DS related project

Berrighi Leonardo

Graph Completion

Link prediction on a real case of study. Qualitative enrichment of knowledge graph. Pre-processing action prior to an entity alignment for matching of the same entities, in the GDPR context. Inferences through KGE models to capture semantics: TransE, TransM and TransR.

Knowledge Base	Machine Learning
KG state of art	Systematic Comparison
Graph Analytics	Algorithms
Graph Embedding	Mathematics
Knowledge Graph Inference	Java

The best models were able to predict **84%** accuracy on relationships, and **93%** accuracy on tail entity.

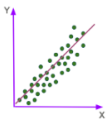
Predicting arrival delay time of commercial flight

Big Data project with Apache Spark on commercial aviation. The application is able to take an input composed of information regarding commercial flights, preprocess it, clean it, transform it, and use the information to construct a model that predicts flight delay time.

Data Analytics	Evaluation & Reporting
Univariate and Multivariate Data Analysis	Scala
Preprocessing	Hadoop
Transformation	SBT
Modeling	ML models

Spark ML

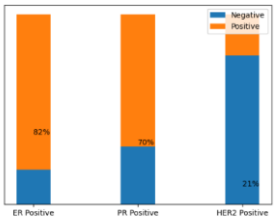
Spark SQL



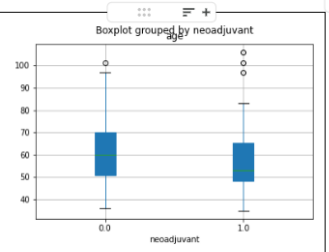
Data cleaning & analysis on breast cancer data

Data Processes project that produces a detailed and accurate analysis from a cleaning and pre-processing perspective, explaining the reasons for the choices.

Data Transformation	Numpy
Data Visualization	Matplotlib
Univariate / Multivariate analysis	Pandas
Python	Seaborn



How many of the patients tested positive for the different receptor and protein markers of the their tumor?



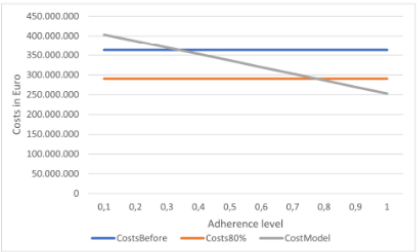
What is the most common age to receive chemotherapy?

Cost analysis on heart attacks data

Data processes project aiming to reduce costs associated with heart attacks in the national healthcare system by a 20%.

Cost analysis	Python
Mathematics	Pandas
Data Processes	Sklearn
ML models	Numpy

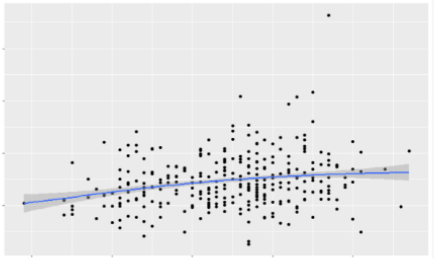
Figure 1: Cost analysis



Interactive dashboard on heart deseases and risk factors

Data visualization project using R and Shiny, to form a more communicative dashboard for the end user.

Data & Task abstraction	**R**
Multivariate & Filtering	**Shiny**
Interaction & Visual encoding	**Data visualization**



Plot of two quantitative variables in a scatter plot with a smooth curve superimposed

Image recognition on xview

The gist of this practical application consists of applying different deep learning techniques over a computer vision classification problem. More advanced techniques are applied to the target problem with each step. From combining the use of incrementally improved techniques and making appropriate configuration decisions, the metrics obtained in the classification problem for each phase are progressively better.

Computer vision	**Keras**
FFNN	**Python**
CNN	**Image recognition**
Transfer learning	**Neural network**

Pre-trained model	Accuracy (%)	Mean recall (%)	Mean precision (%)
VGG16 + DA	79,02	56,5	64,30
InceptionResNetV2 + DA	81,07	62,83	67,10

Table 4. Transfer learning stage 2 experiment results.

ML ranking on Loinc dataset

Information retrieval project for resolution of MLR ranking problem.

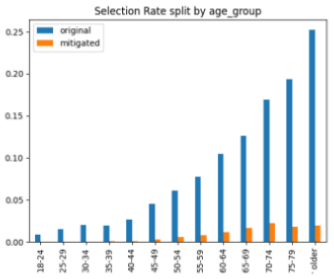
Python	**Sklearn**
Statistical inference	**Numpy**
Relevance	**Pandas**

The results come to have a Mean Average Precision of almost **0.6**.

Testing exploration - Data Integration on bias and fairness dimensions

The objective of this project is to identify the most important difficulties when integrating data, taking into account bias and fairness dimensions.

Data integration	**Aequitas**
Mitigate bias	**Fairlearn**
Discover fairness	**Python**

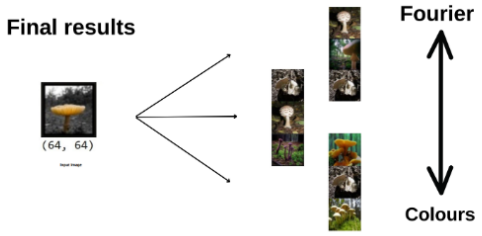


Migitated bias of age-group for heart disease dataset

CBIR app for mushroom recognition

Information Retrieval project. From an image, a series of similar images present in the database are proposed.

Non-textual data extraction	Python
CBIR application	Matplotlib
Color recognition	Numpy
Shape recognition	Scipy
Fourier descriptors	Pandas
Image processing	PIL



Profile-based retrieval engine

The profile-based retrieval system filters and ranks the search results according to the user's preferences. This can help to reduce the amount of irrelevant or unwanted results and provide a more personalized and efficient search experience.

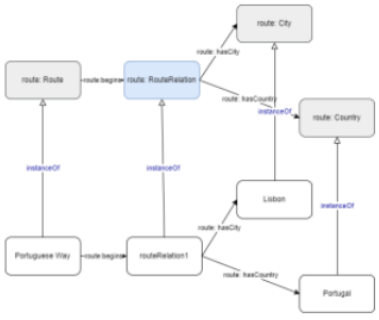
Keyword query	Python
VSM	Numpy
Personalization	Pandas
Cosine similarity	Nltk
Text simplification	Sklearn

The success rate is higher than **85%**.

Identifying Knowledge Patterns and Developing an Ontology

This Intelligent System project is divided in two parts. The object of the first one is to identify at least 4 situations that should be modeled using an N-ary relation pattern. In addition, these new patterns should be classified in the N-ary relation pattern taxonomy. Regarding the objective of the second one, is to develop an ontology (implemented in OWL) in a particular domain. The building of such a network should be done reusing as much as possible existing ontologies and ontology design patterns.

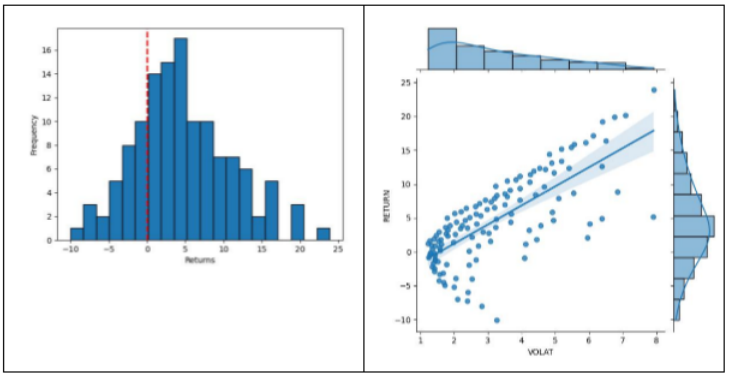
N-ary relation pattern	Ontology
OWL code with 'protege' for the pattern	Ontology design pattern
Taxonomy	CIDOC object-oriented Conceptual Reference Model
Graphical representation	ODP ontology



Investment strategies and analysis on historical assets data

Project related to extraction, processing and analysis of historical assets, with creation and analysis of investment portfolios, strategy performance and data interpretation for analysis.

Harvesting of data	Python
Historical assets data	Selenium
Investment portfolio	Pandas
Investment strategy performance	Matplotlib
Investment strategies analysis	Numpy
Data interpretation	Seaborn



In addition to these **projects**, I have developed several in the **business field**, for example:

- business idea on how to improve engagement and data quality during depression treatments

and others in the context of

- **entrepreneurship & business modellin**
- **technology watch**
- **innovation & entrepreneurship management**

—

All of these projects are provided with ad hoc reports with detailed explanations, the majority of code written and/or supervised for team work. For reasons of privacy and correctness, I will forward such data only to interested parties.