**Breast Cancer Data Analysis**

دیتای مورد بررسی مربوط به نمونه 569 تایی از داده های سرطان سینه می باشد. دیتا دارای 29 فیچر پیوسته و 1 تارگت گسسته می باشد.

# (EDA) Exploratory Data Analysis

تحلیلی اکتشافی ، جهت آماده سازی داده ها و ایجاد فایل پروژه، بارگذاری کتابخانه‌ها ، جمع آوری داده، پاکسازی داده، آمار توصیفی، بررسی بصری داده‌ها و پاکسازی داده های پرت بکاربرده شده است.

طبق بررسی انجام شده دیتا فاقد داده تکراری و از دست رفته و غیرعددی بوده است. فیچرها نیاز به ترکیب یا جداسازی نداشتند. از آنجاییکه کلیه فیچرها عددی می باشند نیاز به هندل کردن داده های کتگوریکال وجود نداشت.

# Splitting the Dataset into the Training set, Test set, and Validation set

بعد از انجام مراحل فوق، داده ها برای یادگیری به نسبت 70 % داده آموزش دیده و 30% تست تقسیم بندی شدند . از آنجایی که بعضی از الگوریتمهای مورد استفاده در این دیتا دارای قدرت زیادی نیستند به نسبت 10% از داده های آموزش دیده و تست برای اعتباربخشی جدا شده اند.

# Handle Outlier Data

بر طبق نمودار Box Plot فیچرهای , **area\_mean worst\_ area** دارای داده ناصحیح می باشند ولی به دلیل پایین بودن نسبت داده های ناصحیح به کل داده ها از حذف آن ها چشم پوشی شده است.

# Feature Selection

دیتای مورد بررسی از نوع پیش بینی و دارای فیچرهای مستقل عددی و تارگت کتگوریکال است، برای انتخاب فیچر ها دو شرط لازم و کافی در نظر گرفته شده است.

* شرط لازم، سنجش میزان همبستگی بین فیچرهای مستقل با فیچر هدف است. در این دیتا از مدل anova استفاده شده است، فیچرهای زیر که کمترین امتیاز را دریافت کردند، کمترین ارتباط را با فیچر هدف دارند و از دیتا حدف شدند.

# "texture\_mean","fractal\_dimension\_mean","texture\_se","smoothness\_se","symmetry\_se","fractal\_dimension\_se"

* شرط کافی شناسایی همبستگی بین فیچرهای مستقل بوسیله ماتریس کورلیشن و تصمیم گیری در خصوص حذف یا نگه داشتن بالاترین همبستگی هاست.
* فیچرهای با خود همبستگی بالای 90% شامل:

# radius\_mean texture\_mean perimeter\_mean area\_mean , radius\_worst, area\_worst, texture\_worst , area\_worst

* تنها فیچر texture\_mean که با ماتریس کورلیشن بدست آمده و با فیچرهای انتخاب شده با روش anova مشترک است از شرط کافی حذف میشود.

Handle Imbalanced Data

با توجه به اینکه تعداد نمونه های آموزش دیده انتخاب شده برای کلاسهای هدف نامتوازن است بنابراین نیاز به یکسان کردن تعداد نمونه ها در هرکلاس است که در اینجا از روش smote استفاده شده است. این مدل علیرغم تراز کردن کلاسها ، دیتای تکراری نیز به وجود نخواهد آورد.

Feature Scaling

با توجه به استفاده از الگوریتمهای فاصله محوری همچون knn  در این پروژه، نیاز به هم مقیاس کردن داده های آموزش دیده، تست و و اعتباری است که در اینجا از روش minmaxscalor استفاده شده است.

**الگوریتم های یادگیری ماشین مورد استفاده در این پروژه:**

K-Nearest Neighbor (KNN)

این الگوریتم بر اساس نزدیکترین همسایه هر نمونه، کلاس مورد نظر را مشخص میکند، به دلیل حجم پایین نمونه ها و پبچیدگی پایین مدل از معیار minkowski درجه دو و 5 همسایگی استفاده شده است.

C-Support Vector Classification (SVC)

مبنای کار این الگوریتم [دسته‌بندی خطی داده‌ها](https://fa.wikipedia.org/wiki/%D8%B7%D8%A8%D9%82%D9%87%E2%80%8C%D8%A8%D9%86%D8%AF%DB%8C_%D8%AE%D8%B7%DB%8C) است و در تقسیم خطی داده‌ها سعی بر انتخاب [ابرصفحه ای](https://fa.wikipedia.org/wiki/%D8%A7%D8%A8%D8%B1%D8%B5%D9%81%D8%AD%D9%87) شده است که حاشیه اطمینان بیشتری داشته باشد.

Decision Tree

معیار سنجش در مدل درخت تصمیم gini و عمق درخت 23 انتخاب شده است.

Random Forest

معیار سنجش در مدل جنگل تصادفی gini و عمق درخت 16 انتخاب شده است. در این مدل از روش grid search , برای تنظیم کردن پارامترها استفاده شده است

**نتایج ارزیابی مدل ها:**

None)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUPPORT | | f1-score | | Recall | | precision | | valid -accuracy | test\_accuracy | EVALUTION  SCORE  ML MODEL |
| M | B | M | B | M | B | M | B |
| **75** | **79** | **0.88** | **0.90** | **0.77** | **1** | **1** | **0.82** | **0.81** | **0.89** | KNN |
| **90** | **64** | **0,78** | **0,80** | **0.64** | **1** | **1** | **0,67** | **0.66** | **0.79** | SVC |
| **66** | **88** | **0.89** | **0,92** | **0.83** | **0,97** | **0.95** | **0,89** | **0.91** | **0.91** | DTREE |
| **59** | **95** | **0,96** | **0.97** | **0.95** | **0.98** | **0.97** | **0.97** | **0,95** | **0,97** | RForest |

همانطور که از نتایج برآمده است جنگل تصادفی با بالاترین accuracy test , precision, recall, f1-score نتایج قابل قبولی برای پیش بینی تشخیص سرطان ارائه داده است.