3 laboratorinio užduotis

Nagrinėsime duomenis, aprašančius siuntų pristatymo bei logistikos įmonės veiklą. Duotajame faile pateikti dalinai struktūrizuoti duomenys apie siuntų išvežiotojų sustojimus, kurie gali būti apibūdinami sekančiais parametrais: "marsrutas","sustojimo data","sandelio id","Firma","Marsruto tipas","Masinos tipas","sustojimo tipas","sustojimo savaites diena","laikas","Sustojimo numeris","siuntu skaicius","svoris","svorio grupe","geografine zona","pasto kodas", "Aptarnavimo grupe","tipas","Laikas iki sustojimo","Laikas po sustojimo", "Uzkrovimo tipas","Ar reikalingos paletes","Laukia","Sustojimo klientu skaicius","sustojimo klientu sarasas","kaina procentas","kaina vienetais".

Papildomai yra duomenys apie maršrutus skirtingomis savaites dienomis (t.y. duomenys sugrupuoti pagal šių parametrų porą: "marsrutas", "sustojimo data") su sekančiais parametrais: "marsrutas", "sustojimo data","M", "BendrasAtstumas","BendrasSvoris","BendrasLaikas","BendraKaina".

Naudodami Apache Spark ir DataFrame duomenų struktūrą, atlikite vieną iš sekančių užduočių:

1. Ištirkite tiesinę priklausomybę parametro "BendraKaina" nuo parametro "siuntu skaicius" (agreguodami pagal maršrutą ir datą, pritaikykite sumos operaciją), kai nagrinėjami duomenys tik su viena ta pačia reikšme "Masinos tipas". Analizės metu pritaikykite tiesinę regresiją.

2. Ištirkite tiesinę priklausomybę parametro "BendraKaina" nuo parametro "svoris" (agreguodami pagal maršrutą ir datą, pritaikykite sumos operaciją), kai nagrinėjami duomenys tik su viena ta pačia reikšme "Masinos tipas". Analizės metu pritaikykite tiesinę regresiją.

3. Ištirkite tiesinę priklausomybę parametro "BendrasLaikas" nuo parametro "siuntu skaicius" (agreguodami pagal maršrutą ir datą, pritaikykite sumos operaciją), kai nagrinėjami duomenys tik su viena ta pačia reikšme "Masinos tipas". Analizės metu pritaikykite tiesinę regresiją.

4. Ištirkite tiesinę priklausomybę parametro "BendrasLaikas" nuo parametro "svoris" (agreguodami pagal maršrutą ir datą, pritaikykite sumos operaciją), kai nagrinėjami duomenys tik su viena ta pačia reikšme "Masinos tipas". Analizės metu pritaikykite tiesinę regresiją.

Užduoties numerį paskaičiuokite pagal formulę $((n-1) \mod 4) + 1$, kur n – Jūsų numeris grupės sąraše.

Sudarykite darbo ataskaitą. Ataskaitoje aprašykite užduotį pagal savo variantą; pateikite sprendimus pseudokodo pavidalu, pekomentuokite gautą rezultatą.

- Spark ML tiesinės regresijos aprašymą galima rasti oficialioje dokumentacijoje: https://spark.apache.org/docs/latest/ml-classification-regression.htm regression