Introduction to Data Science and Analytics
Group 2 / Step 6

In this step, we're required to do predictive analysis. In order to do that, because of the fact that our data set has huge number of features, first, we need to apply feature selection to select the best features that contributes most to our prediction performances. We applied three different feature selection methods to our dataset. First one is the Information Gain method, second one is the FStats method, and the last one is Chi Squared method. We applied scikit-learn's algorithms for feature selection. Before the process, we have 7597 features which represent the protein densities; after the process, we have 20 of them. We can see that the selected features from different algorithms are not identical but there are similar ones.

Table 1.0 - Features that selected by Mutual Information Gain method

#	Feature Name	Description	Type
1	IGLL5	IGLL5 Protein Density	Numeric
2	CCL19	CCL19 Protein Density	Numeric
3	CPSF3	CPSF3 Protein Density	Numeric
4	HSH2D	HSH2D Protein Density	Numeric
5	CENPL	CENPL Protein Density	Numeric
6	IGLV1-44	IGLV1-44 Protein Density	Numeric
7	IGKV1-5	IGKV1-5 Protein Density	Numeric
8	SRI	SRI Protein Density	Numeric
9	KCNAB2	KCNAB2 Protein Density	Numeric
10	GAGE7	GAGE7 Protein Density	Numeric
11	SLC5A2	SLC5A2 Protein Density	Numeric
12	NUDT16L1	NUDT16L1 Protein Density	Numeric
13	XG	XG Protein Density	Numeric
14	WDFY4	WDFY4 Protein Density	Numeric
15	IGL@	IGL@ Protein Density	Numeric
16	PFN2	PFN2 Protein Density	Numeric
17	IL9	IL9 Protein Density	Numeric
18	SRPRA	SRPRA Protein Density	Numeric
19	MAIP1	MAIP1 Protein Density	Numeric
20	ABAT	ABAT Protein Density	Numeric

Table 1.1 - Features that selected with FStats score method

#	Feature Name	Description	Type
1	KDELR3	KDELR3 Protein Density	Numeric
2	IGK	IGK Protein Density	Numeric
3	IGLV1-44	IGLV1-44 Protein Density	Numeric
4	IGHG3	IGHG3 Protein Density	Numeric
5	IGLL5	IGLL5 Protein Density	Numeric
6	KCTD4	KCTD4 Protein Density	Numeric
7	KCTD17	KCTD17 Protein Density	Numeric
8	CTLA4	CTLA4 Protein Density	Numeric
9	IGHV4-31	IGHV4-31 Protein Density	Numeric
10	KDELR2	KDELR2 Protein Density	Numeric
11	DNAJC16	DNAJC16 Protein Density	Numeric
12	IGHG1	IGHG1 Protein Density	Numeric
13	KCTD14	KCTD14 Protein Density	Numeric
14	UBE2V1	UBE2V1 Protein Density	Numeric
15	IGKV1-5	IGKV1-5 Protein Density	Numeric
16	KCNS3	KCNS3 Protein Density	Numeric
17	KCNN3	KCNN3 Protein Density	Numeric
18	IGL@	IGL@ Protein Density	Numeric
19	TRIM21	TRIM21 Protein Density	Numeric
20	KDM2A	KDM2A Protein Density	Numeric

Table 1.3 - Features that selected by Chi Square method

#	Feature Name	Description	Type
1	GMEB1	GMEB1 Protein Density	Numeric
2	BARX1	BARX1 Protein Density	Numeric
3	CTLA4	CTLA4 Protein Density	Numeric
4	BRN	BRN Protein Density	Numeric
5	IGHG1	IGHG1 Protein Density	Numeric
6	IGHG3	IGHG3 Protein Density	Numeric
7	IGL@	IGL@ Protein Density	Numeric
8	IGLL5	IGLL5 Protein Density	Numeric
9	IGLV1-44	IGLV1-44 Protein Density	Numeric
10	KCNMB3	KCNMB3 Protein Density	Numeric
11	KCNN2	KCNN2 Protein Density	Numeric
12	KCNS2	KCNS2 Protein Density	Numeric
13	KCNS3	KCNS3 Protein Density	Numeric
14	KCTD14	KCTD14 Protein Density	Numeric
15	KCTD4	KCTD4 Protein Density	Numeric
16	KDM8	KDM8 Protein Density	Numeric
17	KEL	KEL Protein Density	Numeric
18	KHDRBS2	KHDRBS2 Protein Density	Numeric
19	KIAA0391	KIAA0391 Protein Density	Numeric
20	TRIM21	TRIM21 Protein Density	Numeric

Table 1.4 - Evaluation of Experiments

#	Experiment	Accuracy	F1-Macro	F1-Micro
1	KNN Classification (without fs)	0.8666	0.8795	0.8666
2	Naive Bayes (GaussianNB) (without fs)	0.6666	0.66645	0.6666
3	Decision Tree (With Mutual Info Selection)	0.975	0.955	0.975
4	KNN Classification –2 (With FStats Info Selection)	0.9125	0.9089	0.9125
5	Naive Bayes (GaussianNB)–2 (Chi Square Selection)	0.8875	0.8573	0.8875

We applied 5 different tests and evaluate them. Because of the fact that number of instances in our dataset is not so big, first we apply tests without feature selection. We can see that the highest accuracy score is in decision tree method which gives 97% accuracy. Actually, it's so accurate that we thought there could be a problem, but when we compare the results, it gives correct classification.

On the other hand, we can see the difference of accuracies between KNN Classification and Naive Bayes Classification without feature selection and, the ones with feature selection. Especially, there's a huge improvement on Naive Bayes classification when we use feature selection which confirms that feature selection methods are used to improve the accuracy of the classification/prediction algorithms.