Big Data – Howest 2023

RAPPORT SPACESHIP TITANIC

Desnyder Jasper, Delacroix Tuur, Anthierens Michiel

Inleiding

Het rapport beschrijft het data-onderzoek van groep 5 met betrekking tot de Spaceship Titanic dataset. Wij hebben verschillende Python-pakketten gebruikt om de dataset in Jupyter Notebook te laden en te verkennen. De dataset betreft een fictieve passagiersreis op een ruimteschip genaamd Titanic, waarbij de passagiers uiteindelijk getransporteerd worden naar een andere dimensie.

Rapport

Voorafgaand aan het trainen van het model zijn verschillende stappen genomen om de dataset te verkennen en schoon te maken. Zo zijn ontbrekende waarden vervangen door de gemiddelde of meest voorkomende waarden, afhankelijk van het type kolom. Ook zijn ontbrekende namen vervangen door "None". Daarna is onderzoek gedaan naar de verbanden tussen de numerieke kolommen met behulp van een heatmap.

Vervolgens zijn de belangrijkste kolommen geselecteerd voor de modellering, waarbij de doelvariabele (Transported) is toegewezen aan de y-variabele en de onafhankelijke variabelen zijn toegewezen aan de X-variabelen. De dataset is opgesplitst in train- en testsets, waarbij 20% van de gegevens voor de testset wordt gebruikt.

Er zijn twee modellen getraind: een SVM-classificator en een random forest-classificator. De hyperparameters van elk model zijn geoptimaliseerd met GridSearchCV. De belangrijkste resultaten van de analyse zijn als volgt: het random forest-model had een hogere nauwkeurigheid, precisie, recall en F1-score dan het SVM-model. De belangrijkste functies voor het random forest-model waren Age, RoomService en Spa. Het random forest-model had een testnauwkeurigheid van 78,84%, wat suggereert dat het een betrouwbaar model is.

Slot

Over het algemeen hebben we de dataset van Spaceship Titanic grondig onderzocht en een nauwkeurig model gecreëerd dat voorspellingen kan doen over het transporteren van passagiers naar een andere dimensie. De gebruikte modellen, SVM-classificator en random forest-classificator, werden geoptimaliseerd met behulp van GridSearchCV. De belangrijkste kenmerken voor het random forest-model waren Age, RoomService en Spa. Het model had een testnauwkeurigheid van 78,84%, wat aangeeft dat het model betrouwbaar is.