دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مدیریت علم و فناوری

گزارش کار هفته ششم

طبقه بندی

نگارش

رضا اکبری مقدم

استاد

دکتر مهدی قطعی

آذر ماه ۹۹

فهرست مطالب

فهرست اشکال

مقدمه

دسته بندی یا Classification یک از شاخه های علوم داده **یا** Data Science می باشد. که در بحث پیش بینی آینده predicting the future با استفاده از تجزیه و تحلیل داده ها است. دسته بندی با نام های دیگری همچون طبقه بندی ، کلاس بندی و Classification نیز شناخته می شود. طبقه بندی یک کار علمی داده ها برای پیش بینی مقدار متغیر طبقه بندی شده (هدف یا کلاس) با ساختن یک مدل بر اساس یک یا چند متغیر عددی و / یا دسته ای (پیش بینی کننده یا ویژگی) است.

طبقه بندی (classification) علمی است که بر اساس داده‌های قبلی که دارای برچسب هستند، مدلی برای پیش بینی برچسب داده‌های جدید می‌سازد.

طبقه بندی classification یکی از زیر شاخه های اساسی یادگیری ماشین و داده کاوی است. و اساس آن داده‌های جمع‌آوری شده از اعمال گذشته هستند. اعمالی که بر اساس دانش فرد خبره برچسب گذاری شدند. برای اینکه یک مدل طبقه بند خوب داشته باشیم، باید با داده‌ها و ساختار آنها و نیز تعداد دسته‌ها (برچسب-کلاس-طبقه) اطلاع داشته باشیم. هر چند آشنایی با ساختار و نوع داده‌ها گاها عملی غیر ممکن است اما در صورت وجود یک آشنایی ساده گاها نیز می‌توان مدل طبقه بند درست را انتخاب کرد.

مجموعه داده

مجموعه داده Covertype که به بررسی پیش بینی نوع پوشش جنگل با استفاده از متغیرهای نقشه برداری میپردازد. این نوع پوشش جنگل ها بر اساس مشاهدات انجام شده از داده های سیستم اطلاعات منابع منطقه ۲ سرویس جنگل ایالات متحده تعیین شده است.

این منطقه مورد مطالعه شامل چهار منطقه بیابانی واقع در جنگل ملی روزولت در شمال کلرادو است. این مناطق نشان دهنده جنگل هایی با حداقل آشفتگی ناشی از دخالت های انسان است ، به طوری که انواع پوشش جنگلی موجود بیشتر نتیجه فرآیندهای اکولوژیکی است تا اقدامات مدیریت جنگل.

برخی از اطلاعات زمینه ای برای این چهار منطقه بیابانی: Neota (منطقه ۲)احتمالاً دارای بالاترین مقدار ارتفاعی از ۴ منطقه بیابانی است. Rawah (منطقه ۱) و Comanche Peak (منطقه ۳)دارای یک مقدار میانگین پایین تر از ارتفاع هستند ، در حالی که Cache la Poudre (منطقه ۴) کمترین میانگین ارتفاع را دارد.

همچنین ویژگی پوشش گیاهی شامل ۶ نوع پوشش مختلف به شرح زیر میباشد:

۱. صنوبر (نوع ۱)
۲. کاج گلدان (نوع ۲)
۳. کاج Ponderosa (نوع ۳)
۴. چوب پنبه / بید (نوع ۴)
۵. گون (نوع ۵)
۶. Douglas-fir (نوع ۶)

این مجموعه داده شامل ۵۸۱۰۱۱ سطر و ۵۴ ویژگی میباشد.

گزارش مراحل انجام کار

در این گزارش بنده با استفاده از ابزار پایتون به بررسی و طبقه بندی این مجموعه داده پرداخته ام.

ابتدا به بررسی ستون های اولیه این مجموعه داده میپردازیم.

	0	1	2	3	4	5	6	7	8	9	...	45	46	47	48	49	50	51	52	53	54
0	2596	51	3	258	0	510	221	232	148	6279	...	0	0	0	0	0	0	0	0	0	5
1	2590	56	2	212	-6	390	220	235	151	6225	...	0	0	0	0	0	0	0	0	0	5
2	2804	139	9	268	65	3180	234	238	135	6121	...	0	0	0	0	0	0	0	0	0	2
3	2785	155	18	242	118	3090	238	238	122	6211	...	0	0	0	0	0	0	0	0	0	2
4	2595	45	2	153	-1	391	220	234	150	6172	...	0	0	0	0	0	0	0	0	0	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
581007	2396	153	20	85	17	108	240	237	118	837	...	0	0	0	0	0	0	0	0	0	3
581008	2391	152	19	67	12	95	240	237	119	845	...	0	0	0	0	0	0	0	0	0	3
581009	2386	159	17	60	7	90	236	241	130	854	...	0	0	0	0	0	0	0	0	0	3
581010	2384	170	15	60	5	90	230	245	143	864	...	0	0	0	0	0	0	0	0	0	3
581011	2383	165	13	60	4	67	231	244	141	875	...	0	0	0	0	0	0	0	0	0	3

581012 rows × 55 columns

شکل ۱

همانطور که در جدول مشاهده میشود این مجموعه داده شامل ۵۴ ستون میباشد. با استفاده از متادیتای این مجموعه داده متوجه میشویم که در واقع این مجموعه داده شامل ۱۳ ویژگی میباشد و ۳۱ ستون این مجموعه داده باینری میباشد.

لذا در مرحله پیش پردازش این ستون ها را تبدیل به یک ستون میکنیم و مقادیر نوع هر سطر را برایش در نظر میگیریم به این صورت که ستون ۱۰ تا ۱۳ ویژگی بیانگر ۴ منطقه بیابانی این مجموعه داده میباشد و ستون های ۱۴ تا ۵۳ نمایانگر نوع خاک آن منطقه را بیان میکند.

	Elevation	Aspect	Slope	Horz_Dis_To_Hy	Vert_Dis_To_Hy	Horz_Dis_To_Rw	Hs_9am	Hs_Noon	Hs_3pm	Horz_Dis_To_Fp	Wild_Area	Soil	Cover
0.0	2596	51	3	258	0	510	221	232	148	6279	0	29	5
1.0	2590	56	2	212	-6	390	220	235	151	6225	0	29	5
2.0	2804	139	9	268	65	3180	234	238	135	6121	0	12	2
3.0	2785	155	18	242	118	3090	238	238	122	6211	0	30	2
4.0	2595	45	2	153	-1	391	220	234	150	6172	0	29	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...
581007.0	2396	153	20	85	17	108	240	237	118	837	2	2	3
581008.0	2391	152	19	67	12	95	240	237	119	845	2	2	3
581009.0	2386	159	17	60	7	90	236	241	130	854	2	2	3
581010.0	2384	170	15	60	5	90	230	245	143	864	2	2	3
581011.0	2383	165	13	60	4	67	231	244	141	875	2	2	3

581012 rows × 13 columns

شکل ۲

پس از یکی کردن ستون ها برچسب های ویژگی ها را هم مشخص میکنیم تا بتوانیم بهتر تحلیل انجام دهیم. همانطور که مشاهده میشود ستون های ۱۰ تا ۱۳ به ستون wild_area که شامل ۴ نوع مقدار میباشد و ستون های ۱۴ تا ۵۳ به ستون soil که شامل ۳۰ مقدار میباشد کاهش داده شدند.

```
Cover
1    211840
2    283301
3     35754
4      2747
5      9493
6     17367
7     20510
```

شکل ۳

همانطور که مشاهده میشود در شکل ۳ فراوانی هر پوشش گیاهی مشخص شده است که طبقه بندی را براساس نوع پوشش گیاهی انجام میدهیم.

سپس ۲۰ درصد داده ها را به داده های آموزشی(train) و ۸۰ درصد داده ها را به داده های تست(test) تقسیم میکنیم.

با استفاده از روش انسمبل به بررسی انواع کلاسیفایر ها میپردازیم که کدام کلاسیفایر بر روی داده های ما نتیجه بهتری میدهد.

در اینجا از ۳ کلاسیفایر درخت تصمیم، لگاریتم رگرسیون و k نزدکترین همسایگان استفاده کردیم.

```
train accuracy
Decision Tree :
0.9270844986790359
Logistic Regression :
0.6180133043036755
KNN :
0.9690971833773655

 test accuracy
Desicion Tree :
0.8600780105376618
Logistic Regression :
0.6199428152208757
KNN :
0.92007684877014
```

شکل ۴

همانطور که در شکل ۴ مشاهده میکنید رو رگرسیون لجستیک بر روی داده های ما دقت پایینی دارد و دو روش درخت تصمیم و K نزدیکترین همسایگان دقت قابل قبولی دارند.

لذا برای ادامه کار از درخت تصمیم به عنوان الگوریتم منتخب استفاده میکنیم.

	precision	recall	f1-score	support
1	0.93	0.93	0.93	42275
2	0.94	0.94	0.94	56602
3	0.91	0.91	0.91	7269
4	0.79	0.81	0.80	546
5	0.81	0.80	0.80	1929
6	0.85	0.84	0.85	3496
7	0.94	0.93	0.93	4086
accuracy			0.93	116203
macro avg	0.88	0.88	0.88	116203
weighted avg	0.93	0.93	0.93	116203

شکل ۵

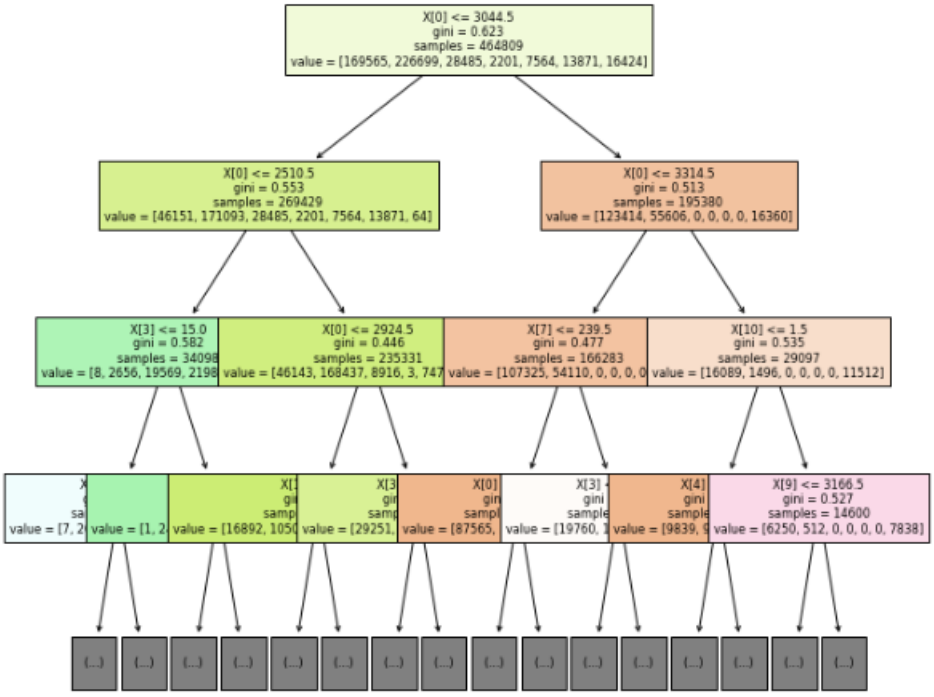
در شکل ۵ مشاهده میشود که مقادیر precision و recall و f1 مجموعه تست محاسبه شده است که مقادیر قابل قبولی را نمایش میدهد.

:

	max-depth	train_acc	test_acc
0	10	0.768064	0.764214
1	12	0.802347	0.796528
2	14	0.829728	0.819075
3	16	0.849700	0.834961
4	18	0.862883	0.846037
5	20	0.870287	0.850641
6	22	0.873161	0.852732
7	24	0.873647	0.853050
8	26	0.873697	0.853102
9	28	0.873742	0.853050
10	30	0.873737	0.853076
11	32	0.873757	0.853093
12	34	0.873757	0.853093

شکل ۶

در ادامه برای بررسی بیش برازش با استفاده از یک حلقه عمق درخت را تغییر میدهیم تا به نقطه تغییر جهت مجموعه داده به بیش برازش برسیم. اما همانطور که در شکل ۶ مشخص میباشد این مجموعه داده بیش برازش ندارد و دقت داده های تست و آموزش با همدیگه رشد میکنند و خلاف همدیگر نمیشوند.



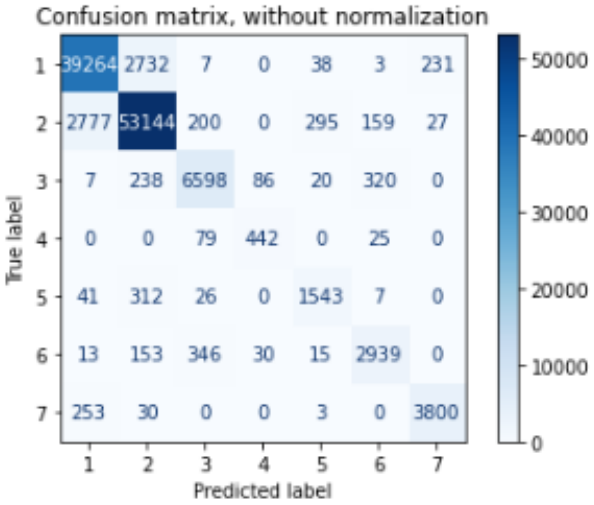
شکل ۷

در شکل ۷ نمایی از درخت تصمیم این مجموعه داده تا عمق ۳ را نمایش داده شده است.

	Actual	Predicted
0	2	2
1	1	1
2	6	2
3	2	1
4	1	1
5	2	2
6	2	2
7	2	2
8	2	2
9	2	2
10	2	2
11	2	1
12	2	2
13	1	1
14	1	1

شکل ۸

در شکل ۸ نمونه ای مشاهده میشود که درخت تصمیم در پیشبینی کلاس های عناصر مجموعه تست چه عملکردی داشته است. که در اکثر سطرها تشخیص درستی داشته است.



شکل ۹

همچنین در شکل ۹ confusion matrix این مجموعه داده مشاهده میشود که اکثر کلاس ها با کمترین خطا پیشبینی شده اند

نتیجه گیری

بر اساس این گزارش میتوان به این نتیجه رسید که الگوریتم های مختلف طبقه بندی بر روی هر مجموعه داده میتوانند خروجی متفاوتی را داشته باشند که میتوان با تغییر مقادیر مختلفی از جمله درصد شرکت داده های آموزشی یا در درخت تصمیم با تغییر عمق درخت مشکلاتی اعم از بیش برازش را رفع کرد.

منابع

١. https://wiki.pathmind.com/accuracy-precision-recall-f١
٢. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html
٣. https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html
٤. https://statinfer.com/204-3-5-information-gain-in-decision-tree-split/
٥. https://chistio.ir/%D8%B7%D8%A8%D9%82%D9%87-%D8%A8%D9%86%D8%AF-%D8%AA%D8%B1%DA%A9%DB%8C%D8%A8%DB%8C-ensemble-classifier-bagging-boosting/
٦. https://archive.ics.uci.edu/ml/datasets/Covertype