گزارش پروژه تحلیل داده‌های بیماری مزمن کلیوی

دانشجو : مائده محمودی
شماره دانشجویی: 401463169

مقدمه

هدف این پروژه تحلیل داده‌های بیماران مبتلا و غیرمبتلا به بیماری مزمن کلیوی با استفاده از روش‌های یادگیری ماشین است

تحلیل اکتشافی داده‌ها(EDA)

Standard Scaler : ویژگی ها با استفاده از standardscalerاستاندارد سازی شدند تا برای الگوریتمهای یادگیری ماشین آماده شوند

ویژگی ها :

ویژگی های عددی شامل
sod,sc,bu,bgr,su,al,sg,bp,age,rbcc,wbcc,pcv,hemo,pot انتخاب شدند

ستون class_encoded به عنوان برچسب استفاده شده است(not ckd = 0 , ckd = 1)

توزیع کلاس ها : نمودار های باکس پلات برای دو گروه notCKD و CKD

تحلیل اکتشافی داده‌ها (EDA)

توزیع کلاس ها:

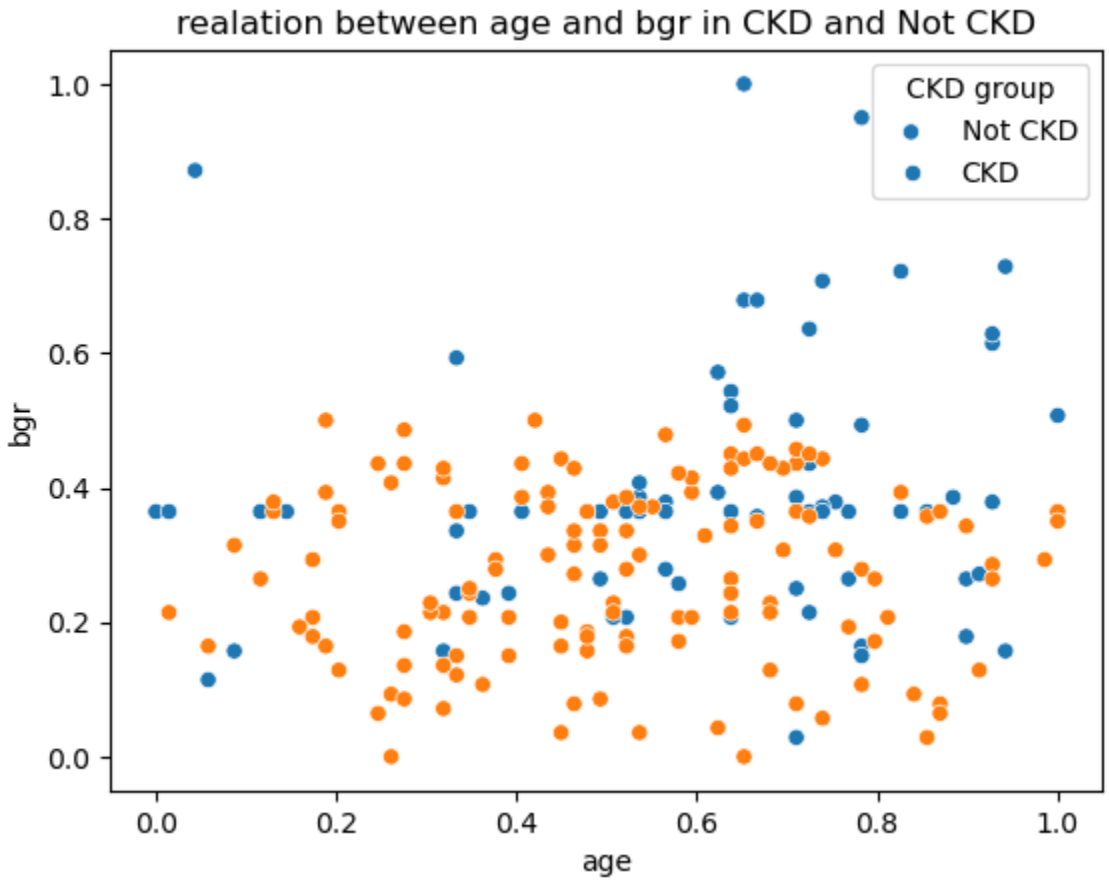
نمودارهای باکس پلات برای مقیاسه (ckd & notckd) با ویژگی های دیگر است

حقیقتا تعداد نمودارهاش زیاد بود به خاطر همین نمودارهای این قسمت رو توی گزارشکار نزاشتم ولی داخل کد میتونید مشاهده کنید

نمودار ckd و not ckd تقریبا متوازن است که برای مدل های طبقه بندی متناسب است

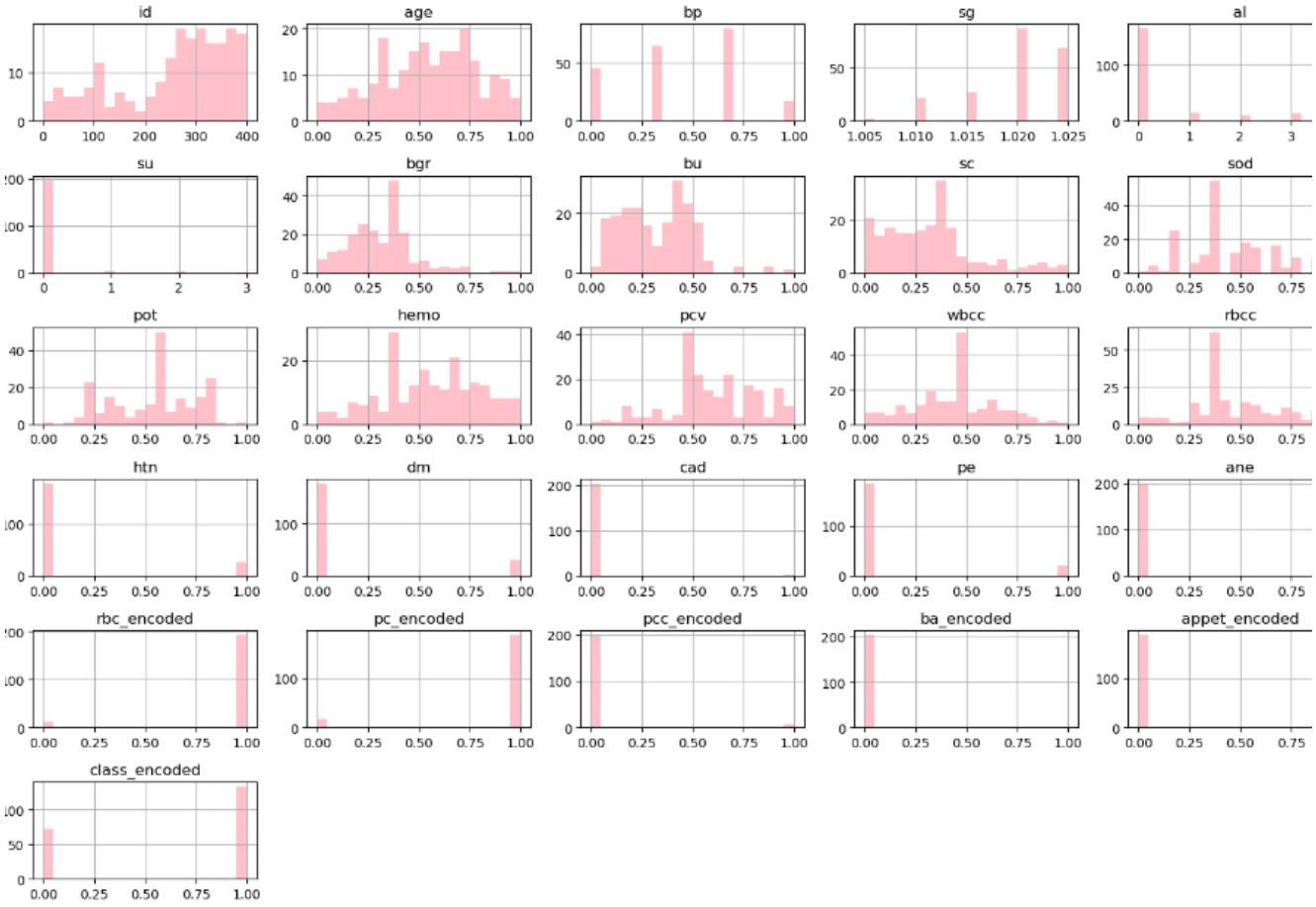
نمودار بعدی که اسکترپلات است برای نشان دادن پراکندگی رابطه ی بین سن و فشار خون (age & bgr)

در بیماران کلیوی و غیر کلیوی (ckd & not ckd)



نمودارهای بعدی که هیستوگرام هستند توزیع کلی ویژگی ها را نمایش دادند

این نمودارها نشان دهنده ی توزیع نرمال و توزیع غیر نرمال ویژگی هاست

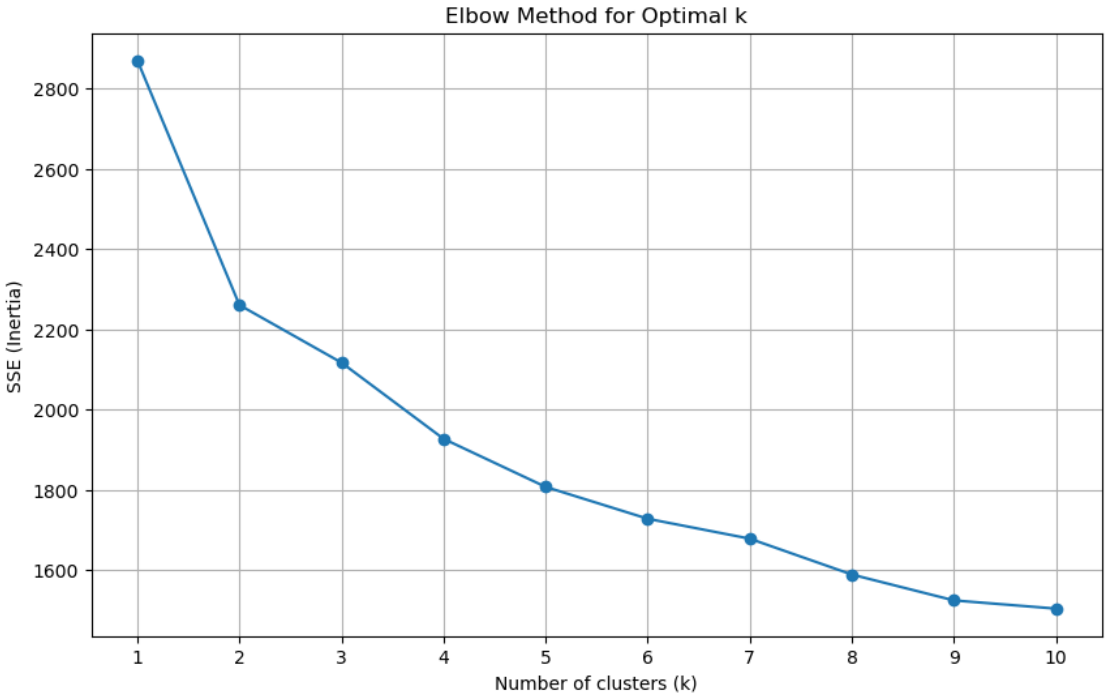


خوشه‌بندی

دو الگوریتم خوشه‌بندی استفاده شدند(k-means & dbscan)

- K-Means:

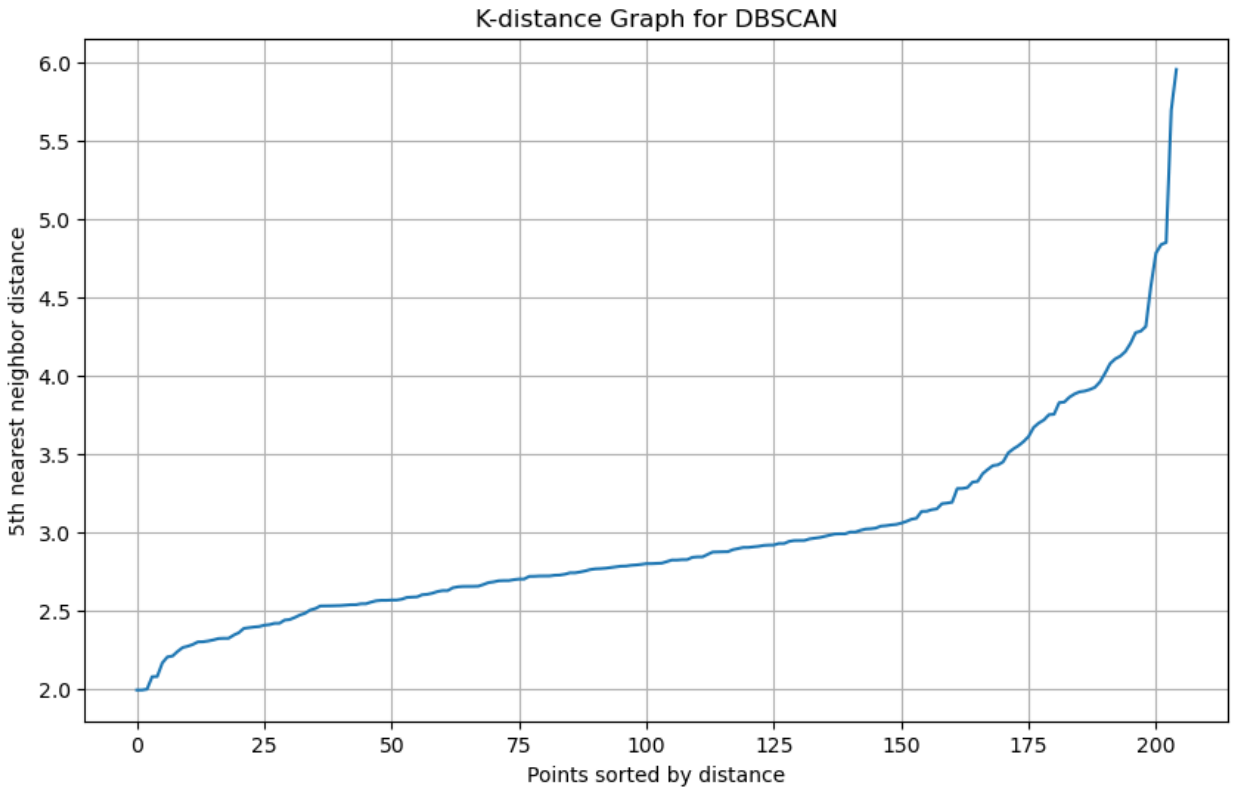
برای تعیین تعداد بهینه خوشه(Elbow) استفاده شده است که همانطور که مشاهده میکنیم
k مناسب برابر 2 (k = 2) پس نتیجه گیری که میتوانیم کنیم این است :
که تعداد خوشه های مناسب برابر 2 است



برای ارزیابی کیفیت خوشه ها از (Silhouette Score) استفاده شده
همانطور که در خروجی نشان داده شده کیفیت خروجی تایید شده

- DBSCAN:

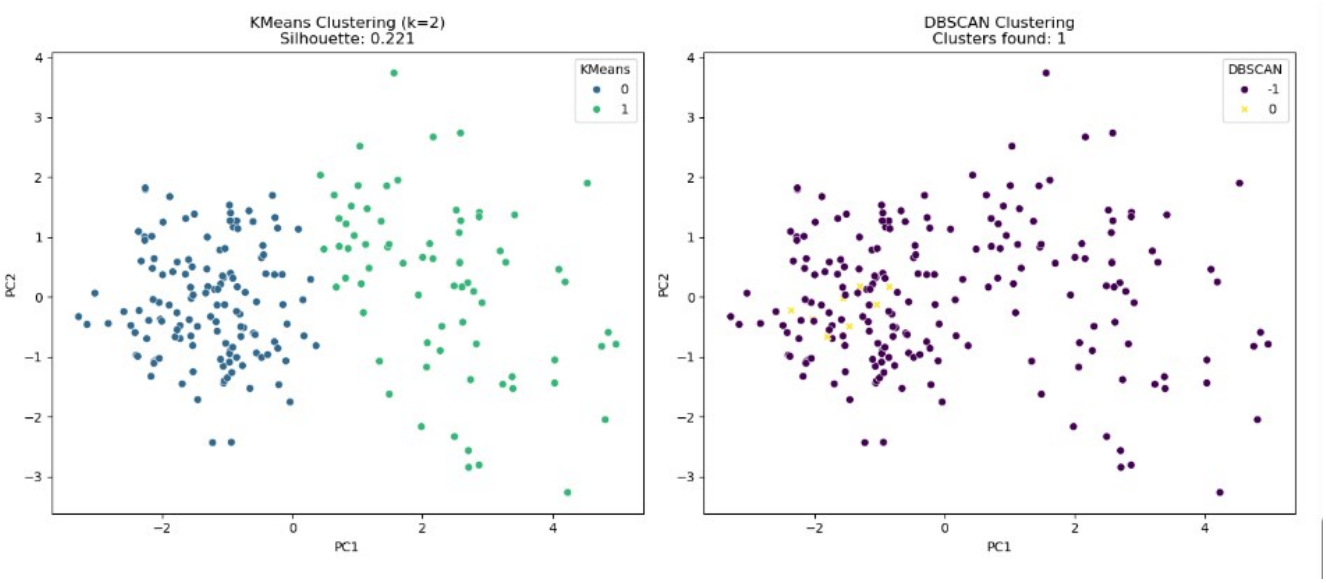
برای تعیین مقدار (esp & min_samples)از نمودار (k-distance)استفاده میکنیم
min_samples = 5 , esp = 0.5
همانطور که مشاهده میشود



سپس حدف نویز میکنیم در واقع در نظر گرفتن خوشه های پیدا شده

و سپس محاسبه ی (Silhouette Score) بدون در نظر گرفتن نقاط نویز

کاهش ابعاد (PCA) ایجاد دیتاگرام و رسم دو نمودار k-means & DBSCAN
مقایسه میانگین ویژگی ها در خوشه های K-means
و سپس مقایسه میانگین ویژگی ها در خوشه های dbscan



همانطور که مشاهده میشودdbscan برخی نقاط را به عنوان نویز شناسایی کرده است(1-)
که ممکن است به دلیل پراکندگی داده ها باشد

```
Mean features per KMeans cluster:
                     age        bp        sg        al        su       bgr  \
cluster_kmeans
0               0.509223  0.393939  1.022386  0.007576  0.000000  0.267695
1               0.614652  0.534247  1.015068  1.178082  0.205479  0.396477

                      bu        sc       sod       pot      hemo       pcv  \
cluster_kmeans
0               0.302642  0.235931  0.576263  0.549825  0.685424  0.701923
1               0.344398  0.499022  0.385388  0.518967  0.317627  0.400421

                    wbcc      rbcc
cluster_kmeans
0               0.404892  0.581930
1               0.486549  0.329274

Mean features per DBSCAN cluster:
                     age        bp      sg   al   su       bgr        bu  \
cluster_dbscan
0               0.530797  0.041667  1.0225  0.0  0.0  0.300893  0.208333

                      sc       sod       pot      hemo     pcv      wbcc  \
cluster_dbscan
0               0.279762  0.766667  0.697115  0.795181  0.5625  0.432229

                    rbcc
cluster_dbscan
0               0.606481
```

طبقه‌بندی

انتخاب سه مدل

svm.1

random forest.2

logistic regression.3

تقسیم داده ها : داده ها به نسبت 70 به 30 تقسیم شدند (70% test & 30% train)

5-fold cross-validation:

برای ارزیابی پایداری مدل ها محاسبه شده است

- Random Forest: 0.985
- SVM: 0.981
- Logistic Regression: 0.981

ارزیابی مدل‌های اولیه

مدل	Accuracy	Precision	Recall	F1-Score
Random Forest	0.968	0.976	0.976	0.976
SVM	0.984	0.977	1.000	0.988
Logistic Regression	0.968	1.000	0.952	0.976

تنظیم هایپرپارامترها:

از GridSearchCV برای بهینه سازی پارامترها استفاده شد

پارامتر ها :

- Random Forest:
  {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}،

- SVM:
  {'C': 10, 'kernel': 'rbf'},

- Logistic Regression:
  {'C': 0.1, 'solver': 'lbfgs'},

ارزیابی مدل‌های بهینه‌شده:

مدل	Accuracy	Precision	Recall	F1-Score
Random Forest (Tuned)	0.968	0.976	0.976	0.976
SVM (Tuned)	0.968	0.976	0.976	0.976
Logistic Regression (Tuned)	0.952	0.953	0.976	0.965

تحلیل:

svm در مدل اولیه بهترین عملکرد را داشت(Recall = 1.00) که نشان دهنده ی تشخیص کامل موارد است

مدل logistic regression بهبود قابل توجهی نشان نداده است و همچنین افت عملکرد داشته است زیرا f1-score کاهش یافته است

ROC: برای مدل های بهینه رسم شدند
منحنی ها مدل های بهینه شده را مقایسه میکند و مقادیرAUC بالا نشان دهنده عملکرد عالی مدل ها در تفکیک کلاس ها است

نتیجه‌گیری

این پروژه با موفقیت داده‌های بیماری مزمن کلیوی را تحلیل کرد

EDA:
تفاوت های کلیدی بین گروه های(CKD & notCKD) در ویژگی هایی مانندal و hemo, sg, شناسایی شد

k-means:
خوشه‌های معنی‌داری تولید کرد که با کلاس‌های واقعی هم‌خوانی داشت، در حالی که DBSCAN به دلیل نویز چالش‌هایی داشت

SVM:
بین مدل های بهبود یافته SVM بهترین عملکرد را داشت
F1-Score 0.988

پیوست‌ها

داده ها : Data.csv

کد پروژه: code.ipynb

نمودار roc : roc_curve