Language difficulty project

Group Nestle

TABLE OF CONTENTS

01

No data cleaning

02

Other methods

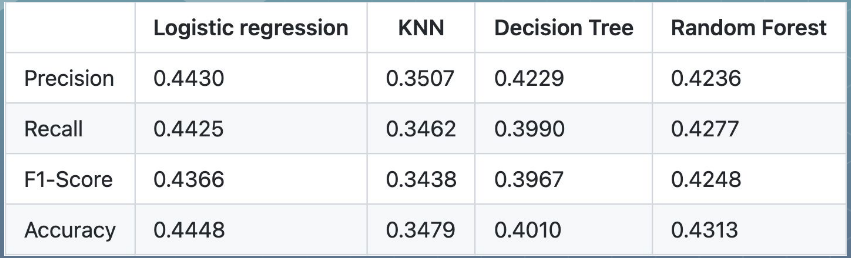
03

Best method

04

Final Score

No data cleaning



	Logistic regression	KNN	Decision Tree	Random Forest
Precision	0.4430	0.3507	0.4229	0.4236
Recall	0.4425	0.3462	0.3990	0.4277
F1-Score	0.4366	0.3438	0.3967	0.4248
Accuracy	0.4448	0.3479	0.4010	0.4313

Logistic Regression is the best (slightly better than Random Forest)

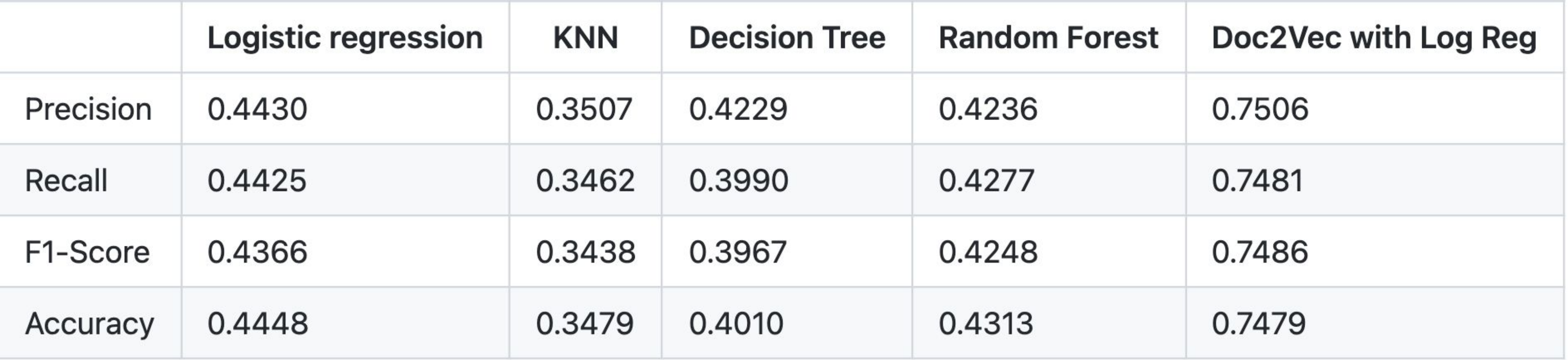
Others Methods

TF-IDF with dimensionality reduction

Standardisation

Spacy & Nltk

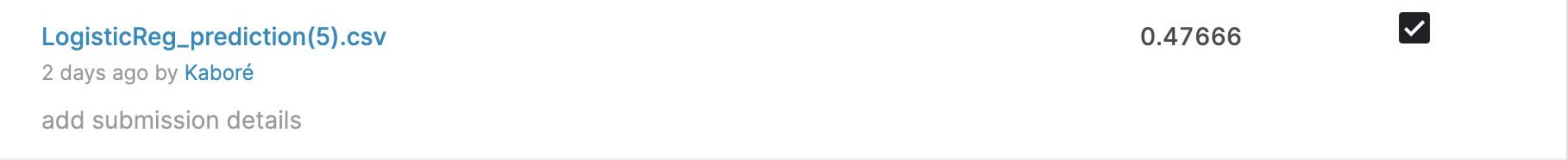
Best Method



	Logistic regression	KNN	Decision Tree	Random Forest	Doc2Vec with Log Reg
Precision	0.4430	0.3507	0.4229	0.4236	0.7506
Recall	0.4425	0.3462	0.3990	0.4277	0.7481
F1-Score	0.4366	0.3438	0.3967	0.4248	0.7486
Accuracy	0.4448	0.3479	0.4010	0.4313	0.7479

DOC2VEC + Log Reg + Nltk

Final score on Kaggle



LogisticReg_prediction(5).csv

2 days ago by Kaboré

add submission details

0.47666

☑