
# Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

# ML Ranking Assignment

*Authors:*

Sergio Marín Sánchez

Irune Monreal Iraceburu

Ander Ros Ollo

Date: Sunday 3rd March, 2024

# 1 Introduction

In the dynamic landscape of healthcare, the efficient retrieval and ranking of medical information in response to specific queries are crucial for informed decision-making. This report proposes a machine learning approach, specifically Multiple Linear Regression (MLR), to address this challenge. By leveraging MLR, we aim to streamline the process, providing healthcare professionals with a tool that enhances both the speed and accuracy of medical query responses.

This report will outline the rationale behind choosing MLR, detail the construction of a pertinent training set, and discuss the implementation using public libraries. Additionally, we'll explore the significance of dataset extension in fortifying the model's performance. Our endeavor is to contribute to the intersection of machine learning and healthcare, presenting a concise exploration of our MLR approach and its potential impact on advancing medical query ranking systems.

# 2 Pointwise MLR approach

In our pursuit of refining medical query ranking, we adopt the Pointwise Multiple Linear Regression (MLR) approach. This methodology addresses the individuality of each query, treating them independently and providing tailored rankings. Focused on queries like "glucose in blood", "bilirubin in plasma" and "white blood cell count" the Pointwise MLR approach ensures adaptability to diverse query characteristics. By merging the interpretability, simplicity, and efficiency of MLR with the nuanced individual ranking facilitated by the pointwise method, our approach offers a precise and accessible solution to the complex landscape of medical information retrieval.

# 3 Building the Appropriate Training Set

Our approach to medical query ranking involves constructing a robust training set using advanced techniques in natural language processing and machine learning. A pivotal aspect of this process includes converting textual queries and data rows into numerical vectors through TF-IDF vectorization, accentuating the significance of terms in the dataset. This allows for a quantitative evaluation of similarity using the cosine similarity metric.

The cosine similarity scores, ranging from 0 to 1 in the results, elucidate the relevance between the queries and diverse medical tests. A score of 1 indicates a perfect match, scores above 0.7 imply high relevance, while those between 0.5 and 0.7 suggest moderate relevance. Scores below 0.5 signify a lower level of relevance, and a score of 0 indicates no similarity to the query. These scores are then seamlessly integrated into a new DataFrame, streamlining the medical query ranking process for efficient analysis and interpretation.

# 4 Implement model using public libraries

We employed Bayesian Ridge regression models to predict the similarity of medical queries based on relevant features. It combines the principles of ridge regression with Bayesian methods. Ridge regression is a regularization technique that helps prevent overfitting in regression models by adding a penalty to the coefficients of features. Bayesian methods, on the other hand, treat unknown variables as probability distributions and use Bayes' theorem to update these distributions as new data is observed.

In Bayesian Ridge Regression, a probability distribution is assumed over the model parameters and updated using Bayes' theorem as data is observed. This allows for robust estimation of the model parameters and provides a measure of the uncertainty associated with the model predictions. In summary, Bayesian Ridge Regression is useful when accounting for uncertainty in the data and obtaining more robust estimates in the presence of limited or noisy data. Key steps included TF-IDF vectorization of textual data and model training on the encoded features. Evaluation metrics, such as root mean squared error, were used to assess the models' performance.

Example Results for "Glucose in Blood" Query:

| loinc_num | y_test | y_pred |
|---|---|---|
| 49926-9 | 0.47 | 0.59 |
| 15076-3 | 0.62 | 0.77 |
| 74774-1 | 0.58 | 0.36 |
| Model error (RMSE): 0.21 | | |

Example Results for "White Blood Cells Count" Query:

| loinc_num | y_test | y_pred |
|---|---|---|
| 1003-3 | 0.48 | 0.53 |
| 14578-9 | 0.32 | 0.48 |
| 1988-5 | 0.53 | 0.51 |
| Model error (RMSE): 0.19 | | |

# 5 Extend dataset

In an effort to enhance the diversity and comprehensiveness of our dataset, we undertook the task of expanding the queries by exploring additional entries in the LOINC database Search LOINC Home[1]. This extension aims to broaden the scope of medical tests and queries, enabling our models to capture a more extensive range of relationships between queries and corresponding tests.

# 6 Predicted Similarity Scores for new documents

These predicted similarity scores offer insights into the model's perception of similarity between new documents and the given medical queries

| loinc_num | long_common_name | component | system | property | similarity_glucose_in_blood | similarity_bilirubin_in_plasma | similarity_White_blood_cells_count |
|---|---|---|---|---|---|---|---|
| 12345-6 | Glucose in Blood | Glucose | Blood | Mass | 0.63 | -0.17 | 0.20 |
| 12346-7 | Bilirubin in Plasma | Bilirubin | Plasma | Mass | 0.63 | -0.17 | 0.05 |
| 12347-8 | White Blood Cells Count | Blood Cells | Blood | Count | 0.64 | -0.16 | 0.26 |

# References

[1] "SearchLOINC home," LOINC. (), [Online]. Available: https://loinc.org/search/.