CASE STUDY 2 : PERSONALIZED CANCER DIAGNOSIS

BOW VECTORIZATION

MODEL	TRAINING LOSS	CV LOSS	TESTING LOSS	MISCLASSIFIED %
Naïve Bayes (OHE)	0.862	1.23	1.24	38.9
KNN (Response coding)	0.633	1.03	1.05	37.4
Logstic Regression (OHE)	0.619	1.17	1.1	35.9
Logistic Regression + Balancing (OHE)	0.618	1.15	1.08	37.5
Linear SVM (OHE)	0.746	1.17	1.11	35.9
Random Forest (OHE)	0.712	1.19	1.13	40.4
Random Forest (Response Coding)	0.051	1.23	1.22	42.3
Stacking (OHE)	0.681	1.17	1.11	34.8
Maximum Voting Classifier (OHE)	0.928	1.2	1.19	35.1
Logistic Regression + BOW uni, bi gram vectorization + feature engg.	0.724	1.09	1.05	40.4

F.E + TFIDF + FEATURES →

TFIDF VECTORIZATION (3000 feature, ngram_range=(1,5), min_df=10)

MODEL	TRAINING LOSS	CV LOSS	TESTING LOSS	MISCLASSIFIED %
Naïve Bayes (OHE)	0.573	1.25	1.26	40.6
KNN (Response coding)	0.835	1.13	1.15	37.2
Logstic Regression (OHE)	0.53	0.976	0.993	35.7
Logistic Regression + Balancing (OHE)	0.389	0.972	0.979	35.1
Linear SVM (OHE)	0.447	0.993	1.01	33.8
Random Forest (OHE)	0.889	1.211	1.16	43.8
Random Forest (Response Coding)	0.02	1.92	1.89	74.6
Stacking (OHE)	0.496	1.16	1.15	36.2
Maximum Voting Classifier (OHE)	0.616	1.05	1.07	35.3