اصغر فرهادی

دانشگاه صنعتی بنگلادش
دانشکده کامپیوتر - گروه علوم کامپیوتر
شماره دانشجویی

مستندات پروژه پیشبینی و طبقه بندی سرطان ریه

۱۸ دی ماه ۱۴۰۲

مقدمه

این پروژه به پیش‌بینی سطوح خطر سرطان ریه در بیماران بر اساس مختلف عوامل دموگرافیک و شیوه زندگی می‌پردازد. مجموعه داده اطلاعاتی در مورد بیماران ارائه می‌دهد، از جمله سن، جنسیت، آلودگی هوا، مصرف الکل و سایر ویژگی‌های مرتبط با سلامت. هدف اصلی آموزش و مقایسه مدل‌های یادگیری ماشین برای ارزیابی کارآیی آنها در پیش‌بینی سطوح خطر سرطان ریه است.

توضیحات مجموعه داده

این مجموعه داده حاوی اطلاعاتی در مورد بیماران مبتلا به سرطان ریه، از جمله سن، جنسیت، قرار گرفتن در معرض آلودگی هوا، مصرف الکل، آلرژی گرد و غبار، خطرات شغلی، خطر ژنتیکی، بیماری مزمن ریوی، رژیم غذایی متعادل، چاقی، وضعیت سیگار کشیدن، وضعیت سیگاری غیرفعال، درد قفسه سینه است. سرفه خونی، میزان خستگی، کاهش وزن، تنگی نفس، خس خس سینه، مشکل در بلع، چاق شدن ناخن انگشتان، سرماخوردگی مکرر، سرفه های خشک و خروپف. با تجزیه و تحلیل این داده‌ها می‌توانیم بینشی در مورد عوامل ایجاد سرطان ریه و بهترین روش درمان آن به دست آوریم.

ستون‌ها (ویژگی ها)

1. Age: سن بیمار.
2. Gender: جنسیت بیمار.
3. Air Pollution: سطح تعرض به آلودگی هوا بیمار.
4. Alcohol Use: سطح مصرف الکل بیمار.
5. Dust Allergy: سطح حساسیت به گرد و غبار بیمار.

6. Occupational Hazards: سطح خطرات شغلی بیمار.

7. Genetic Risk: سطح خطر ژنتیکی بیمار.

8. Chronic Lung Disease: سطح بیماری مزمن ریه بیمار.

9. Balanced Diet: سطح رژیم غذایی متعادل بیمار.

10. Obesity: سطح چاقی بیمار.

11. Smoking: سطح سیگار کشیدن بیمار.

12. Passive Smoker: سطح سیگار کشیدن غیرفعال بیمار.

13. Chest Pain: سطح درد سینه بیمار.

14. Coughing of Blood: سطح خون‌ریزی از دهان بیمار.

15. Fatigue: سطح خستگی بیمار.

16. Weight Loss: سطح افت وزن بیمار.

17. Shortness of Breath: سطح تنگی نفس بیمار.

18. Wheezing: سطح خس خس سینه بیمار.

19. Swallowing Difficulty: سطح مشکل در بلع بیمار.

20. Clubbing of Finger Nails: سطح تغییرات در ناخن‌های بیمار.

مراحل اجرای پروژه

1. تحلیل و پیش‌پردازش داده:

- بارگیری مجموعه داده و بررسی اولیه چند ردیف اول.

- تبدیل نام ویژگی‌ها به حروف کوچک.

- حذف ویژگی های اضافی

- حذف رکوردهای تکراری

- بررسی وجود یا عدم وجود missing value در دیتاست

- بررسی balance و یا imbalance بودن دیتا

- بررسی آماری دیتا و بررسی همبستگی بین ویژگی ها

2. تقسیم داده به مجموعه‌های آموزش و آزمایش:

- تقسیم مجموعه داده به مجموعه‌های آموزش و آزمایش برای ارزیابی مدل.

3. انتخاب مدل:

- انتخاب مدل‌های یادگیری ماشین مناسب برای مسئله طبقه‌بندی.

- آموزش و ارزیابی مدل‌ها از جمله:

- رگرسیون لجستیک

- جنگل تصادفی

- شبکه عصبی چند لایه (MLP)

- درخت تصمیم

- k-نزدیک ترین همسایه (KNN)

4. ارزیابی مدل:

- ارزیابی عملکردی مدل با استفاده از معیارهایی مانند accuracy, precission, recall, f1 score, confusion matrix

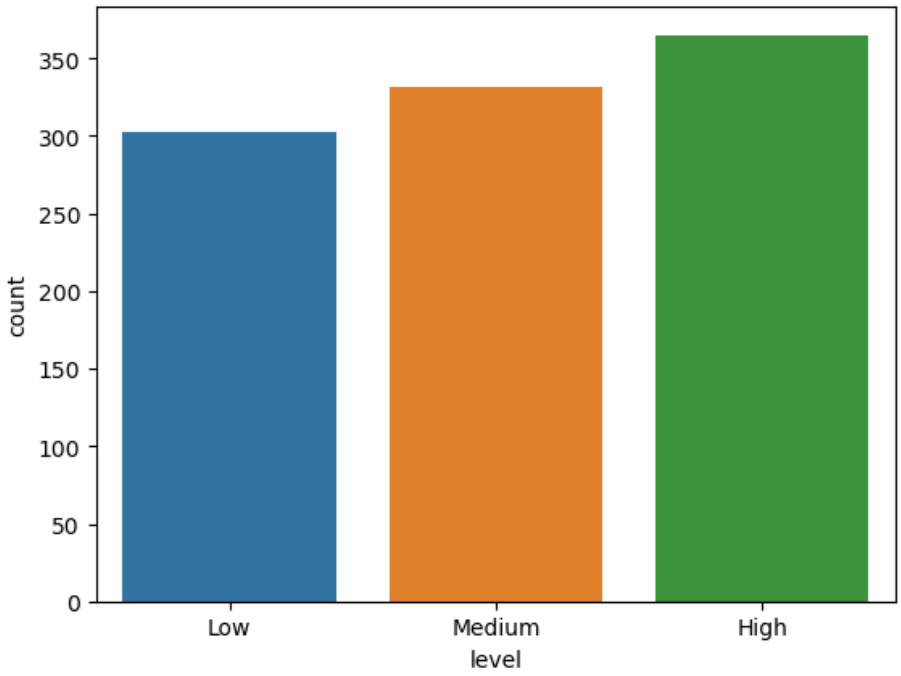
5. نتیجه گیری:

- بررسی مدل ها و مقایسه آنها

تحلیل و پیش پردازش داده ها

پس از انجام برخی پیش پردازش ها بر روی دیتا به بررسی ستون تارگت یعنی level میپردازیم.

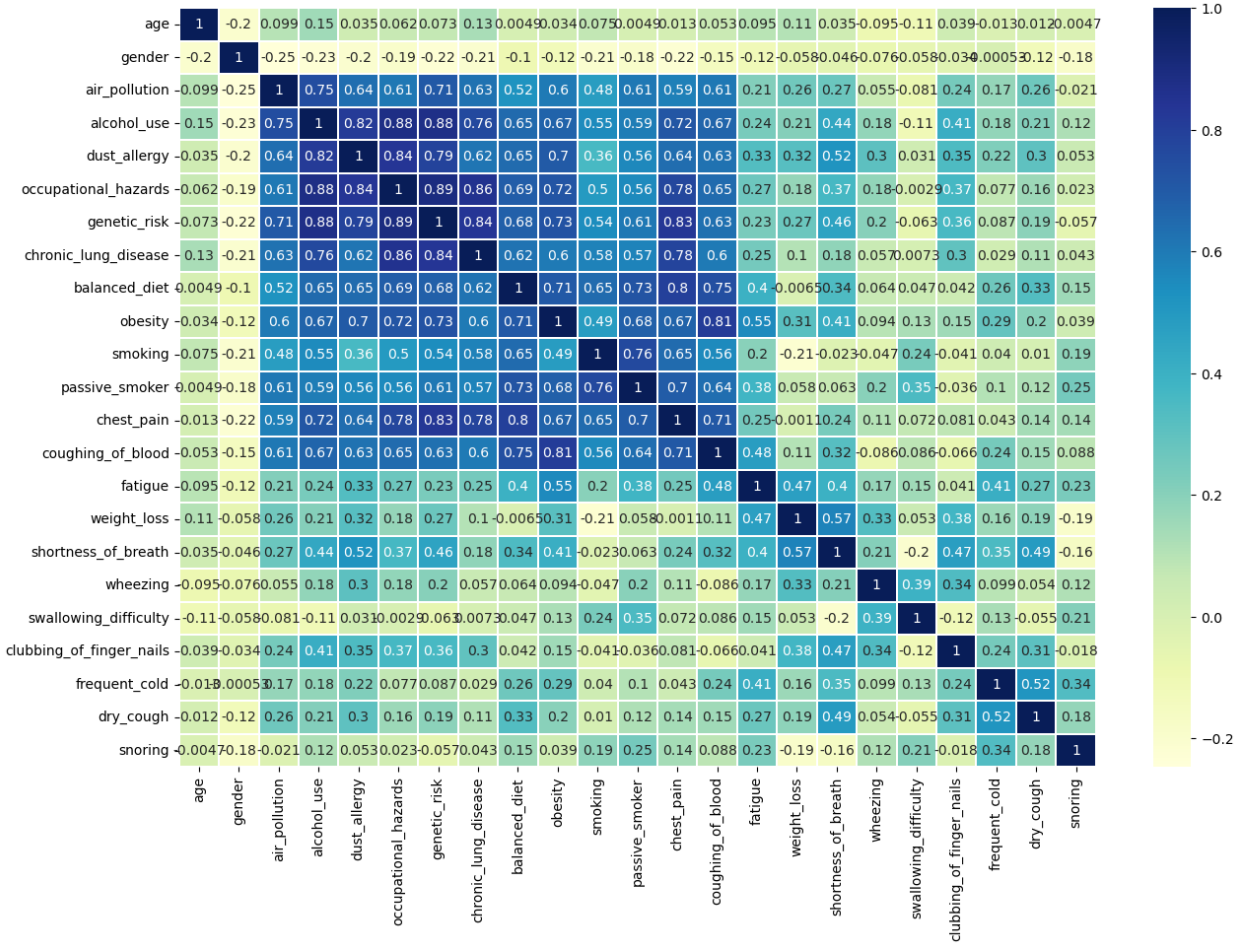
با پلات کردن این ستون مشاهده میکنیم که تعداد نمونه ها در آن تقریبا باهم برابرند و دیتا balance است.



باتوجه به اینکه بیشتر ویژگی های دیتاست دیتای categorical دارند، نمیشود خیلی به نتایج آماری تابع describe توجه کرد. اما با این حال میبینیم که بیشترین سن در این دیتاست ۷۳ سال است و میانگین سن افراد ۳۷ است،که این نشان میدهد که جامعه آماری نسبتا جوانی داریم.

	mean	std	min	25%	50%	75%	max
age	37.174000	12.005493	14.000000	27.750000	36.000000	45.000000	73.000000
gender	1.402000	0.490547	1.000000	1.000000	1.000000	2.000000	2.000000
air_pollution	3.840000	2.030400	1.000000	2.000000	3.000000	6.000000	8.000000
alcohol_use	4.563000	2.620477	1.000000	2.000000	5.000000	7.000000	8.000000
dust_allergy	5.165000	1.980833	1.000000	4.000000	6.000000	7.000000	8.000000
occupational_hazards	4.840000	2.107805	1.000000	3.000000	5.000000	7.000000	8.000000
genetic_risk	4.580000	2.126999	1.000000	2.000000	5.000000	7.000000	7.000000
chronic_lung_disease	4.380000	1.848518	1.000000	3.000000	4.000000	6.000000	7.000000
balanced_diet	4.491000	2.135528	1.000000	2.000000	4.000000	7.000000	7.000000
obesity	4.465000	2.124921	1.000000	3.000000	4.000000	7.000000	7.000000
smoking	3.948000	2.495902	1.000000	2.000000	3.000000	7.000000	8.000000
passive_smoker	4.195000	2.311778	1.000000	2.000000	4.000000	7.000000	8.000000
chest_pain	4.438000	2.280209	1.000000	2.000000	4.000000	7.000000	9.000000
coughing_of_blood	4.859000	2.427965	1.000000	3.000000	4.000000	7.000000	9.000000
fatigue	3.856000	2.244616	1.000000	2.000000	3.000000	5.000000	9.000000
weight_loss	3.855000	2.206546	1.000000	2.000000	3.000000	6.000000	8.000000
shortness_of_breath	4.240000	2.285087	1.000000	2.000000	4.000000	6.000000	9.000000
wheezing	3.777000	2.041921	1.000000	2.000000	4.000000	5.000000	8.000000
swallowing_difficulty	3.746000	2.270383	1.000000	2.000000	4.000000	5.000000	8.000000
clubbing_of_finger_nails	3.923000	2.388048	1.000000	2.000000	4.000000	5.000000	9.000000
frequent_cold	3.536000	1.832502	1.000000	2.000000	3.000000	5.000000	7.000000
dry_cough	3.853000	2.039007	1.000000	2.000000	4.000000	6.000000	7.000000
snoring	2.926000	1.474686	1.000000	2.000000	3.000000	4.000000	7.000000
level	2.062000	0.815365	1.000000	1.000000	2.000000	3.000000	3.000000

به علاوه باتوجه به اینکه تعداد دیتاپوینت ها بسیار محدود است (۱۰۰۰) به نتایج همبستگی نیز نمیتوان خیلی استنادکرد و ممکن است به دلیل ماهیت دیتا باشد؛ با این حال برخی از آنها قابل توجیه است برای مثال اینکه با افزایش آلودگی هوا حساسیت به گرد و غبار هم بیشتر شود کاملا منطقی است.



تقسیم داده به مجموعه های آموزش و آزمون

در این بخش داده های خود را به دو بخش میشکنیم؛ به این صورت که ۳۰ درصد از دیتا را برای تست کنار میگذاریم و مابقی را برای آموزش مدل نگه میداریم.

انتخاب مدل

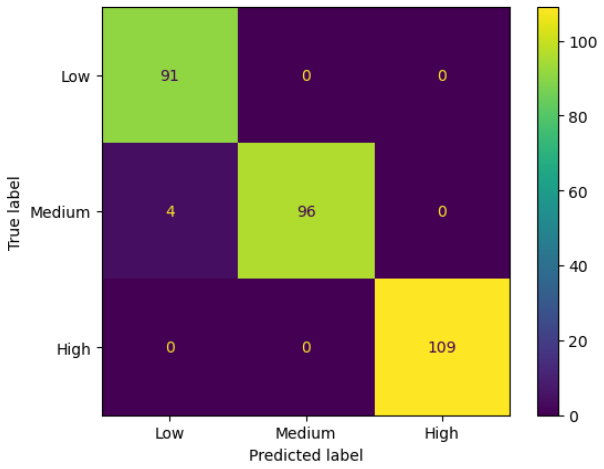
در این بخش به بررسی چند مدل مختلف میپردازیم و آنها را آموزش میدهیم و پس از آن به ارزیابی آنها میپردازیم.

- رگرسیون لجستیک

اولین مدلی که بررسی می‌کنیم،لاجستیک رگرشن است که مدل دقت خوبی دارد. جزئیات دقت آن با معیار های مختلف در زیر مشخص است.

```
Accuracy of LogisticRegression: 0.9866666666666667
Precision of LogisticRegression: 0.9866666666666667
Recall of LogisticRegression: 0.9866666666666667
F1 of LogisticRegression: 0.9866666666666668

                  Classification Report
              precision    recall  f1-score   support

           1       0.96      1.00      0.98        91
           2       1.00      0.96      0.98       100
           3       1.00      1.00      1.00       109

    accuracy                           0.99       300
   macro avg       0.99      0.99      0.99       300
weighted avg       0.99      0.99      0.99       300
```

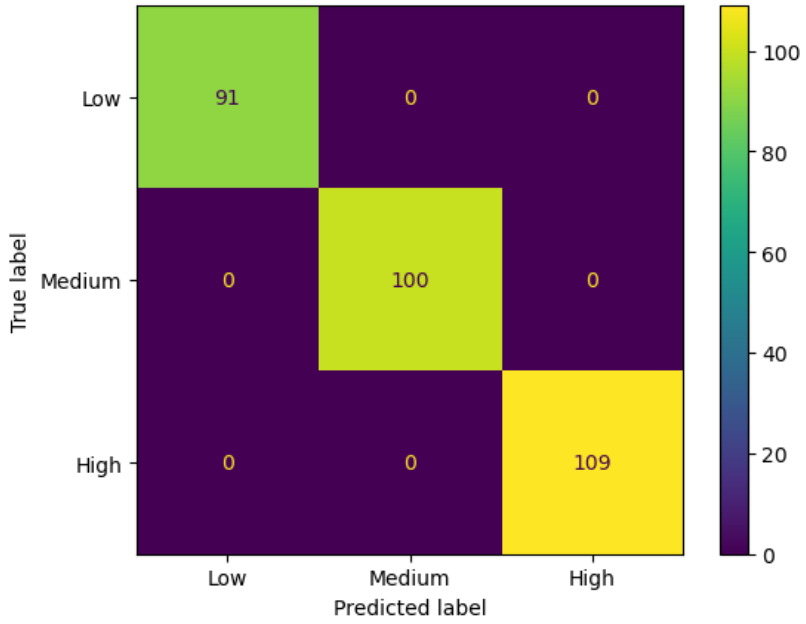
با بررسی ماتریس گمراهی مشاهده می‌کنیم که مدل در پیشبینی سطح medium اشتباه کرده و به اشتباه ۴ دیتاپوینت را سطح low پیشبینی کرده است.



- جنگل تصادفی

مدل بعدی را از بین مدل های ensemble انتخاب کردیم. میبینیم که مدل به خوبی قادر است که سطوح مختلف را با دقت بالا پیشبینی کند. باتوجه به اینکه پارامتر bootstrap در هایپرپارمتر های مدل مقدار True دارد مدل overfit نشده است. جزئیات دقت و ماتریس گمراهی به تفکیک در زیر آمده است.

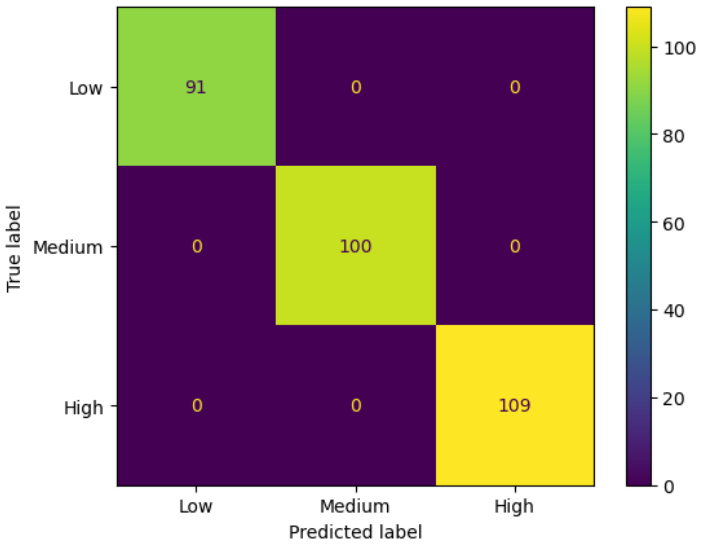
```
Accuracy of RandomForest: 1.0
Precision of RandomForest: 1.0
Recall of RandomForest: 1.0
F1 of RandomForest: 1.0

                Classification Report
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        91
           2       1.00      1.00      1.00       100
           3       1.00      1.00      1.00       109

    accuracy                           1.00       300
   macro avg       1.00      1.00      1.00       300
weighted avg       1.00      1.00      1.00       300
```



- شبکه عصبی چند لایه (MLP)

باتوجه به عملکرد خوب مدل جنگل تصادفی، میتوان انتظار داشت که نتیجه مشابه با یک شبکه عصبی multi layer perceptron تکرار خواهد شد.

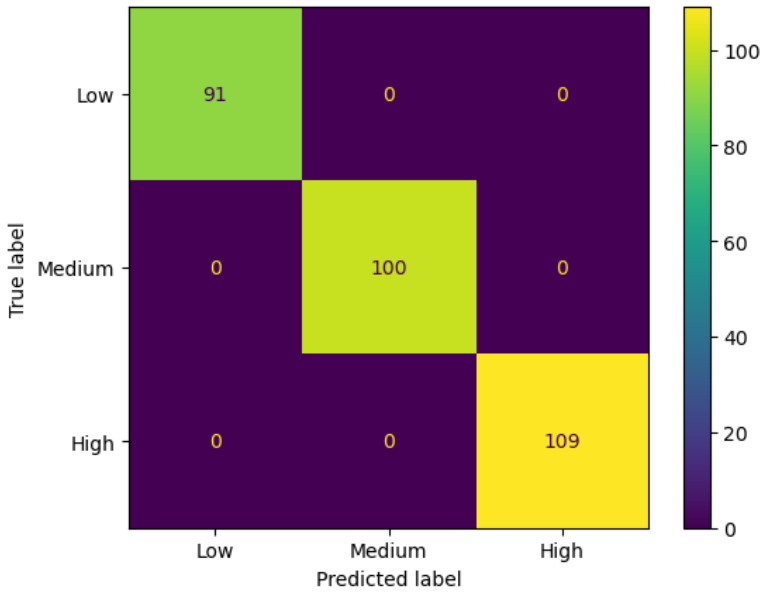
```
Accuracy of Multi-Layer-Perceptron: 1.0
Precision of Multi-Layer-Perceptron: 1.0
Recall of Multi-Layer-Perceptron: 1.0
F1 of Multi-Layer-Perceptron: 1.0

_________________Classification Report_________________
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        91
           2       1.00      1.00      1.00       100
           3       1.00      1.00      1.00       109

    accuracy                           1.00       300
   macro avg       1.00      1.00      1.00       300
weighted avg       1.00      1.00      1.00       300
```



- درخت تصمیم

با توجه به اینکه مدل جنگل تصادفی به راحتی توانست به دقت ۱۰۰ همگرا شود،این سوال به وجود می‌آید که آیا یک درخت تصمیم هم میتواند به تنهایی عملکرد خوبی داشته باشید؟ نتایج ارزیابی که این را نشان می‌دهد.

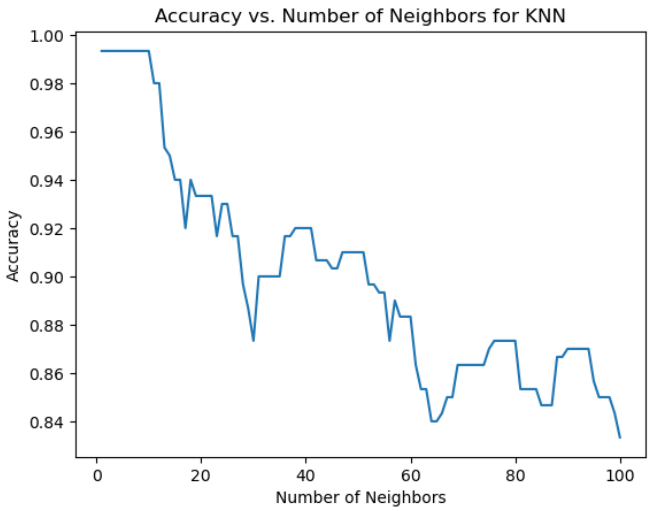
```
Accuracy of Decision Tree: 1.0
Precision of Decision Tree: 1.0
Recall of Decision Tree: 1.0
F1 of Decision Tree: 1.0

_________________Classification Report_________________
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        91
           2       1.00      1.00      1.00       100
           3       1.00      1.00      1.00       109

    accuracy                           1.00       300
   macro avg       1.00      1.00      1.00       300
weighted avg       1.00      1.00      1.00       300
```



k-نزدیک ترین همسایه (KNN) -

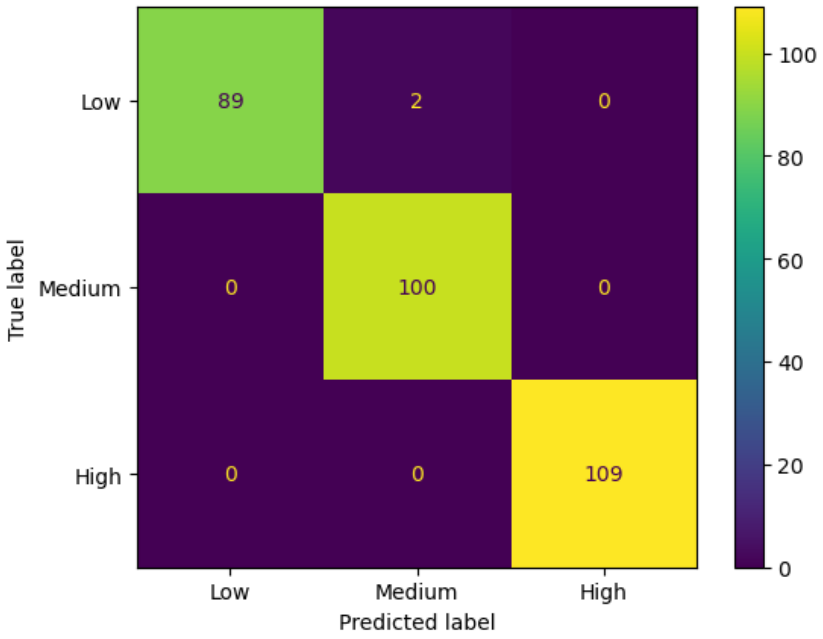
آخرین مدلی که بررسی می‌کنیم، knn خواهد بود. برای پیدا کردن بهترین تعداد همسایه ها مدل را ۱۰۰ بار به ازای مقادیر مختلف تست می‌کنیم؛ مشاهده می‌کنیم با افزایش تعداد همسایه ها دقت مدل متناوبا کاهش می‌یابد.

```
1, Accuracy of K-Nearest Neighbors: 0.9933333333333333
2, Accuracy of K-Nearest Neighbors: 0.9933333333333333
3, Accuracy of K-Nearest Neighbors: 0.9933333333333333
4, Accuracy of K-Nearest Neighbors: 0.9933333333333333
5, Accuracy of K-Nearest Neighbors: 0.9933333333333333
6, Accuracy of K-Nearest Neighbors: 0.9933333333333333
7, Accuracy of K-Nearest Neighbors: 0.9933333333333333
8, Accuracy of K-Nearest Neighbors: 0.9933333333333333
9, Accuracy of K-Nearest Neighbors: 0.9933333333333333
10, Accuracy of K-Nearest Neighbors: 0.9933333333333333
11, Accuracy of K-Nearest Neighbors: 0.98
12, Accuracy of K-Nearest Neighbors: 0.98
13, Accuracy of K-Nearest Neighbors: 0.9533333333333334
14, Accuracy of K-Nearest Neighbors: 0.95
15, Accuracy of K-Nearest Neighbors: 0.94
16, Accuracy of K-Nearest Neighbors: 0.94
17, Accuracy of K-Nearest Neighbors: 0.92
18, Accuracy of K-Nearest Neighbors: 0.94
19, Accuracy of K-Nearest Neighbors: 0.9333333333333333
20, Accuracy of K-Nearest Neighbors: 0.9333333333333333
```



پس یکی از مقادیر 1 تا ۸ را انتخاب میکنیم و مجدد مدل را آموزش می‌دهیم.

```
Accuracy of K-Nearest Neighbors: 0.9933333333333333
Precision of K-Nearest Neighbors: 0.9933333333333333
Recall of K-Nearest Neighbors: 0.9933333333333333
F1 of K-Nearest Neighbors: 0.9933333333333333

_________________Classification Report_________________
              precision    recall  f1-score   support

           1       1.00      0.98      0.99        91
           2       0.98      1.00      0.99       100
           3       1.00      1.00      1.00       109

    accuracy                           0.99       300
   macro avg       0.99      0.99      0.99       300
weighted avg       0.99      0.99      0.99       300
```



نتیجه گیری

باتوجه به نتایج حاصل شده در بخش قبل متوجه میشویم که بهترین مدل ها برای این مسئله درخت تصمیم، جنگل تصادفی و شبکه عصبی چند لایه هستند، هرچند دو مدل دیگر هم دقت بالایی داشتند اما این مدل ها به مقدار ۱۰۰ درصد دقت همگرا شدند. البته که باتوجه به اینکه دیتا بسیار دیتای تمیزی بود این مورد خیلی دور از ذهن هم نبود. باید توجه داشت که برای دستیابی به نتیاج قابل اتکا باید این مدل ها بر روی دیتاست ها بزرگ تر و صنعتی تست شوند و این دیتا صرفا برای مقاصد آموزشی بود.