Take-Home Assessment for Data Scientist/Bioinformatician Role

Select one of the two options below as your answer. For either choice, please present your progress using a well-documented GitHub repository. It is essential that your code, data processing steps, and analysis are clearly laid out and commented upon for reviewers. While you may choose to present your results in a Jupyter notebook, we highly encourage the use of Streamlit or an equivalent platform for a more interactive and user-friendly presentation of your findings.

Should you have any questions email bara.badwan@intelligencia.ai or gerry@intelligencia.ai

Option 1

Objective: Develop a machine learning model to predict the IC50 value for drug-cell line combinations using provided datasets.

Background: The Genomics of Drug Sensitivity in Cancer (GDSC) project provides a valuable resource for understanding drug responses in various cancer cell lines. This assessment will make use of a subset of this data.

Datasets:

GDSC data: This dataset contains the response of various cancer cell lines to different drugs, measured as the IC50 value.

- Provided to you in train/test split:
  - gdsc_cell_line_ic50_test_fraction_0.1_id_997_seed_42.csv
  - gdsc_cell_line_ic50_train_fraction_0.9_id_997_seed_42.csv
  - Contains drug identifier, cell-line identifier and the IC50 value

Transcriptomic profile: Gene expression profiles for each cell line

- Provided as gdsc-rnaseq_gene-expression
- List of genes to focus on provided in 2128_genes.pkl
- Cell_lines infos provides information on each cell line

SMILES structure of the drug: A textual representation of the drug's molecular structure.

- Provided in gdsc.smi a two column tsv with the SMILE structure in the first column and its identifier in the second

Task:

Using the provided datasets, construct a machine learning model to predict IC50 values.

Data preprocessing: Handle missing values, normalization, and any other necessary preprocessing steps for the datasets.

Feature engineering: Extract useful features from the transcriptomic profile and SMILES structure to be used in the model.

- SMILES can be converted into Morgan Fingerprints or a Transformer can be used
- The transcriptomic profiles of the cell lines should be embedded

Model development: Choose an appropriate machine learning model, train it, and evaluate its performance using metrics such as RMSE (Root Mean Square Error) or MAE (Mean Absolute Error).

Interpretation: Briefly describe your process and any roadblocks

Option 2

Objective: Use the PyKEEN (Python KnowlEdge EmbeddiNg) pipeline and the Hetionet dataset to train a model and predict indications for a newly introduced drug entity.

Background: Hetionet is an integrative network of biology, disease, and pharmacology. PyKEEN is a software package designed for training and evaluating knowledge graph embedding models. In this assessment, we'll combine the two to predict potential drug indications.

Datasets:

Hetionet dataset: This dataset contains various entities and relationships from the domain of biology, disease, and pharmacology.

Task:

Data Preparation:

- Load the Hetionet dataset into the appropriate format for the PyKEEN pipeline.
- Introduce a new drug entity named "drug new" and establish relationships to gene targets IL15 and IL16.

Model Training:

- Use the PyKEEN pipeline to train a knowledge graph embedding model using the Hetionet dataset.
- You can choose any suitable model architecture available in PyKEEN. Justify your choice in your report.

Prediction:

- Using the trained model, predict which indications "drug new" is expected to be most effective in.

Report:

- Describe the steps taken during data preparation, model training, and prediction.
- Discuss any challenges faced and how they were addressed.
- Analyze and interpret the predicted indications for "drug new". Provide your thoughts on the results.