Titanic

SUBTASK 1

„Analyze given dataset (performing basic stats, missing values analysis, detecting outliers, etc.). Document your discoveries.“

Za uređivanje dataset-a koristimo bibloteku *Pandas*. Postoje neke osnovne tehnike kako popuniti vrijednosti koje nedostaju u dataset-u. Prije nego što to navedem, htio bih da napišem ideju kako bih ja to uradio (zamisao prije istraživanja).

Kada sam otvorio dataset, vidio sam nazive kolona: Name, PClass, Age, Sex, Survived. Razmišljajući kako da popunim redove u kojima nedostaje element "Age" imao sam dvije opcije – ili preskočiti taj red u analizi u kojoj koristim godine putnika, ili pronaći srednju vrijednost svih putnika. Treća opcija bi bila na osnovu radnog mjesta na brodu, ali nemamo informaciju o tome. Na primjer, ukoliko bi pozicija radnika bila *konobar*, vjerovatno osoba ne bi bila mlađa od nekih 20 godina, niti starija od 40. Onda bi se mogla uzeti opcija srednje vrijednosti svih konobara ili srednja vrijednost godina svih putnika na brodu između 20 i 40 godina.

Ideja je da pretražim svaku kolonu i provjerim da li se u njoj nalazi ijedan element čija je vrijednost NA. Rezultati prvog koda prikazani su na slici 1.

```
Broj putnika sa nepoznatim imenom:  0
Broj putnika sa nepoznatom putničkom klasom:  0
Broj putnika sa nepoznatim godinama:  553
Broj putnika sa nepoznatim spolom:  0
Broj putnika sa nepoznatim ishodom (prezivjeli ili ne):  0
```

Slika 1 – Rezultat analize dataset-a

Nakon ove provjere, smatram da bi dobro bilo pronaći srednju vrijednost godina svih putnika na brodu, a potom srednju vrijednost putnika za svaku od klasa. Na osnovu toga umetnuti srednju vrijednost godina na osnovu klase, ili na osnovu ukupne srednje vrijednosti godina. Prije toga provjeriti postoji li neka druga klasa mimo 1^{st}, 2^{nd} I 3^{rd}.

PITANJE: Šta ukoliko nemam podatke koji su tipa "string"?

Outlier je neki podatak koji se drastično razlikuje od ostalih u dataset-u. Moj intuitivan način kojim bi detektovao outliere bi bio da sortiram po veličini (ukoliko je riječ o brojevima) i izračunam razliku između svaka susjedna elementa liste. Ukoliko je ta razlika veća od nekog praga, element proglašavam outlierom i ne uzimam ga u razmatranje.

PITANJE: Kako pronaći outliere za tipove koji nisu numerički?

Missing values links:

https://machinelearningmastery.com/handle-missing-data-python/

https://medium.com/ibm-data-science-experience/analyze-open-data-sets-using-pandas-in-a-python-notebook-64e93776370a

https://datascience.stackexchange.com/questions/22740/missing-values-in-data

http://www.dummies.com/programming/big-data/data-science/data-science-how-to-deal-with-missing-data-in-python/

Outlier links:

http://colingorrie.github.io/outlier-detection.html

Why care about outliers? There are a couple of reasons:

Outliers distort the picture of the data we obtain using descriptive statitics and data visualization. When our goal is to understand the data, it is often worthwhile to disregard outliers.

Outliers play havoc with many machine learning algorithms and statistical models. When our goal is to predict, our models are often improved by ignoring outliers.

Outliers can be exactly what we want to learn about, especially for tasks like anomaly detection.

Tri metode za detekciju outlier-a: Z-score method, and the modified Z-score method, and the IQR (interquartile range) method.