INFORMATION RETRIEVAL
EXTRACTION AND INTEGRATION

LAB ML RANKING ASSIGNMENT

José Antonio Ruiz Heredia
Joseph Tartivel
Álvaro Honrubia

TABLE OF CONTENTS

INTRODUCTION

- **Challenge**: Ranking lab tests efficiently and accurately
- **Solution**: *Machine learning ranking (MLR)* methods to optimize search results
- **Focus**: Using *AdaRank* to improve retrieval of lab tests in response to queries
- **Goal**: Build and train a ranking model on documents related to specific medical tests

ADARANK: A BOOSTING ALGORITHM FOR INFORMATION RETRIEVAL

- **Concept**: *AdaRank* improves search results through iterative learning
- **Approach**: Adjusts focus over time, giving more weight to incorrectly ranked cases

- **Advantage over Traditional Methods**:
  - **Pointwise methods**: Score documents individually, don't optimize ordering
  - **Pairwise methods** (*Ranking SVM*, *RankBoost*): Compare document pairs but don't optimize full list
  - ***AdaRank***: Considers entire list at once, directly optimizes performance metrics like *NDCG*

CONSTRUCTING THE TRAINING DATASET: PRE-PROCESSING

	loinc_num	long_common_name	component	system	property
4	1988-5	C reactive protein [Mass/volume] in Serum or Plasma	C reactive pro	Ser/Plas	MCnc
5	1959-6	Bicarbonate [Moles/volume] in Blood	Bicarbonate	Bld	SCnc
6	10331-7	Rh [Type] in Blood	Rh	Bld	Type
7	18998-5	Trimethoprim+Sulfamethoxazole [Susceptibility]	Trimethoprim	Isolate	Susc
8	1975-2	Bilirubin.total [Mass/volume] in Serum or Plasma	Bilirubin	Ser/Plas	MCnc
9	890-4	Blood group antibody screen [Presence] in Serum or Plasma	Blood group a	Ser/Plas	ACnc
10	20565-8	Carbon dioxide, total [Moles/volume] in Blood	Carbon dioxid	Bld	SCnc
11	18906-8	Ciprofloxacin [Susceptibility]	Ciprofloxacin	Isolate	Susc
12	2143-6	Cortisol [Mass/volume] in Serum or Plasma	Cortisol	Ser/Plas	MCnc
13	2075-0	Chloride [Moles/volume] in Serum or Plasma	Chloride	Ser/Plas	SCnc
14	4671-4	Protein C [Mass/volume] in Plasma	Protein C	Plas	MCnc

We need a clean dataset to calculate relevance score

- **“Dirty” Text:** Ponctuation, special characters, capital letters…
- **Useless Information:** Grammar, *LOINC* number, repetition…
- **Abbreviations:** “*Bld*” instead of “*Blood*”, “*Plas*” instead of “*Plasma*”…
- **Bad Structuring:** Units of measurement in the “*name*” column…

CONSTRUCTING THE TRAINING DATASET: PRE-PROCESSING

- **“*Dirty*” Text -> Text Cleaning**

  Convert text to lowercase, remove punctuation and special characters.
- **Useless Information -> Lemmatization**

  Eliminate common stop words and reduce words to their base forms.
- **Abbreviations -> Mapping Dictionary**

  Standardize the terminology.
- **Bad Structuring -> Column Creation**

  Creation of a new column for the unit of measurement.

CONSTRUCTING THE TRAINING DATASET: HYBRID SCORING

Traditional Scoring

Keyword Matching

+ Embeddings Scoring

Semantic Similarity

= Hybrid Scoring

Capture relationships beyond exact keyword matches.

CONSTRUCTING THE TRAINING DATASET: TRADITIONAL SCORING

$$\text{Score}_{\text{traditional}}(q, d) = \sum_{f \in \{component, system\}} \omega_f \cdot \text{Match}_f(q, d)$$

Where:

- ω_f is the weight for field f (e.g., 6.0 for component, 3.0 for system)
- $\text{Match}_f(q, d)$ equals:

Component

  * if there's an exact match between query term and component field:

$$\text{weight}_{\text{component}} \cdot \text{weight}_{\text{component}}$$

  * if the query term is contained within the component field:

$$\text{weight}_{\text{component}} \cdot \frac{\text{weight}_{\text{component}}}{2}$$

  * 0 if there's no match

System

  * if there's an exact match between query term and system field:

$$\text{weight}_{\text{system}} \cdot \text{weight}_{\text{system}}$$

  * if the query term is contained within the system field:

$$\text{weight}_{\text{system}} \cdot \frac{\text{weight}_{\text{system}}}{2}$$

  * 0 if there's no match

For column "*Component*" and "*System*"

To capture exact keyword matches.

CONSTRUCTING THE TRAINING DATASET: EMBEDDINGS SCORING

$$\text{Score}_{\text{embedding}}(q, d) = \sum_{f \in F} \omega_f \cdot \frac{\cos(\vec{q}, \vec{d_f}) + 1}{2} \cdot 5$$

Where:

- F is the set of all fields in the document
- ω_f is the weight for field f
- $\vec{q}$ is the embedding vector of the query
- $\vec{d_f}$ is the embedding vector of field f in document d
- $\cos(\vec{q}, \vec{d_f})$ is the cosine similarity between the query embedding and field embedding
- The term $\frac{\cos(\vec{q}, \vec{d_f})+1}{2}$ normalizes the cosine similarity from [-1,1] to [0,1]
- The multiplier 5 scales the normalized similarity to match the scale of the traditional score

With the pre-trained biomedical embedding model:

BioBERT-MNLI

To calculate semantic similarity between the query and document fields

CONSTRUCTING THE TRAINING DATASET: FINAL RESULT

Query	LOINC Coc	Name	Component	System	Property	Measurement	Normalized_Scor
glucose in blood	1988-5	component reactive protein serum plasma	component reactive protein	serum plasma	mass concentration	mass volume	0.14186244
glucose in blood	1959-6	bicarbonate blood	bicarbonate	blood	substance concentration	mole volume	0.40183312
glucose in blood	10331-7	rh blood	rh	blood	type	type	0.36898455
glucose in blood	18998-5	trimethoprim sulfamethoxazole	trimethoprim sulfamethoxazole	isolate	susceptibility	susceptibility	0.015217709
glucose in blood	1975-2	bilirubin total serum plasma	bilirubin	serum plasma	mass concentration	mass volume	0.6513374
glucose in blood	890-4	blood group antibody screen serum plasma	blood group antibody screen	serum plasma	amount concentration	presence	0.22904973
glucose in blood	20565-8	carbon dioxide total blood	carbon dioxide	blood	substance concentration	mole volume	0.4322821
glucose in blood	18906-8	ciprofloxacin	ciprofloxacin	isolate	susceptibility	susceptibility	0.015049424
glucose in blood	2143-6	cortisol serum plasma	cortisol	serum plasma	mass concentration	mass volume	0.17161444
glucose in blood	2075-0	chloride serum plasma	chloride	serum plasma	substance concentration	mole volume	0.174336
glucose in blood	4671-4	protein component plasma	protein component	plasma	mass concentration	mass volume	0.25051117
glucose in blood	18864-9	ampicillin	ampicillin	isolate	susceptibility	susceptibility	0.042238485
glucose in blood	15076-3	glucose urine	glucose	urine	substance concentration	mole volume	0.7523206
glucose in blood	1798-8	amylase serum plasma	amylase	serum plasma	cell concentration	enzymatic activity volur	0.21758804
glucose in blood	26474-7	lymphocyte blood	lymphocyte	blood	number concentration	volume	0.43848428
glucose in blood	1920-8	aspartate aminotransferase serum plasma	aspartate aminotransferase	serum plasma	cell concentration	enzymatic activity volur	0.22135702
glucose in blood	13317-3	methicillin resistant staphylococcus aureus unspecified specimen organism spe	staphylococcus aureus methicillin resistant isola	xxx	amount concentration	presence	0.019305676

Normalized Relevance Score using Min-Max.

Allowing to train the *AdaRank* Algorithm.

MODEL DEVELOPMENT AND IMPLEMENTATION

Dataset Preparation:

- **Encoded categorical features** to numerical representations
- **Score labels** by scaling normalized scores to integers
- **Split data**: 75% training, 25% testing

Implementation:

- ***LightGBM*** with ***rank_xendcg*** objective
- Simulated ***AdaRank's*** boosting and reweighting mechanism
- **Parameters**: *boosting_type*, *num_leaves*, *learning_rate*, *max_depth*, *feature_fraction*, *label_gain*, *regularization*...

EXPANDING THE DATASET

Query Expansion:

Added new queries: "**calcium in serum**", "**cells in urine**"

Added variations based on component: "*bilirubin*", "*cells*", "*leukocytes*", "*calcium*", "*glucose*"

Added variations based on system: "*blood*", "*serum or plasma*", "*urine*"

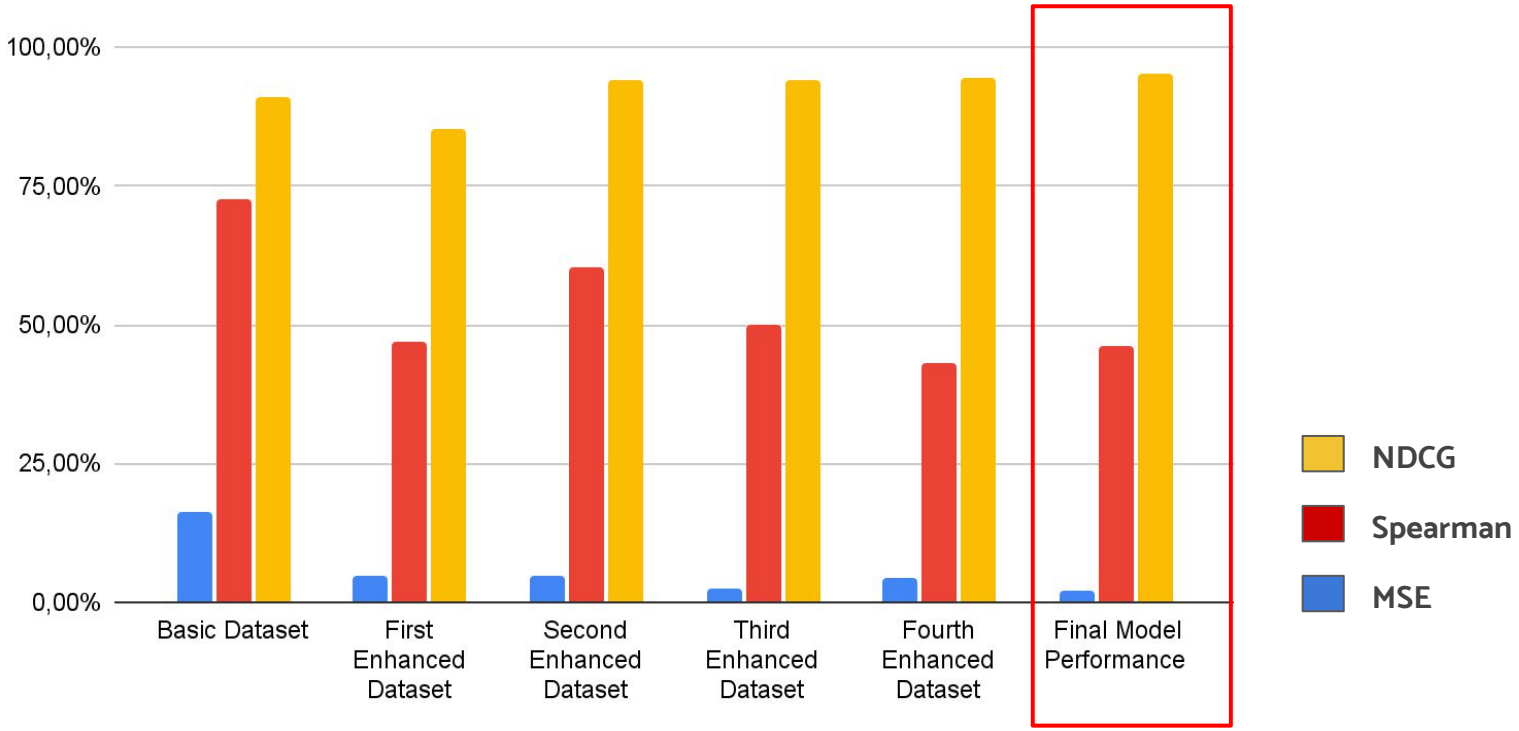
Model Optimization:

Adjusted **hyperparameters** and **learning rates** depending on the dataset

Evaluated each version of the dataset to **measure impact on ranking quality**

EVALUATION

NDCG: 0.9517
Spearman's: 0.4615
MSE: 0.0191



CONCLUSION

Project Insights

- **Dataset diversity** impacts ranking performance
- **Adding medical queries** improves model adaptability
- ***AdaRank*** effectively optimizes ranking for medical retrieval

Implications

- **Improved search efficiency** supports better clinical decision-making
- **Structured search optimization** more important as medical information grows