dobijene rezultate.

Datasetovi koji se mogu koristiti:

https://www.football-data.co.uk/data.php
https://www.kaggle.com/hugomathien/soccer
https://www.kaggle.com/yonilev/the-most-predictable-league
https://kladjenje.rs/sportsko-kladenje/sistemi-za-kladjenje-na-fudbal/
https://help.mozzartbet.com/sta-su-to-sistemi/

Bodovna raspodela:

Sistemi	6
Predvidivost liga i provera isplativosti	6
Structured streaming	6