INFORMATION RETRIEVAL

ML Ranking System using Logistic Regression + Pairwise

JAVIER MARIA ARTEAGA PUELL, RODRIGO CASTAÑÓN MARTÍNEZ Y DAKOTA MELLISH

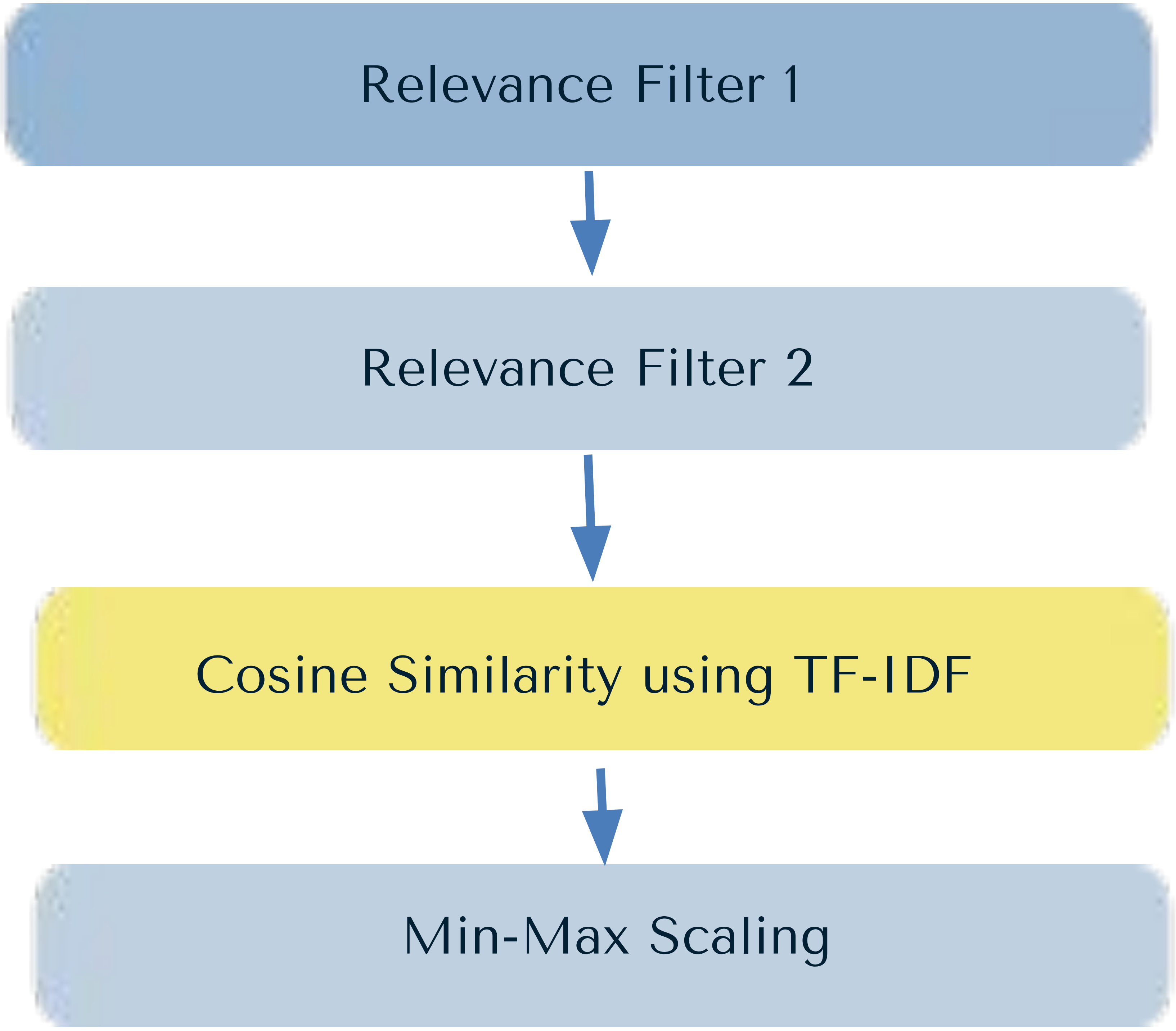
Background

LOINC (Full dataset used, over 100k records)

Libraries such as NLTK and Scikit

"insulin in blood", "Cancer Ag 125 Pleural Fluid" "MRA Thigh vessels contrast"

Preprocessing Steps



Preprocessing Steps

Relevance Filter 1

NLTK stop words function

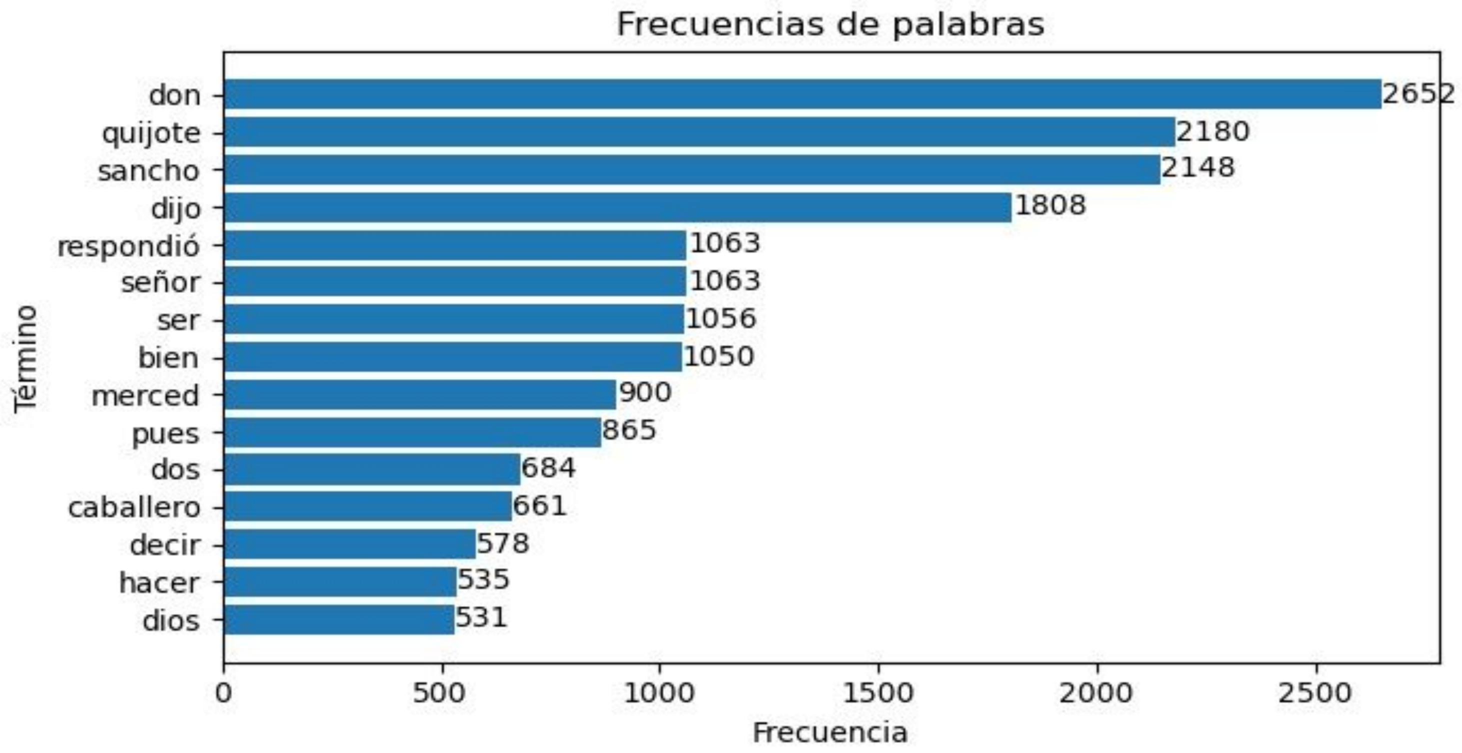
"to", "and", "in" etc.

Checked against "LOINC Common Name"

BEQ	LBTESTCD	LBLOINC	LBTEST	LBCAT	LBORRES	LBORRESU	LBOR
	PLAT	26515-7	Platelet	HEMATOL...	284	THOU/uL	130
	PLAT	26515-7	Platelet	HEMATOL...	266	THOU/uL	130
	PLAT	26515-7	Platelet	HEMATOL...	261	THOU/uL	130
	PLAT	26515-7	Platelet	HEMATOL...	260	THOU/uL	130
	PLAT	26515-7	Platelet	HEMATOL...	293	THOU/uL	130
	PLAT	26515-7					130
	PLAT	26515-7					130
	PLAT	26515-7					130
	RBC	26453-1					4
	RBC	26453-1	Erythrocytes	HEMATOL...	4.40	MILL/uL	4
	RBC	26453-1	Erythrocytes	HEMATOL...	4.30	MILL/uL	4
	RBC	26453-1	Erythrocytes	HEMATOL...	4.30	MILL/uL	4
	RBC	26453-1	Erythrocytes	HEMATOL...	4.40	MILL/uL	4
	RBC	26453-1	Erythrocytes	HEMATOL...	4.40	MILL/uL	4
	RBC	26453-1	Erythrocytes	HEMATOL...	4.30	MILL/uL	4
	RBC	26453-1	Erythrocytes	HEMATOL...	4.20	MILL/uL	4

26515-7 (LBLOINC)
LOINC Name: Platelets:NCnc:Pt:Bld:Qn
LOINC Common Name: Platelets [#/volume] in Blood
Example UCUM Units: 10*3/uL

If a match is found, we give a score equal to the character length of the term (more complex words are rewarded).



We use all terms with a query_score>6 + a random sample

Preprocessing Steps

Relevance Filter 2

Using the results of the previous data, we then performed a more general “check" to see if any of the query terms were found in the loinc_common_name field. The result is a binary variable of 0 or 1 called “relevance".

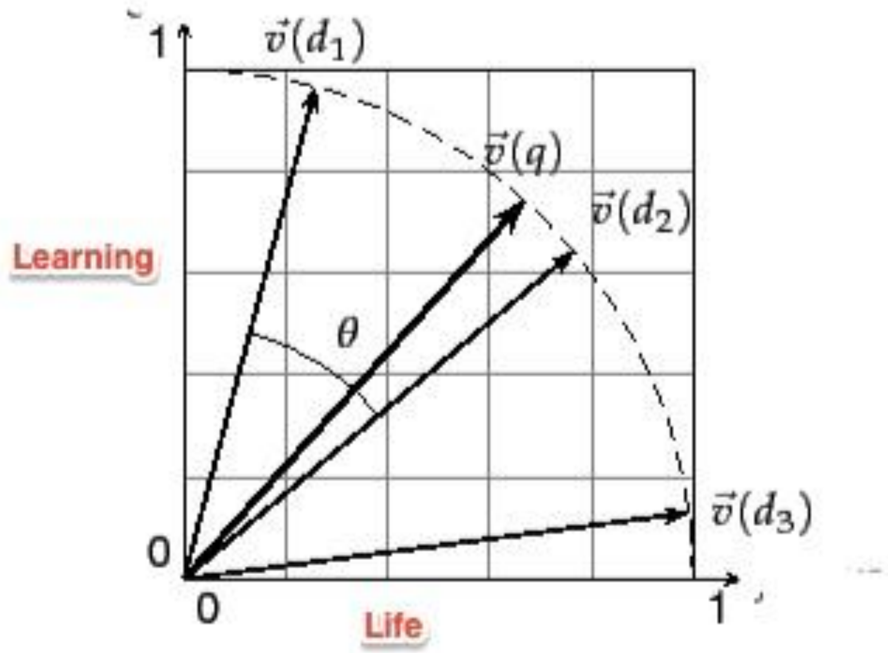
Preprocessing Steps

Cosine Similarity using TF-IDF

We combined the variables *loinc_common_name* and *component* and compare this with the query vector to create **W** by calculating how often the query terms appear in the loinc_common_name + component field

We then use the matrix **W** with the query vector to calculate the cosine similarity matrix.

Cosine similarity score >=.5 is considered "Relevant". We transform this column into a binary variable and use it as our label variable.



$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

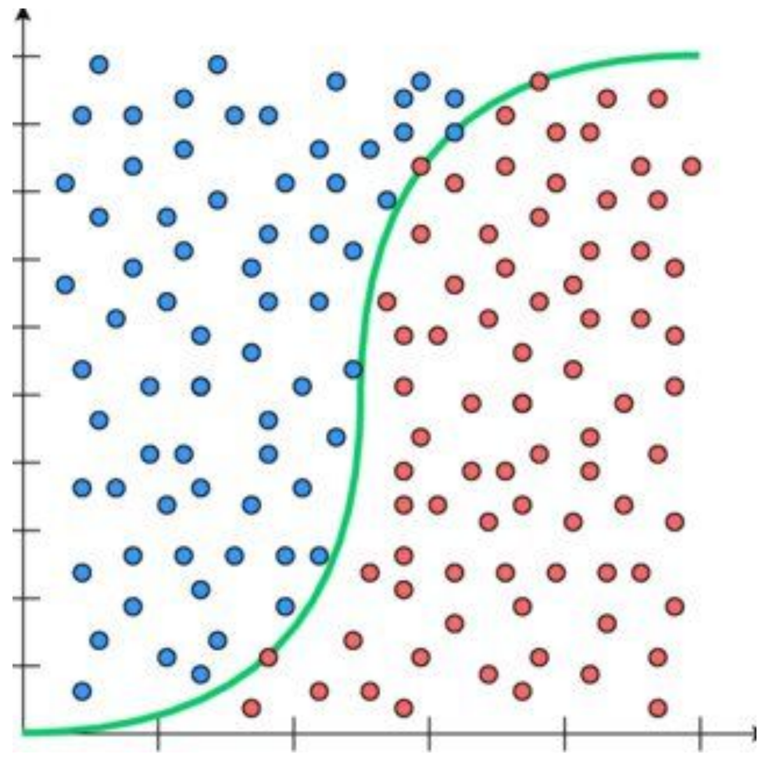
$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Preprocessing Steps

Min-Max Scaling

We apply min-max scaling use scikit-learn to normalize the values of the variable query_score, as it has a high amount of variation

Model Chosen



- Logistic Regression classifier + query_score

- The general model formula looks like

Relevancy Label = query_score + relevance

- Query score is formed from filter 1, Relevancy flag is formed from filter 2, and the Relevancy label is based off of the **cosine similarity** score shown previously.

Model Results

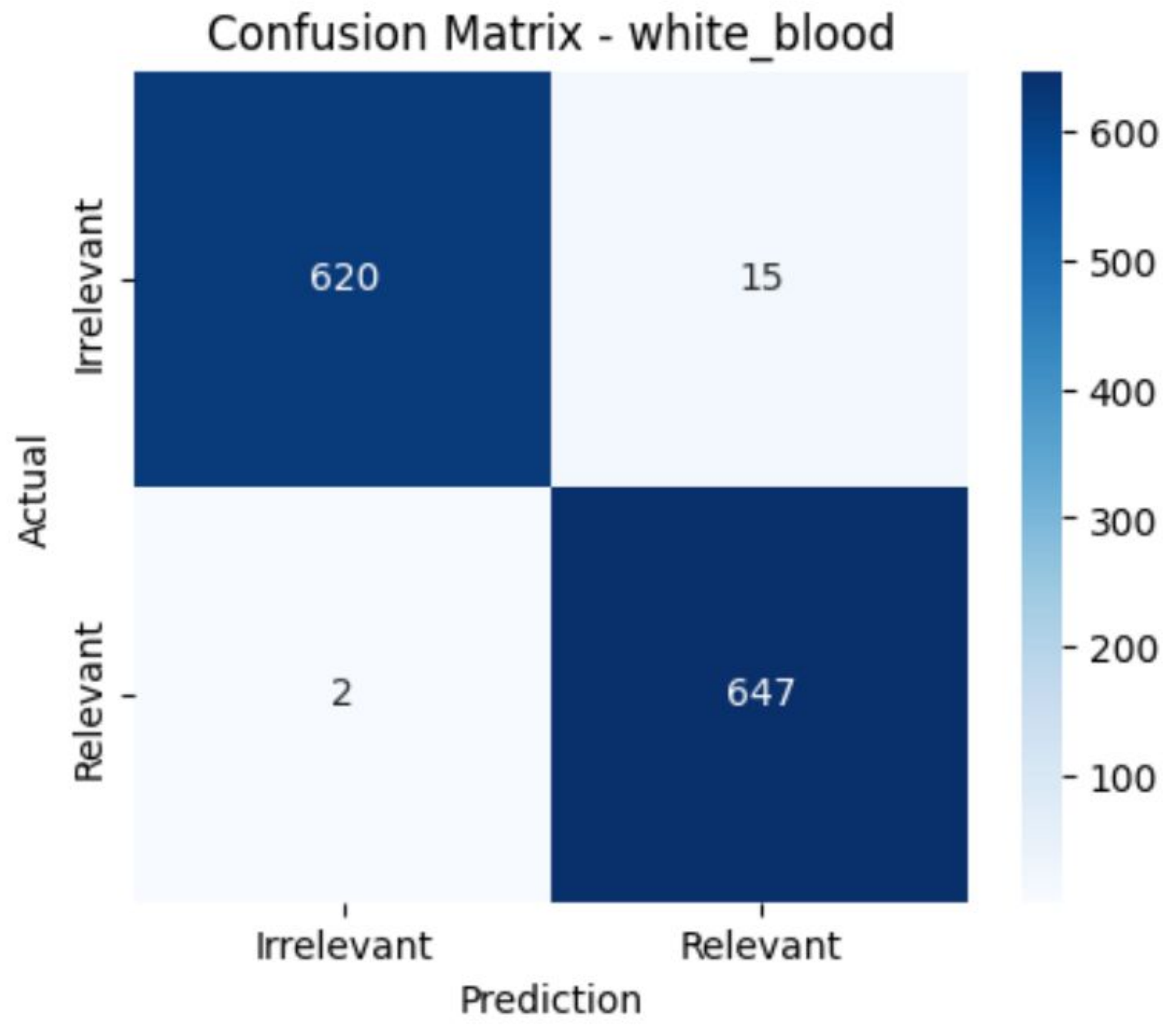
- We consider Confusion Matrix, ROC Curve and Probability Distribution for each query separately
- Overall relatively strong classification results, particularly when considering AUC, though this is somewhat deceptive
- Aided by the fact that the filter 1 provides a balanced set of records
- Model does well with specificity, considering some queries provide more "irrelevant terms" than relevant
- Results highly dependent on query

Model Results by Query

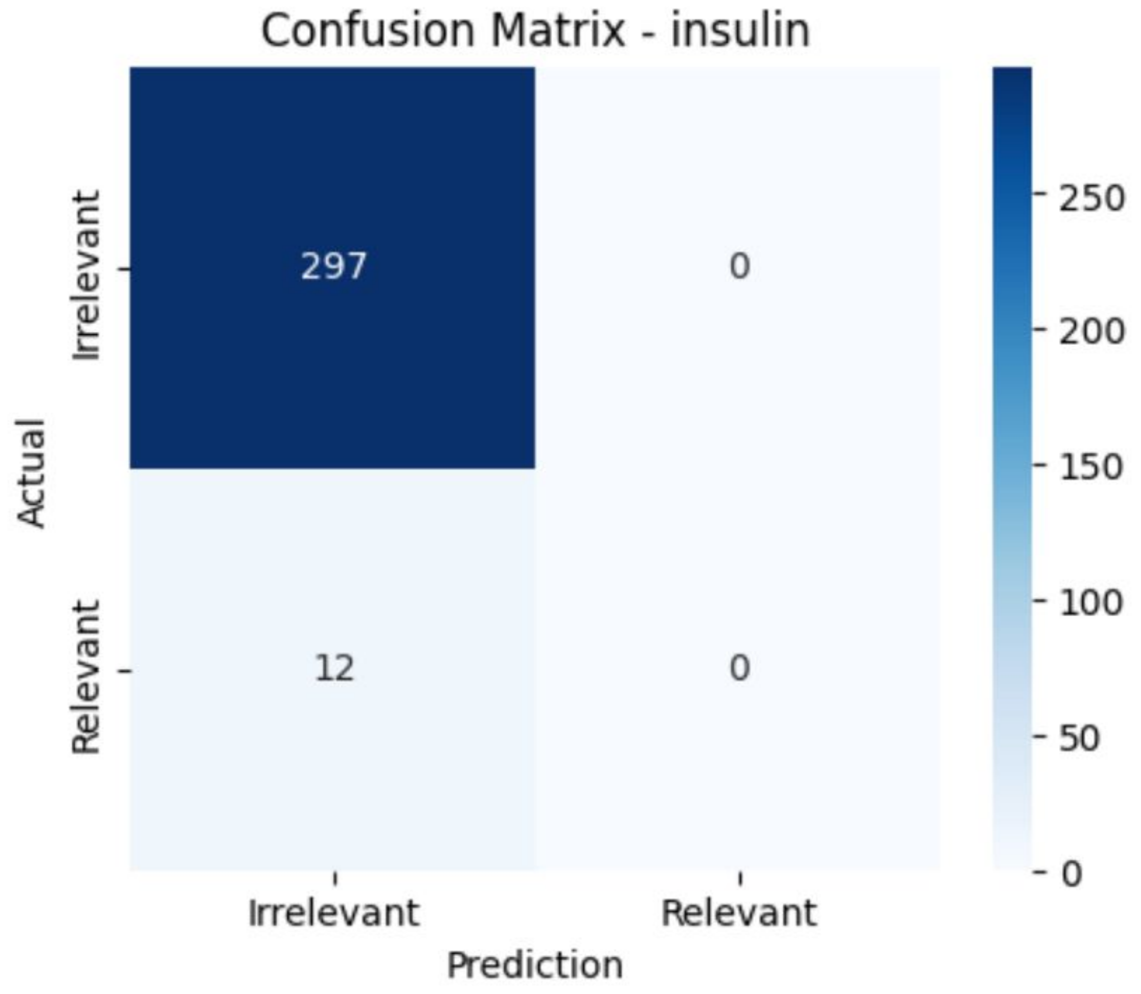
Query	Results	Precision	Recall	F1 Score
Glucose in Blood	0.9672	0.9429	1.0000	0.9706
Bilirubin in Plasma	0.9860	0.9787	0.9938	0.9862
White blood cell count	**0.9860**	**0.9787**	**0.9938**	**0.9862**
Insulin in Blood	**0.9579**	**0.0000**	**0.0000**	**0.0000**
Cancer AG 125 Pleural Fluid	0.8441	0.1163	0.4762	0.1869
MRA Thigh vessels contrast	0.9502	0.4426	0.4821	0.4615
Deoxycortisol in serum	0.8667	0.9333	0.8235	0.8750

Model Results: Confusion Matrix

White Blood Cell Count

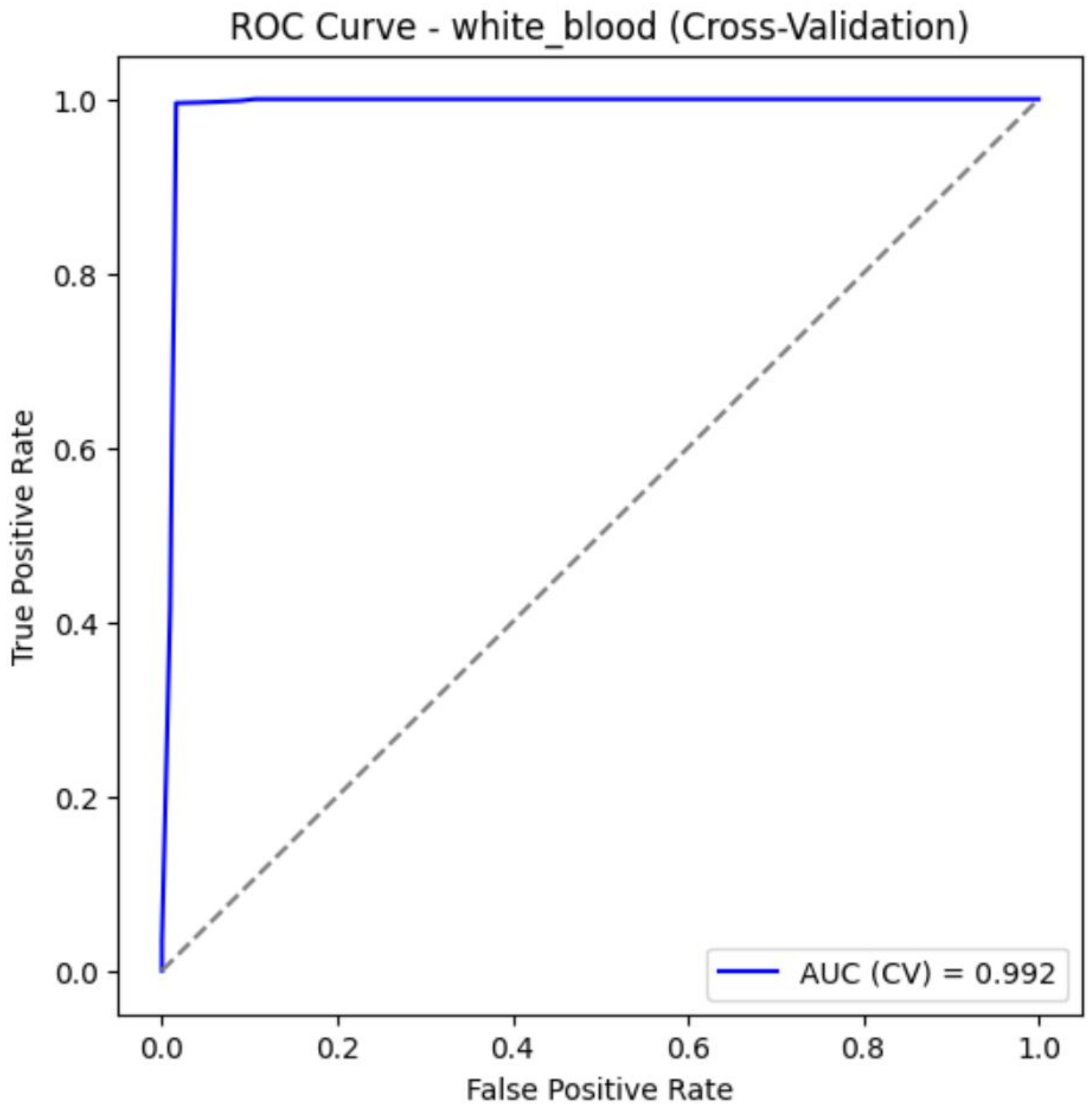


Insulin in Blood

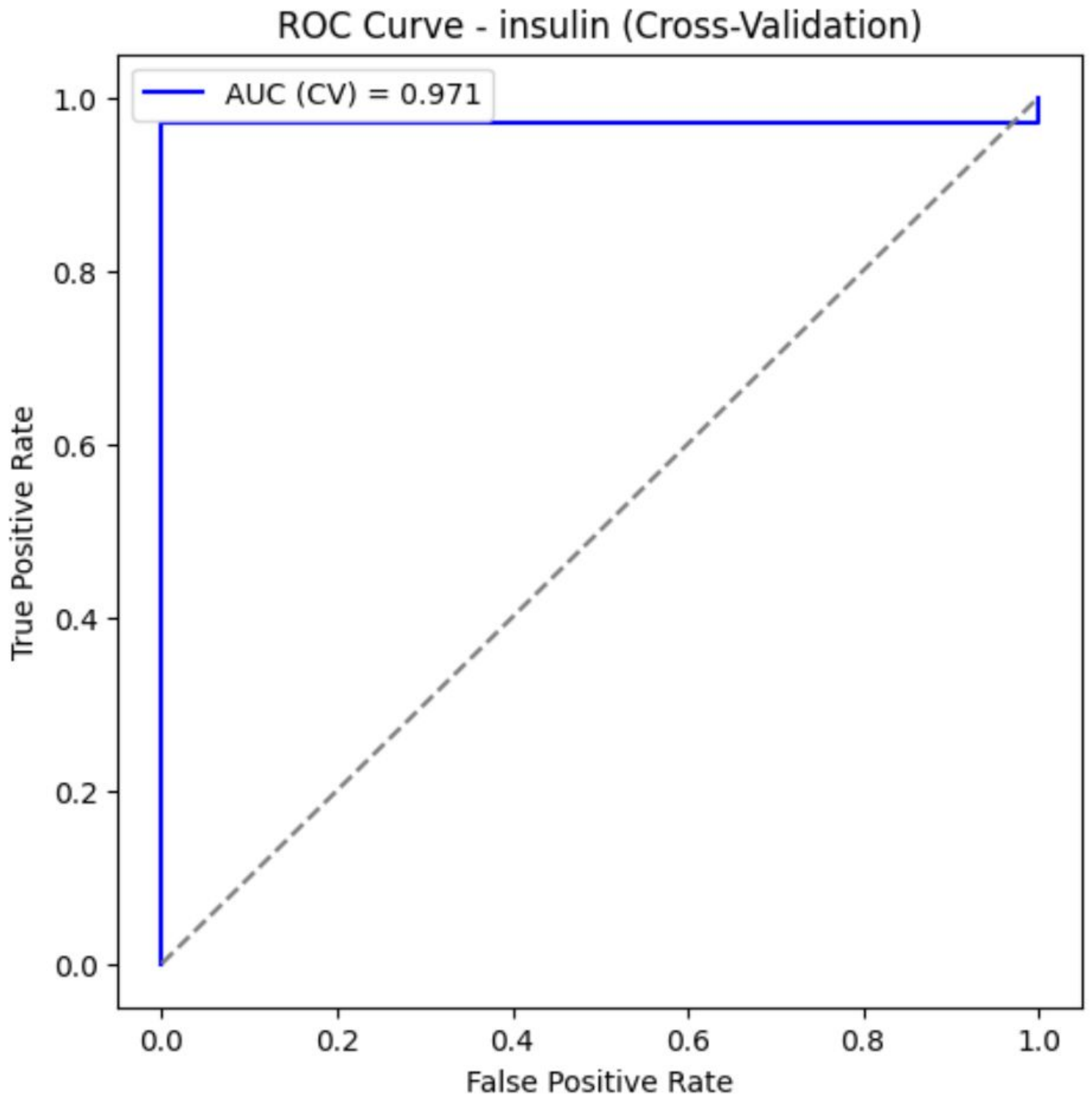


Model Results: Confusion Matrix

White Blood Cell Count



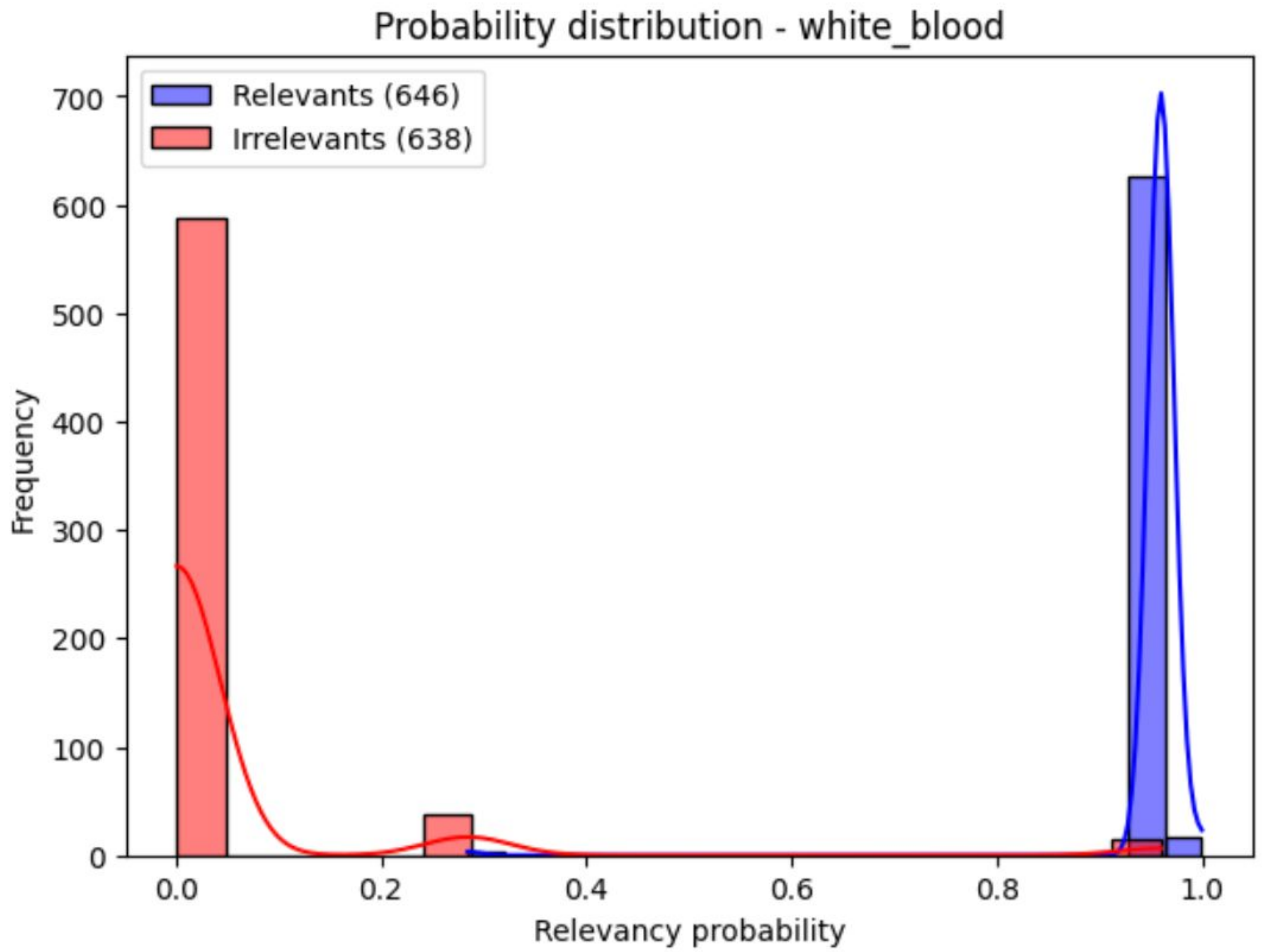
Insulin in Blood

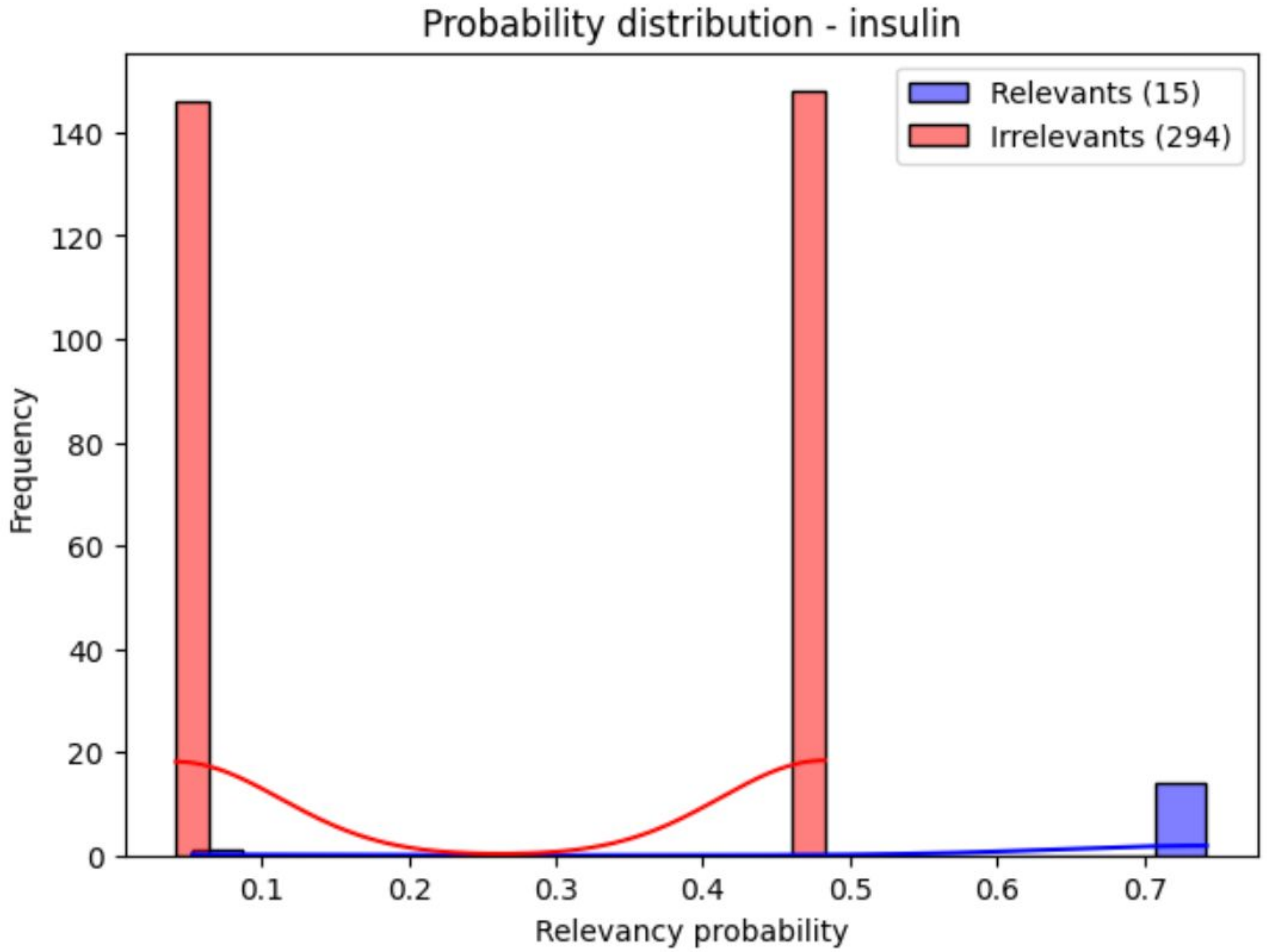


Model Results: Confusion Matrix

White Blood Cell Count

Insulin in Blood





Future Considerations and Conclusion

- Room to expand features, possibly perform subword match to account for misspellings or small differences in query vs results

- Inclusion of variables *System* and *Property*

- Adjust cosine similarity threshold of .5 for labeling

- Adjust query score of >6

- Experiment with other models

- Overall, logistic regression + pointwise provides a fast, straightforward and highly scalable implementation

THANK YOU FOR YOUR ATTENTION!

INFORMATION RETRIEVAL