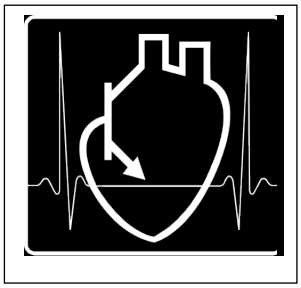
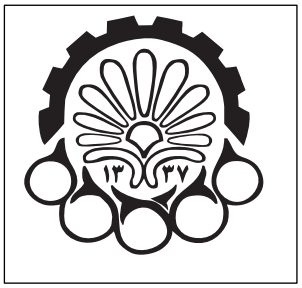




پروژه میانی ۲ درس یادگیری ماشین

استاد درس: دکتر بغدادی

آریان افشار

۴۰۱۳۳۰۰۴

مقدمه

کد کامل در پوشه به همراه تمام تصاویر و توضیحات مربوطه موجود است

تشخیص زودهنگام و دقیق سرطان سینه یکی از مهم ترین چالش‌های حوزه سلامت و پزشکی محسوب می شود. در این پروژه، هدف اصلی استفاده از مدل‌های مختلف طبقه بندی برای پیش بینی نوع تومور (بدخیم یا خوش خیم) در داده های مجموعه Breast Cancer Wisconsin (Diagnostic) می باشد.

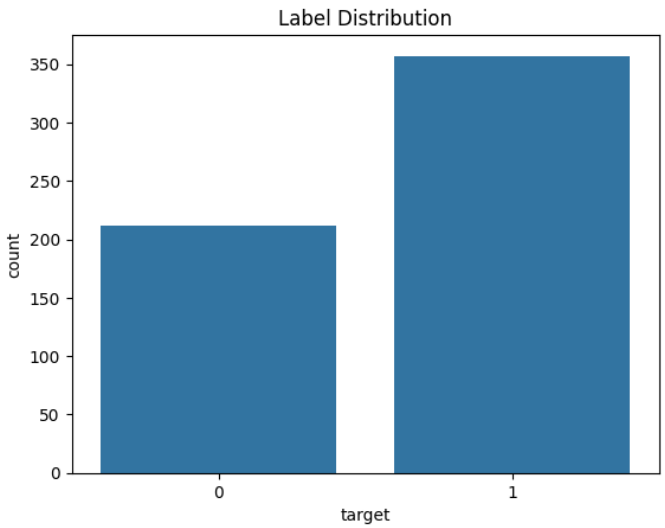
این دیتاست شامل مجموعه ای از ویژگی های عددی مرتبط با تصاویر دیجیتال از نمونه های زیستی بیماران است که شامل اندازه، شکل، بافت و دیگر خصوصیات هسته سلولی می شود. با بهره گیری از الگوریتم های مختلف ، تلاش شده است که برچسب نهایی تومور بر اساس این ویژگی‌ها با بیشترین دقت ممکن پیش بینی شود.

در پاسخگویی به انتظارات پروژه، مراحل زیر به طور دقیق انجام شده‌اند:

- پیاده‌سازی مدل های مختلف طبقه بندی و **مقایسه عملکرد آن ها** با استفاده از معیارهای ارزیابی گوناگون.
- **انتخاب ویژگی های مؤثر** با بهره گیری از پنج روش مختلف کاهش ابعاد و تحلیل اهمیت ویژگی ها.
- بررسی **تأثیر انتخاب ویژگی ها بر عملکرد مدل ها** و تغییر دقت پس از اعمال روش های انتخاب ویژگی.
- استفاده از **اعتبارسنجی متقابل (Cross Validation)** جهت سنجش پایداری و دقت نهایی مدل ها.
- تحلیل عملکرد مدل‌ها با متریک‌هایی همچون **F1 Score ،AUC-ROC، دقت، یادآوری، و ماتریس درهم ریختگی**.

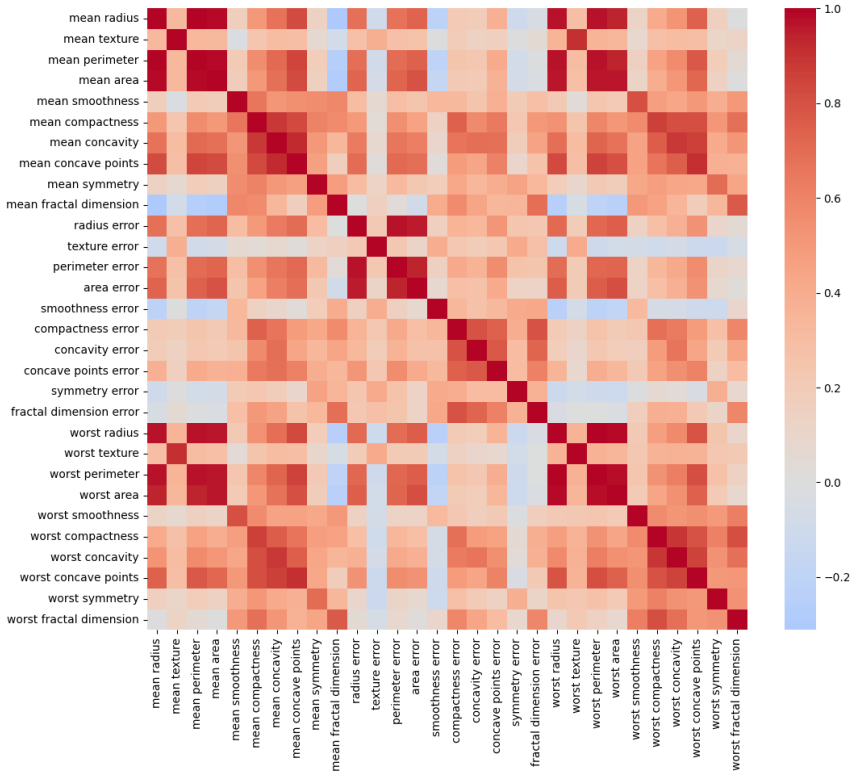
تحلیل کد

در ابتدا من تمام من کتابخانه های مورد نیاز را نوشتم و دیتا را بارگزاری کردم و سپس یک دید کلی از میزان توازن داده ها رسم کردم که به صورت زیر است:



همانطور که میبینیم حدود **۳۵۰ نفر سرطان** دارند و حدود **۲۱۰ نفر سالم.**

حال برای بررسی همبستگی هر ویژگی نسبت به هم, **confusion matrix** را رسم کردم که هر چه مقدار به سمت ۱ برود یعنی همبستگی بالایی دارد یعنی محتوا داده های ان ۲ شبیه به همند و ممکن است از کارامدی اطلاعات مفید کم کند, برای مثال worst radius , mean radius خیلی مقادیر شبیه به همی را دارند:



پس از آمادهسازی اولیه دادهها و حذف مقادیر نامعتبر یا نامناسب (در صورت وجود)، اولین گام اساسی در فرآیند یادگیری ماشین، تقسیم داده به دو بخش مجزای **دادههای آموزش (Training Set)** و **دادههای آزمون (Test Set)** است. این تقسیم بندی با هدف ارزیابی واقعی عملکرد مدلها انجام می شود تا از **بروز بیشبرازش (Overfitting)** جلوگیری شود و عملکرد مدل روی دادههای دیده نشده سنجیده شود.

از آنجایی که داده های پزشکی شامل ویژگی هایی با مقیاس های مختلف هستند (برای مثال، یک ویژگی ممکن است در بازه 0 تا 1 باشد و دیگری در بازه 0 تا 1000)، استفاده از **نرمالسازی استاندارد (Standard Scaling)** ضروری است. در این روش، هر ویژگی طوری تغییر می کند که میانگین آن برابر صفر و انحراف معیار آن برابر یک شود.و من در کد این کار را پس از تقسیم داده انجام دادم.

این نرمال سازی به ویژه برای مدل هایی نظیر SVM و Logistic Regression اهمیت زیادی دارد که به مقیاس عددی داده‌ها حساس هستند.

گام بعدی **پیاده سازی و ارزیابی مدل های طبقه بندی** در مرحله **اولیه** بدون انتخاب ویژگی است که من ابتدا این کار را انجام دادم و مدل ها را evaluate کردم و بعد طبق خواسته سوال با feature selection بررسی تأثیر انتخاب ویژگی ها بر عملکرد مدل ها و تغییر دقت پس از اعمال روش های انتخاب ویژگی را انجام دادم.

در این مرحله، مجموعه ای از مدل‌های طبقه بندی پرکاربرد بر روی داده های نرمال سازی شده اعمال گردید تا عملکرد آن ها در پیش‌بینی نوع تومور (بدخیم یا خوش‌خیم) مورد ارزیابی و مقایسه قرار گیرد. هدف از این تحلیل، بررسی دقت هر مدل، توانایی تفکیک کلاس‌ها، میزان خطای مدل و تحلیل نقاط قوت و ضعف هر الگوریتم می‌باشد. همچنین از کد های تمرین ۴ نیز کمک گرفتم.

۱ .مدل درخت تصمیم (Decision Tree)

مدل درخت تصمیم با معیار شاخص جینی (Gini Impurity) و استفاده از بهترین تقسیم کننده (splitter='best') پیاده سازی شد. همچنین عمق درخت به صورت کنترل‌شده (max_depth=4) و حداقل تعداد نمونه برای تقسیم (min_samples_split=100) تعیین گردید تا از بیش برازش جلوگیری شود.

ماتریس درهم ریختگی مربوط به این مدل برای تحلیل نوع خطا های طبقه بندی نیز ترسیم گردید:

دقت مدل (Accuracy) : به صورت عددی گزارش گردید.

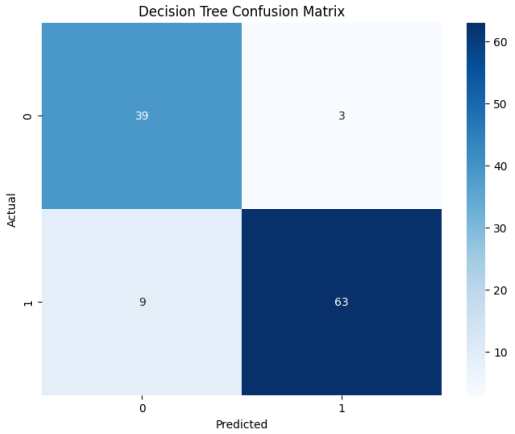
گزارش طبقه‌بندی (Classification Report) : شامل معیارهای دقت (Precision) ، یادآوری (Recall)، و امتیاز F1برای هر کلاس می باشد.

ماتریس درهم ریختگی: نشان می دهد مدل در پیش‌بینی هر کلاس چه میزان درست یا نادرست عمل کرده است.

```
Decision Tree Accuracy: 0.8947368421052632
              precision    recall  f1-score   support

           0       0.81      0.93      0.87        42
           1       0.95      0.88      0.91        72

    accuracy                           0.89       114
   macro avg       0.88      0.90      0.89       114
weighted avg       0.90      0.89      0.90       114
```



۲. سایر مدل ها

همین روند برای مدل های زیر نیز به‌کار گرفته شد:

مدل	تنظیمات مهم
Random Forest	تعداد درخت ها (n_estimators)، کنترل عمق و تصادفی بودن
Naive Bayes	مدل ساده سازی‌شده بر اساس توزیع گاوسی
SVM (Linear/Kernel)	استفاده از هسته های مختلف و تنظیم C
Logistic Regression	با استفاده از حل‌کننده liblinear و تنظیم L2
Bagging Classifier	مبتنی بر مدل پایه (اغلب درخت تصمیم) با نمونه گیری تصادفی
Ensemble (Voting Classifier)	تجمیع چندین مدل پایه با رأی گیری اکثریتی یا میانگین احتمال
LDA (Linear Discriminant Analysis)	مدلی خطی برای جداسازی کلاس ها بر پایه بیشینه سازی فاصله بین‌کلاسی

برای هر مدل موارد زیر محاسبه و گزارش شدند:

- دقت کلی (Accuracy)
- معیارهای گزارش طبقه بندی (Precision, Recall, F1-score)
- ماتریس درهم ریختگی
- تحلیل نقاط ضعف و قوت بر اساس خطاهای مشاهده شده

کار بعدی **تحلیل cross validation** بود که پروژه از ما خواسته بود.

اعتبارسنجی متقابل یا **Cross-Validation** روشی برای ارزیابی پایایی و دقت مدل های یادگیری ماشین است که به جای استفاده از یک بار تقسیم داده ها به آموزش و آزمون، داده ها را به چند بخش تقسیم کرده و چندین بار فرآیند آموزش و ارزیابی را تکرار می‌کند.

یکی از رایج ترین روش‌ها در این زمینه، **K-Fold Cross-Validation** است.

در این پروژه، از **Stratified K-Fold Cross-Validation** با تعداد پنج بخش استفاده شد. در این روش:

- داده ها به پنج بخش تقریباً مساوی تقسیم می شوند.
- در هر تکرار، یکی از این بخش ها به عنوان داده‌ی آزمون و چهار بخش دیگر به عنوان داده ی آموزش در نظر گرفته می شوند.
- این فرآیند پنج بار تکرار می شود، به گونه ای که هر بخش یک بار به عنوان داده ی آزمون انتخاب می‌شود.
- در نسخه‌ی Stratified، نسبت کلاس ها در هر بخش حفظ می شود (به‌خصوص در مسائل طبقه بندی که کلاس ها نامتوازن‌اند، این موضوع اهمیت زیادی دارد)

مزایای استفاده از Cross-Validation

1. **کاهش وابستگی به تقسیم خاص داده ها :**چون ارزیابی در چندین تقسیم مختلف انجام می شود، نتایج قابل اعتماد تر خواهند بود.
2. **برآورد دقیق‌تر دقت مدل :**میانگین نتایج اعتبارسنجی، نمایانگر عملکرد کلی مدل روی داده های نادیده گرفته شده است.
3. **کشف مدل های پایدارتر :**با محاسبه ی انحراف معیار نتایج، می توان مدل هایی را انتخاب کرد که نه تنها دقت بالاتری دارند، بلکه عملکرد پایدارتری نیز در مواجهه با تغییر داده دارند.

در این پروژه، اعتبارسنجی متقابل روی داده‌های آموزش یافته با استفاده از مدل های مختلفی نظیر درخت تصمیم، جنگل تصادفی، بیز ساده، ماشین بردار پشتیبان، رگرسیون لجستیک، تحلیل افتراقی خطی و مدل های ترکیبی انجام شد. برای هر مدل، **میانگین دقت (Accuracy Mean)** و **انحراف معیار (Standard Deviation)** محاسبه شد تا هم دقت و هم پایداری مدل در شرایط مختلف سنجیده شود.

	Mean Accuracy	Std
LogisticRegression	0.978022	0.009829
SVM	0.964835	0.016150
RandomForest	0.962637	0.017855
LDA	0.960440	0.017855
Bagging	0.949451	0.011207
NaiveBayes	0.934066	0.028656
DecisionTree	0.920879	0.008223

حال با استفاده از دقت کلی, معیارهای گزارش طبقه بندی و اعتبارسنجی متقابل, دقت مدل ها را به ترتیب عالی به بد سورت کردم که نتایج به صورت زیر شد:

1. رگرسیون لجستیک(Logistic Regression)

بهترین عملکرد را داشت. سادگی، قابلیت تفسیر، و سازگاری با داده‌های استاندارد شده از دلایل اصلی موفقیت آن است.

2. ماشین بردار پشتیبان(SVM)

دقت بالا و پایداری مناسبی داشت. برای داده‌های با ویژگی های قابل تفکیک بسیار مناسب است.

3. جنگل تصادفی(Random Forest)

عملکرد خوب و مقاوم در برابر بیش برازش. اما به‌دلیل پیچیدگی بیشتر، کمی پایین تر از مدل های خطی قرار گرفت.

4. تحلیل افتراقی خطی(LDA)

عملکرد قابل قبول، ولی وابسته به فرض‌های آماری خاص (نرمال بودن داده‌ها و کوواریانس یکسان کلاس‌ها)

5. بگینگ(Bagging)

عملکرد متوسط، با کاهش واریانس از طریق ترکیب مدل‌ها، اما بدون بهبود چشم گیر در این مسئله خاص.

6. بیز ساده(Naive Bayes)

به دلیل فرض استقلال ویژگی ها، نتوانست روابط میان ویژگی ها را به‌درستی مدل کند و عملکرد پایین‌تری داشت.

7. درخت تصمیم(Decision Tree)

ضعیف‌ترین عملکرد به دلیل تنظیمات محدود کننده (عمق کم و حداقل نمونه زیاد)، که باعث کاهش قدرت مدل شد.

مدل‌های **لجستیک رگرشن و SVM** برای این قسمت پروژه بهترین انتخاب بودند.

حال میخواهیم با استفاده از چندین روش **feature selection** بهترین ویژگی ها را انتخاب کنیم و اینبار روی ان ها مدل ها را پیاده سازی کنیم.

یکی از مراحل کلیدی در طراحی مدل های یادگیری ماشین، به ویژه در مسئله های پزشکی مانند تشخیص سرطان، انتخاب صحیح ویژگی‌ها (Feature Selection) است. وجود ویژگی های غیرضروری، تکراری یا دارای ارتباط ضعیف با خروجی می تواند باعث کاهش دقت مدل، افزایش پیچیدگی و ایجاد بیش برازش شود. از این رو، در این پروژه پنج روش معتبر برای انتخاب ویژگی به کار گرفته شد و نتایج آن‌ها با هم ترکیب شدند تا مجموعه‌ای از ویژگی‌های مؤثر شناسایی شود.

۱. اطلاعات متقابل(Mutual Information)

این روش میزان اطلاعات مشترک بین هر ویژگی و متغیر هدف را اندازه گیری می‌کند و قادر به شناسایی وابستگی های غیرخطی است. برخلاف روش‌های آماری کلاسیک، اطلاعات متقابل نیازمند فرض نرمال بودن داده‌ها نیست و می تواند روابط پیچیده و پنهان بین ویژگی ها و خروجی را آشکار سازد.
مزیت کلیدی :قدرت تشخیص روابط غیرخطی و مستقل بودن از مدل.

۲. حذف بازگشتی ویژگی‌ها (RFE) با استفاده از رگرسیون لجستیک

در این روش، با استفاده از یک مدل پایه (در اینجاLogistic Regression) ویژگی ها به‌صورت تدریجی و بر اساس تأثیرشان بر دقت مدل حذف می شوند. ابتدا مدل با تمام ویژگی ها آموزش می بیند و سپس در هر مرحله ضعیف ترین ویژگی‌ها (با وزن یا اهمیت کمتر) حذف می شوند تا در نهایت تنها مهم ترین ویژگی ها باقی بمانند.
مزیت کلیدی :لحاظ کردن تعامل بین ویژگی‌ها در ساختار یک مدل پیش‌بینی‌کننده.

۳ .آزمون آماری ANOVA با معیارChi-Square

روش SelectKBest همراه با آزمون خی دو برای محاسبه رابطه آماری بین هر ویژگی و برچسب خروجی استفاده شد. این روش برای داده های طبقه‌ای و گسسته مناسب است. به دلیل نیاز این روش به داده های غیرمنفی، ابتدا ویژگی ها با مقیاس‌گذار MinMax نرمال سازی شدند.
مزیت کلیدی :ساده، سریع و مناسب برای ارزیابی وابستگی آماری بین ویژگی ها و خروجی.

۴ .اهمیت ویژگی‌ها در مدل جنگل تصادفی(Random Forest Feature Importance)

جنگل تصادفی با استفاده از ساختار مجموعه ای از درخت‌های تصمیم گیری، میزان مشارکت هر ویژگی در کاهش معیار عدم قطعیت (مانند Gini یا Entropy) را اندازه گیری می کند. خروجی این روش، امتیاز اهمیت برای هر ویژگی است که به‌صورت مستقیم از مدل استخراج می شود.
مزیت کلیدی :در نظر گرفتن روابط پیچیده و غیرخطی بین ویژگی ها و قابلیت تفسیر مدل.

۵ .انتخاب ویژگی از طریق مدل لجستیک با جریمه(L1 (Lasso

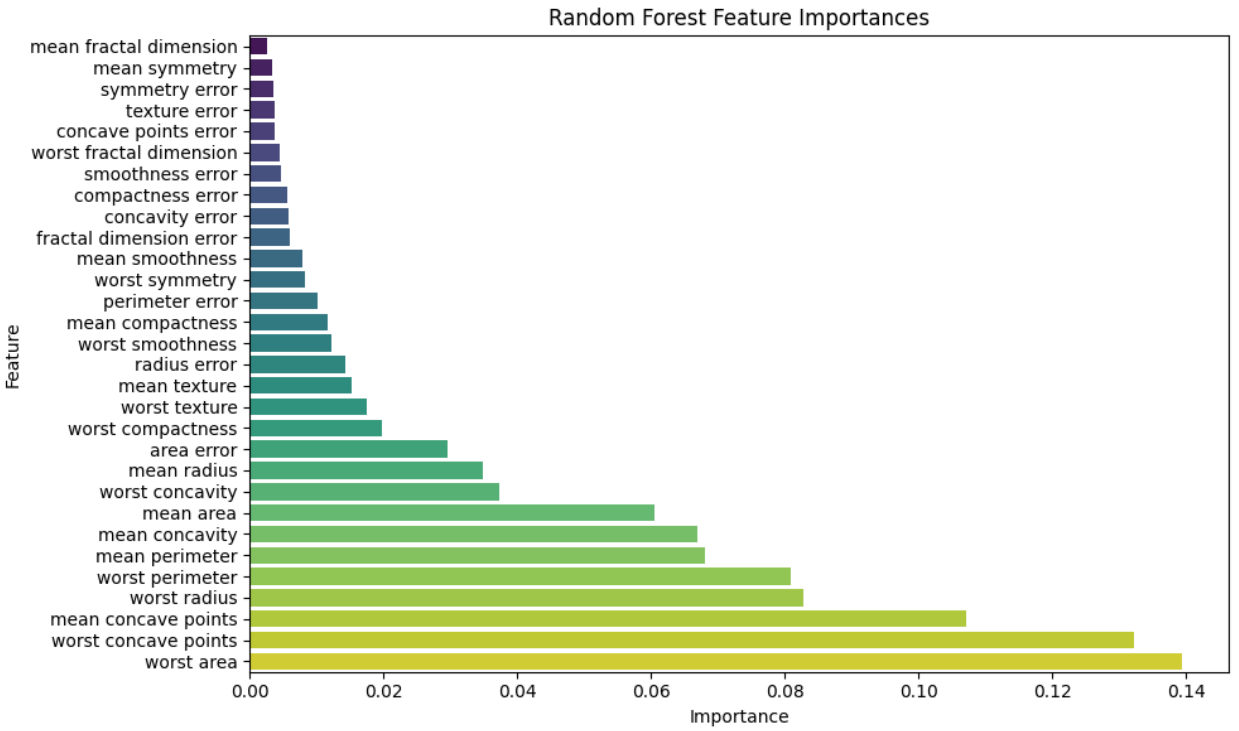
در این روش، مدل Logistic Regression با جریمه L1 آموزش داده شد که باعث صفر شدن ضرایب ویژگی های غیرمؤثر می شود.سپس تنها ویژگی هایی که ضریب آن‌ها غیر صفر باقی مانده بود، به‌عنوان ویژگی‌های منتخب حفظ شدند.
مزیت کلیدی :انتخاب خودکار ویژگی‌ها در ساختار مدل و کاهش پیچیدگی به‌کمک تنبیه.

پس از اجرای هر پنج روش، لیستی از ۱۰ ویژگی برتر از هر روش استخراج شد. در ادامه با ترکیب همه لیست ها و شمارش تعداد دفعات تکرار هر ویژگی در بین روش‌ها، ۲۰ ویژگی نهایی انتخاب شدند. این فرآیند که نوعی رأی گیری چند مرحله ای محسوب می‌شود، باعث افزایش پایداری انتخاب ویژگی ها شده و دقت نهایی مدل ها را ارتقاء می‌دهد.

استفاده از ترکیب روش‌های متنوع – از روش‌های آماری مانند chi2 تا روش‌های مدل محور مانند RFE و – L1 تضمین می‌کند که مجموعه ویژگی های نهایی، هم از نظر آماری و هم از نظر تجربی، بیشترین تأثیر را در پیش‌بینی برچسب «بدخیم» یا «خوش‌خیم» بودن تومور دارند.

```
✅ Top 20 Final Selected Features Based on Voting Across 5 Methods:
1. worst concave points
2. mean concave points
3. worst radius
4. worst area
5. mean concavity
6. worst perimeter
7. worst concavity
8. mean area
9. mean radius
10. mean perimeter
11. area error
12. radius error
13. compactness error
14. worst texture
15. mean fractal dimension
16. texture error
17. smoothness error
18. fractal dimension error
19. worst smoothness
20. worst symmetry
```

در ادامه من برای هر ۵ روش نمودار را نیز رسم کردم که در کد موجود است. برای مثال برای random forest: نمودار زیر رسم شد که نشان داد worst concave point , worst area بهترین ویژگی ها بر اساس این روش بود و برای بقیه روش ها هم به همین ترتیب رسم شد.



همچنین به توضیح راجع به چند ویژگی برتر نیز میپردازم.

worst concave points

نشان دهنده بیشترین میزان فرورفتگی در مرز تومور است. فرورفتگی های زیاد اغلب با رشد غیر طبیعی سلول ها مرتبط هستند و یکی از قوی ترین شاخص ها برای تومور بدخیم به شمار می رود.

mean concave points

میانگین تعداد نقاط فرورفته در مرز سلول. مانند ویژگی قبلی، این پارامتر نیز با نامنظم بودن مرز سلول و احتمال بدخیمی مرتبط است.

worst perimeter

بزرگ‌ترین مقدار محیط ثبت شده از مرز تومور. تومورهای بدخیم اغلب مرزهای بزرگ تر و نامنظم تری دارند.

worst radius

بزرگ ترین شعاع ثبت شده در نمونه. این ویژگی می تواند نشان دهنده اندازه بزرگ تر تومورهای بدخیم باشد.

mean perimeter

میانگین محیط سلول در نمونه‌گیری ها. محیط بزرگ تر می تواند حاکی از ساختار غیرطبیعی و افزایش رشد سلولی باشد.

mean radius

میانگین شعاع تومور. اندازه کلی تومور یکی از پارامترهای مهم در تعیین نوع آن است.

worst area

بیشترین مقدار مساحت سلول. تومورهای بدخیم معمولاً دارای سلول هایی با اندازه بزرگ تر هستند.

mean area

میانگین مساحت نمونه های سلولی. افزایش مساحت با تهاجمی بودن تومور مرتبط است.

worst texture

واریانس در شدت پیکسل ها در بزرگ ترین ناحیه تصویر. ناهمگنی بیشتر معمولاً در تومورهای بدخیم دیده می شود.

mean concavity

میانگین میزان انحنای فرو رفته در مرز تومور. انحناهای بیشتر معمولاً از رشد غیرطبیعی و غیر متقارن ناشی می شوند

حال که لیستی از ۲۰ ویژگی برتر داریم مدل ها را بر این اساس ارزیابی میکنیم که دوباره به دو بخش آزمون و آموزش تقسیم و آن ها را استاندارد میکنیم.
پس از اینکه مدل ها را روی آن ها پیاده سازی کردم به نتایج زیر رسیدم که AUC را که برابر با مساحت زیر منحنی ROC است و بین ۰ و ۱ هست را نیز قرار دادم.
خود ROC نموداری است که True Positive Rate حساسیت را در مقابل False Positive Rate نرخ خطای مثبت کاذب در مقادیر مختلف آستانه (threshold) قرار می دهد. به عبارتی نشان می دهد که مدل چگونه در تشخیص نمونه‌های مثبت و منفی عمل می کند.
در این مرحله به نتایج زیر رسیدیم:

Model: DecisionTree (با 20 ویژگی انتخاب‌شده)

	precision	recall	f1-score	support
0	0.93	0.86	0.89	43
1	0.92	0.96	0.94	71
accuracy			0.92	114
macro avg	0.92	0.91	0.91	114
weighted avg	0.92	0.92	0.92	114

AUC: 0.9091057975761546

Model: RandomForest (با 20 ویژگی انتخاب‌شده)

	precision	recall	f1-score	support
0	0.95	0.93	0.94	43
1	0.96	0.97	0.97	71
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

AUC: 0.9510317720275139

Model: NaiveBayes (با 20 ویژگی انتخاب‌شده)

	precision	recall	f1-score	support
0	1.00	0.93	0.96	43
1	0.96	1.00	0.98	71
accuracy			0.97	114
macro avg	0.98	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

AUC: 0.9651162790697674

Model: SVM (با 20 ویژگی انتخاب‌شده)

	precision	recall	f1-score	support
0	0.98	0.98	0.98	43
1	0.99	0.99	0.99	71
accuracy			0.98	114
macro avg	0.98	0.98	0.98	114
weighted avg	0.98	0.98	0.98	114

AUC: 0.9813298395021289

Model: Bagging (با 20 ویژگی انتخاب‌شده)

	precision	recall	f1-score	support
0	0.95	0.93	0.94	43
1	0.96	0.97	0.97	71
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

AUC: 0.9510317720275139

Model: LogisticRegression (با 20 ویژگی انتخاب‌شده)

	precision	recall	f1-score	support
0	0.98	0.98	0.98	43
1	0.99	0.99	0.99	71
accuracy			0.98	114
macro avg	0.98	0.98	0.98	114
weighted avg	0.98	0.98	0.98	114

AUC: 0.9813298395021289

Model: LDA (با 20 ویژگی انتخاب‌شده)

	precision	recall	f1-score	support
0	0.98	0.93	0.95	43
1	0.96	0.99	0.97	71
accuracy			0.96	114
macro avg	0.97	0.96	0.96	114
weighted avg	0.97	0.96	0.96	114

AUC: 0.9580740255486406

حال دوباره از cross validation استفاده کردم و به نتایج زیر رسیدم:

	Mean Accuracy	Std
SVM	0.978022	0.006950
LogisticRegression	0.978022	0.006950
LDA	0.962637	0.014906
RandomForest	0.953846	0.025441
Bagging	0.947253	0.032894
NaiveBayes	0.931868	0.012815
DecisionTree	0.909890	0.035027

این بار نیز SVMو Logistic Regression بهترین مدل طبقه بندی بودند .

بررسی عملکرد مدل ها پیش و پس از انتخاب ویژگی ها نشان می دهد که تأثیر این فرایند بر دقت و پایداری مدل‌ها متغیر است.

مدل **Logistic Regression** پس از انتخاب ویژگی ها، همان دقت قبلی خود **(٪۹۷/۸)** را حفظ کرد، اما با کاهش انحراف معیار از ۰/۰۰۹۸ به ۰/۰۰۶۹، پایداری بیشتری در اعتبارسنجی متقابل نشان داد. این پایداری به این معناست که مدل در فولد های مختلف عملکرد تقریباً یکسانی داشته و نسبت به ویژگی های اضافی حساسیت کمتری دارد.

مدل **SVM** نیز از انتخاب ویژگی به طور محسوسی سود برد. دقت آن از **٪۹۶/۴ به ٪۹۷/۸** افزایش یافت و انحراف معیار آن نیز کاهش پیدا کرد. این نشان می دهد که انتخاب ویژگی باعث تمرکز مدل روی اطلاعات مهمتر شده و قابلیت تعمیم آن بهبود یافته است.

در مقابل، عملکرد **Random Forest** با کمی کاهش همراه بود؛ دقت آن از **٪۹۶/۲ به ٪۹۵/۳** رسید. با توجه به اینکه این مدل ذاتاً از اهمیت ویژگی ها برای ساخت درخت ها استفاده می‌کند، کاهش دقت می تواند ناشی از حذف برخی ویژگی های مکمل باشد که در کنار هم اطلاعات مؤثری تولید می کردند. همچنین انحراف معیار بیشتر نیز کاهش پایداری را نشان می دهد.

LDA عملکرد نسبتاً ثابتی داشت و دقت آن کمی افزایش یافت. این مدل که مبتنی بر فرض توزیع نرمال ویژگی هاست، از انتخاب ویژگی هایی که با این فرض سازگارتر هستند، بهره مند شد و دقت آن از **٪۹۶/۰ به ٪۹۶/۲** رسید.

برای مدل **Bagging** ، تغییرات چشمگیری مشاهده نشد؛ اما دقت اندکی کاهش و انحراف معیار کمی افزایش یافت که نشان‌دهنده نوسانات بیشتر در بین فولد هاست. از آنجا که این روش ترکیبی است و به مدل پایه درخت تصمیم وابسته است است، اثر انتخاب ویژگی محدودتر بوده است.

مدل **Naive Bayes** نیز کمی افت دقت را تجربه کرد. این مدل بر اساس فرض استقلال ویژگی ها عمل می کند و بنابراین حذف برخی ویژگی‌های مرتبط می تواند ساختار اطلاعاتی آن را برهم بزند. با این حال، کاهش انحراف معیار نشانه ای از پایداری بیشتر است.

در نهایت، مدل **Decision Tree** بیشترین افت عملکرد را تجربه کرد. دقت آن از **٪۹۲/۰ به ٪۹۰/۹** کاهش یافت و انحراف معیار نیز افزایش یافت. این مدل برای تقسیم بندی دقیق به مجموعه ای کامل از ویژگی ها نیاز دارد و حذف برخی از آن‌ها باعث ساده سازی بیش از حد ساختار درخت و کاهش توان تفکیک مدل شده است.

در مجموع، **انتخاب ویژگی تأثیر مثبتی بر مدل‌های خطی مانند Logistic Regression و SVM** داشته است، در حالی که **برای مدل‌هایی با ساختار پیچیده‌تر و وابستگی داخلی بیشتر مانندDecision Tree و Random Forest**، اثر آن در برخی موارد **منفی** بوده است.