تشخیص سرطان سینه با علم داده

چرا علم داده؟

کاهش ریسک خطا درتشخیص

کمک به پزشکان در
روند تشخیص بیماری

کاهش زمان انتظار بیماران برای
شروع روند درمانی

دیتاست جمع آوری شده

ویژگی ها			
id	concavity_mean	smoothness_se	perimeter_worst
diagnosis	concave points_mean	compactness_se	area_worst
radius_mean	symmetry_mean	concavity_se	smoothness_worst
texture_mean	fractal_dimension_mean	concave points_se	compactness_worst
perimeter_mean	radius_se	symmetry_se	concavity_worst
area_mean	texture_se	fractal_dimension_se	concave points_worst
smoothness_mean	perimeter_se	radius_worst	symmetry_worst
compactness_mean	area_se	texture_worst	fractal_dimension_worst

مراحل پردازش پروژه

مرحله اول

- Table Analysis

مرحله دوم

- EDA

مرحله سوم

- Data prepration

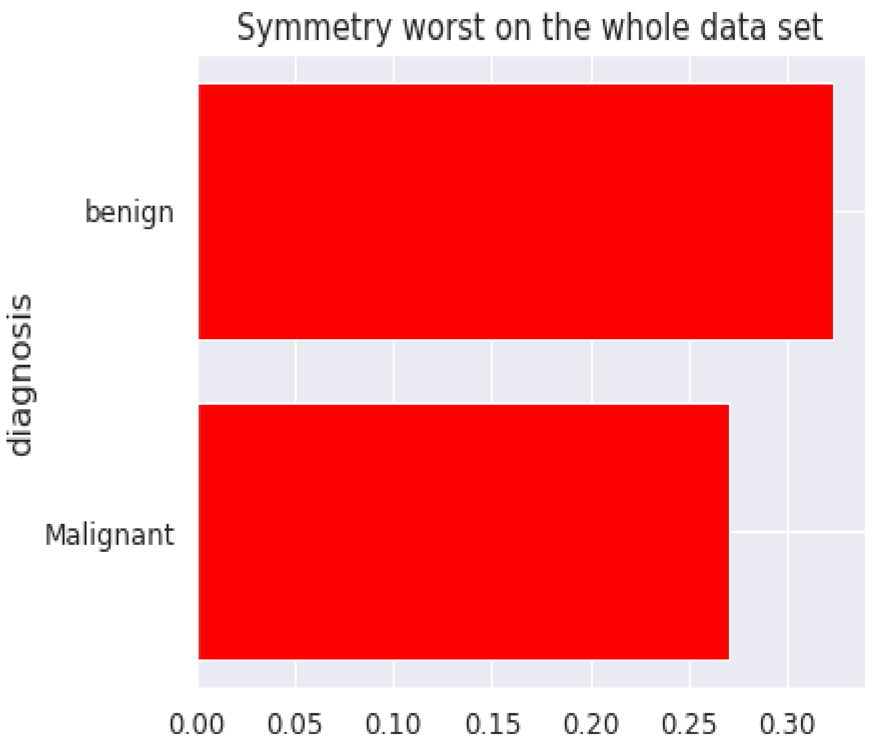
مرحله چهارم

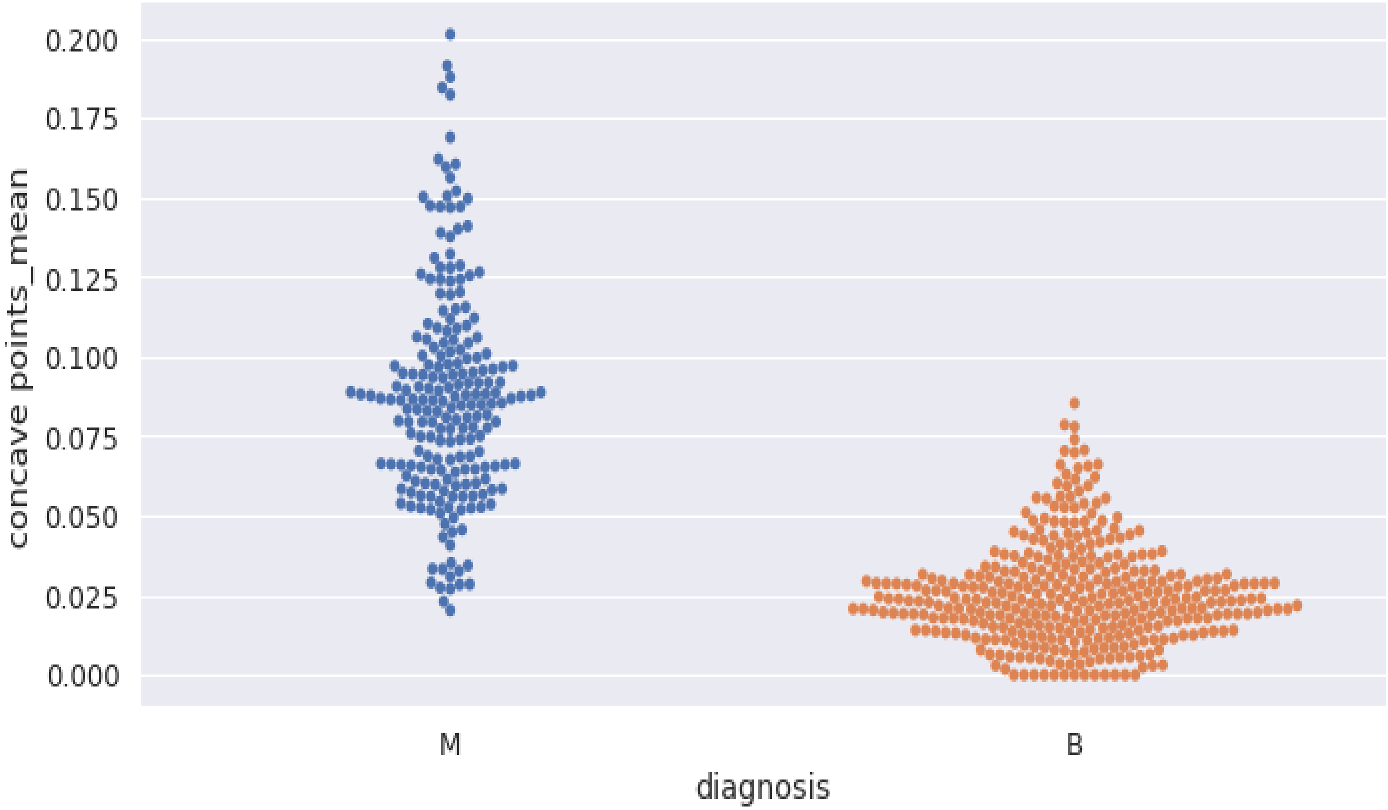
- Machine learning

Table Analysis

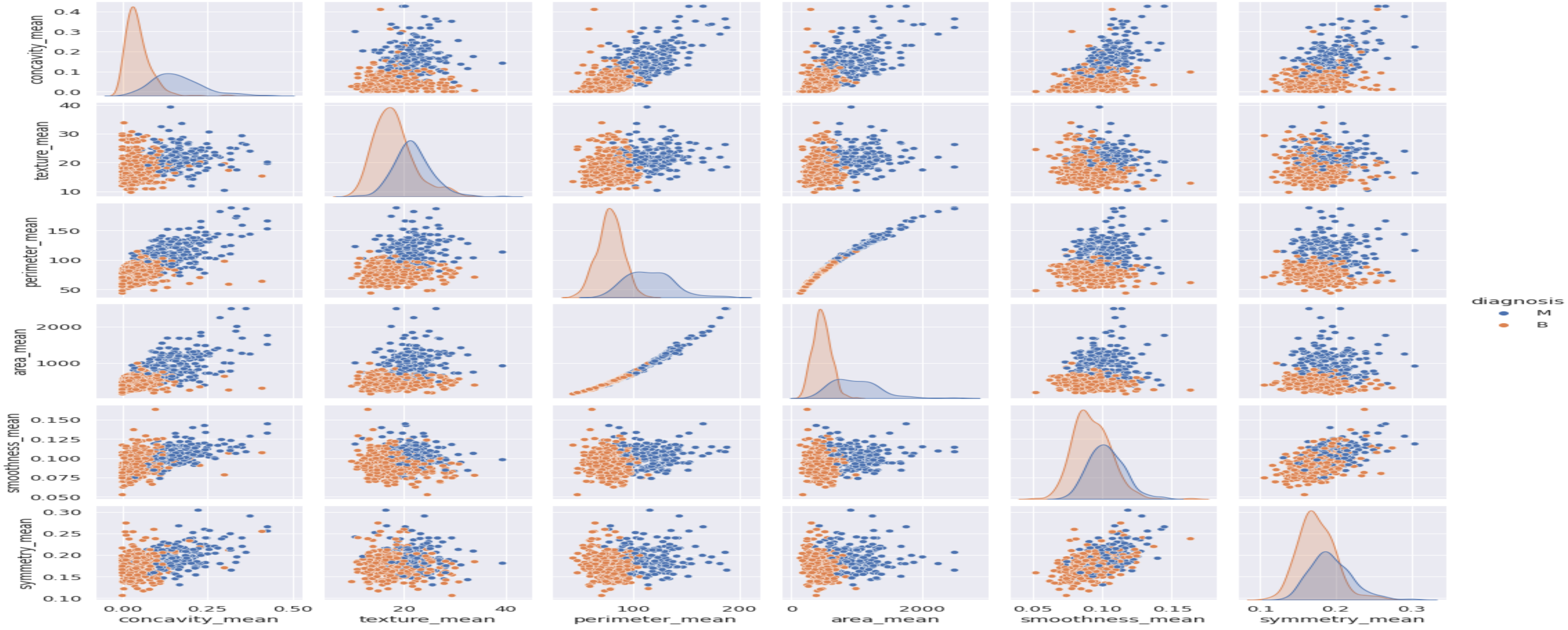
	radius_mean
count	357.000000
mean	12.146524
std	1.780512
Min	6.981000
25%	11.080000
50%	12.200000
75%	13.370000
max	17.850000

EDA with python



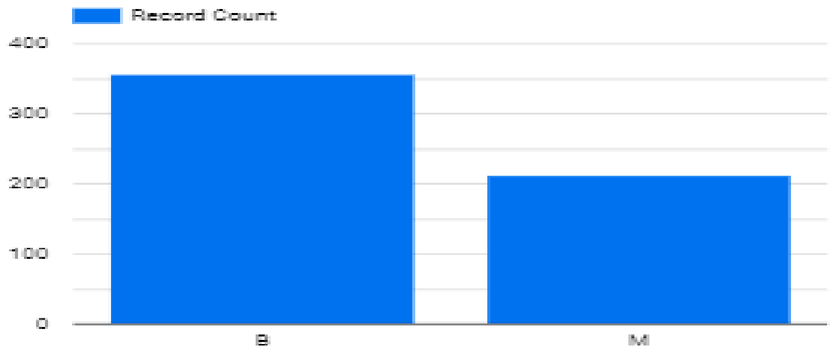


EDA with python



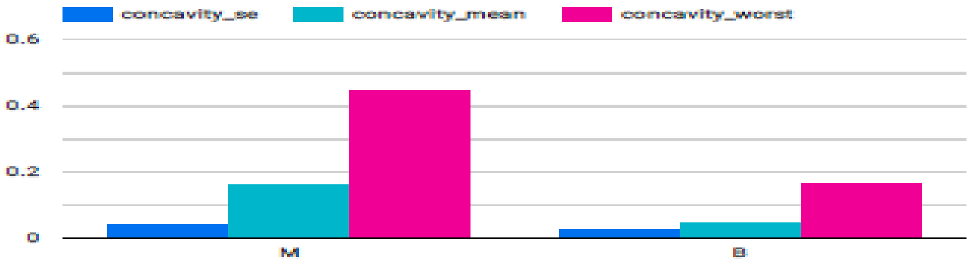
EDA with google data studio

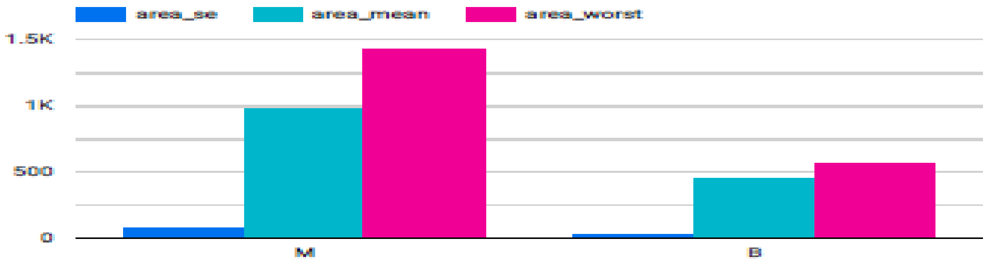
The ratio of the number of benign and Malignant tumors

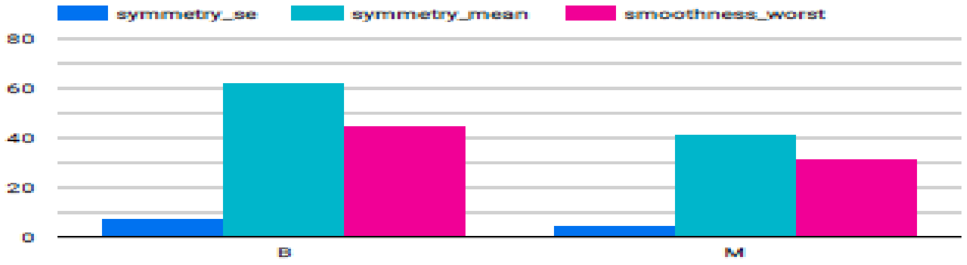


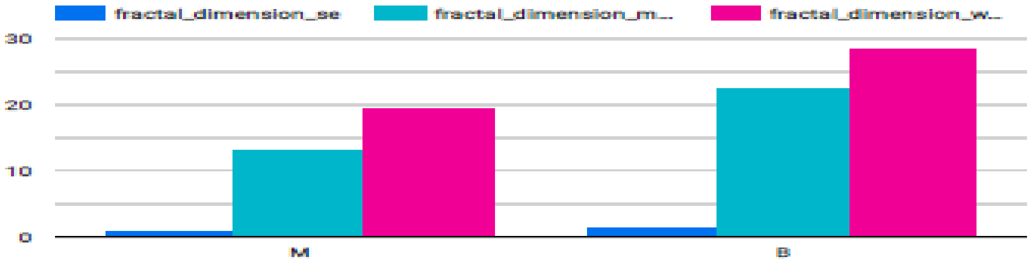
EDA with google data studio

comparison between the estimated standard error and mean and worst values of features



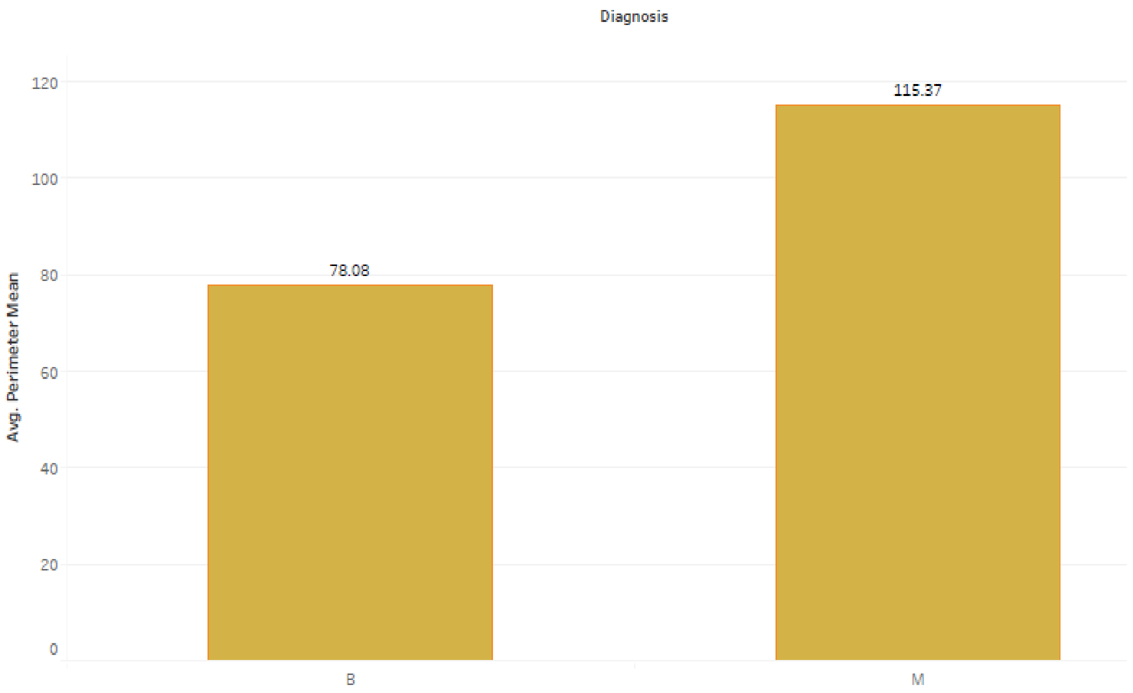






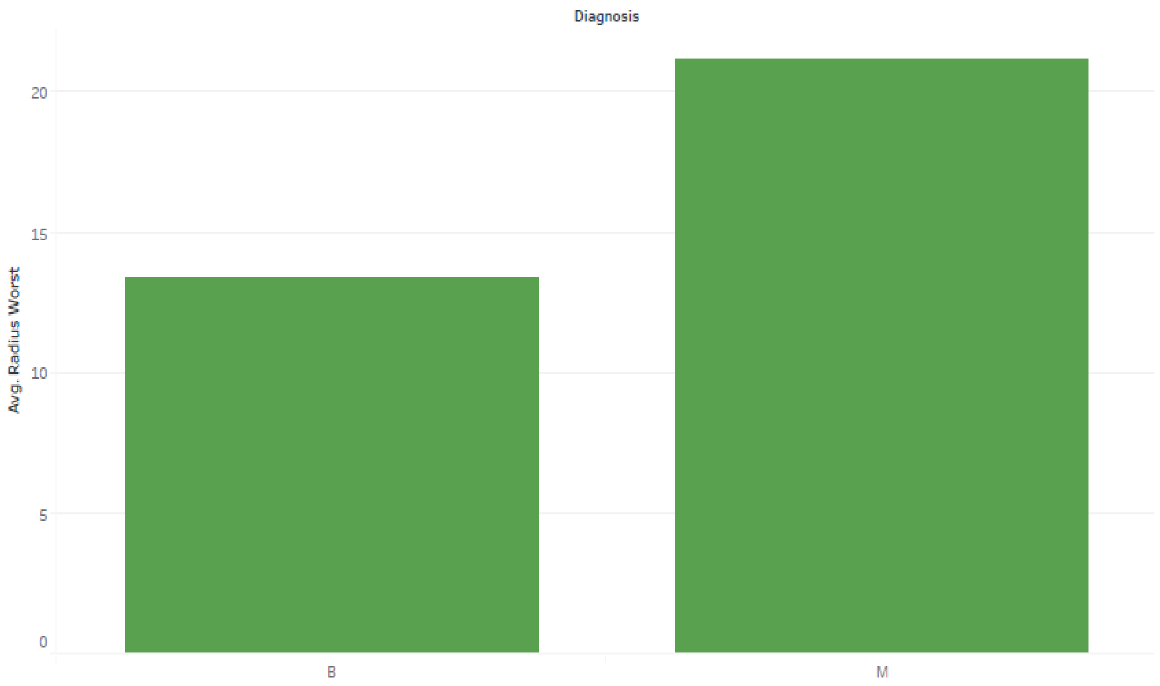
EDA with Tableau

perimeter mean Comparison



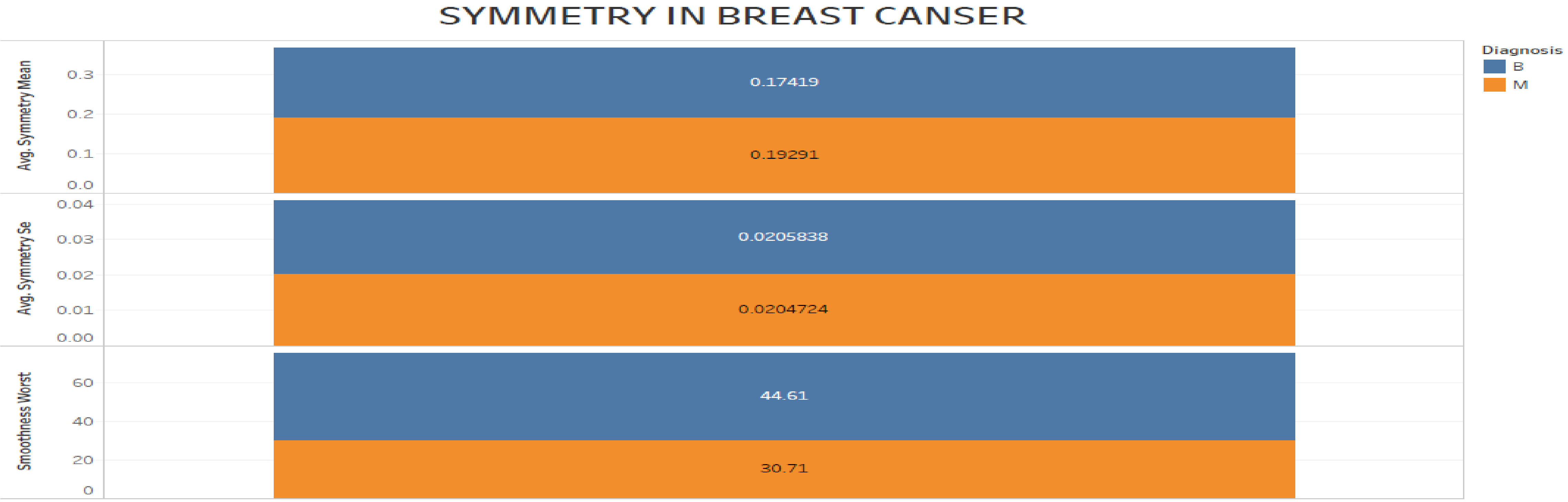
Average of Perimeter Mean for each Diagnosis.

Radius Worst Comparison



Average of Radius Worst for each Diagnosis.

EDA with Tableau



Average of Symmetry Mean, average of Symmetry Se and sum of Smoothness Worst. Color shows details about Diagnosis.

Data prepration

❑ **Data proccecing**

- Handle categorical
- Handle missing value
- Handle outlier
- Handle duplicate

❑ **Feture scaling**

- MinMax scaler

❑ **Feature selection**

- Correlation
- PCA

Machin learning

□ **spliting data into train test**

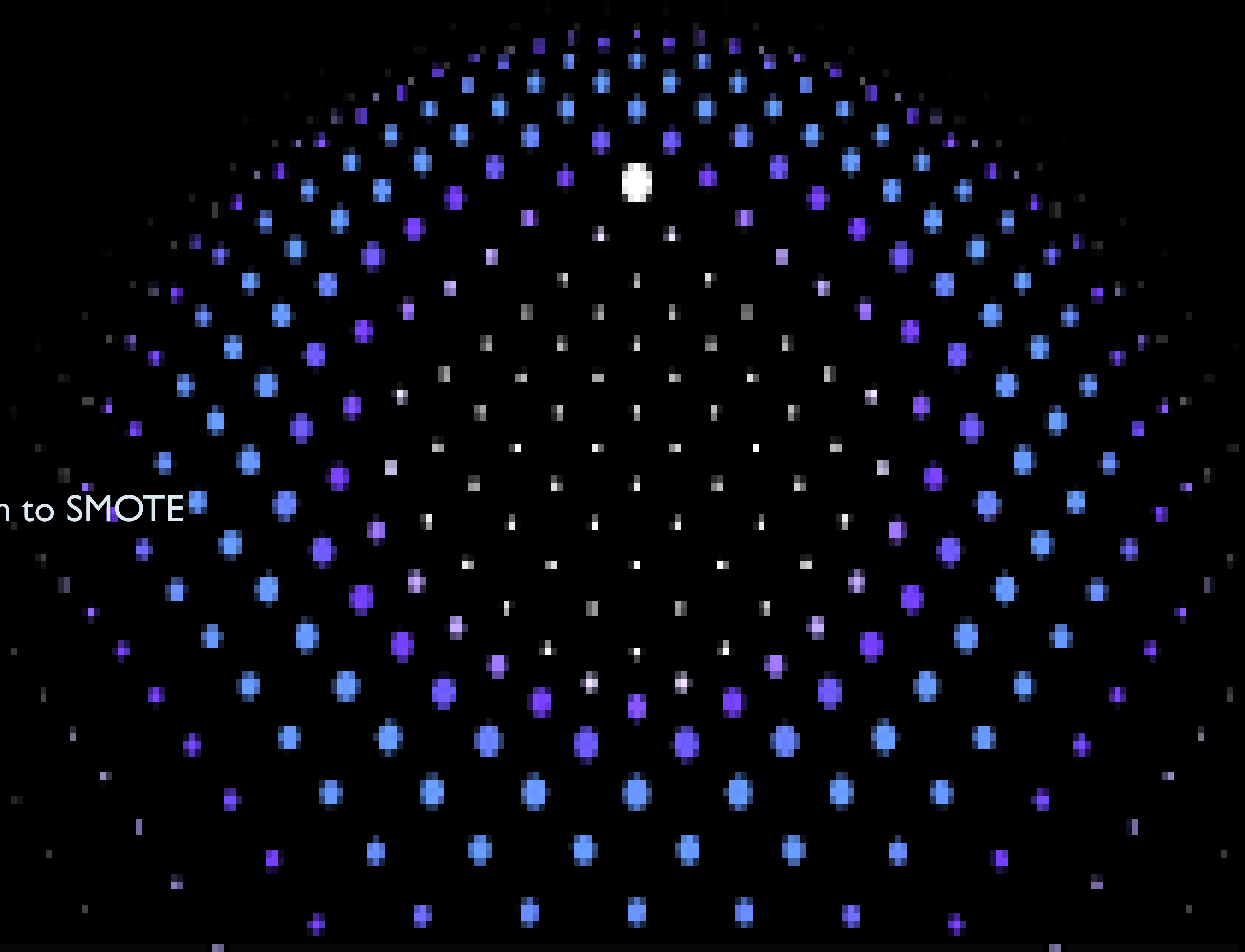
□ **Training the model**

□ **Handle Imbalance data**

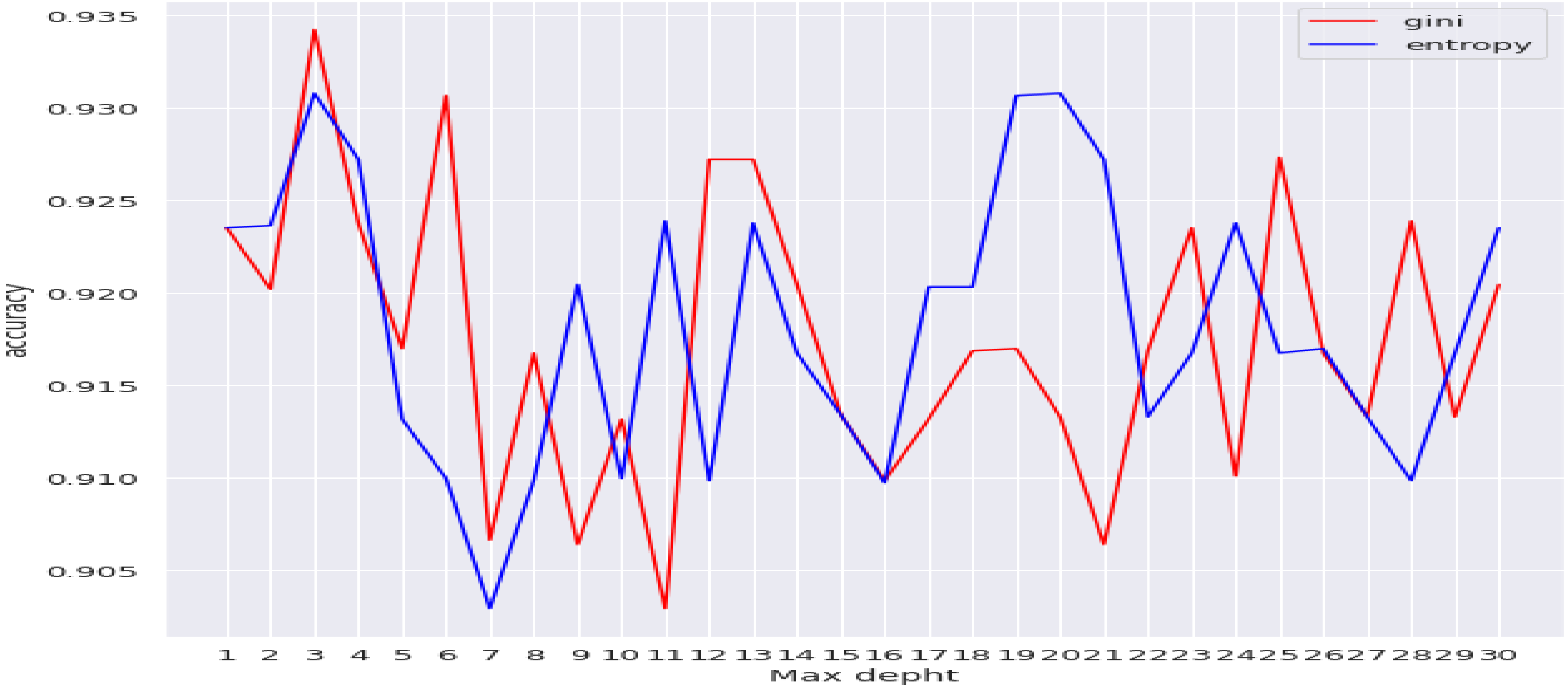
we used over sampling method but
it wasn't good enough in comparison to SMOTE

□ **Training the model**

- gini
- entropy



نمودار دقت الگوریتم نسبت به عمق



الگوریتم های استفاده شده

- ✓ **DecisionTree**
- ✓ **RandomForest**

اعتبار سنجی الگوریتم

- **Confusion matrix**

decision tree:

$$\begin{bmatrix} 101 & 2 \\ 5 & 30 \end{bmatrix}$$

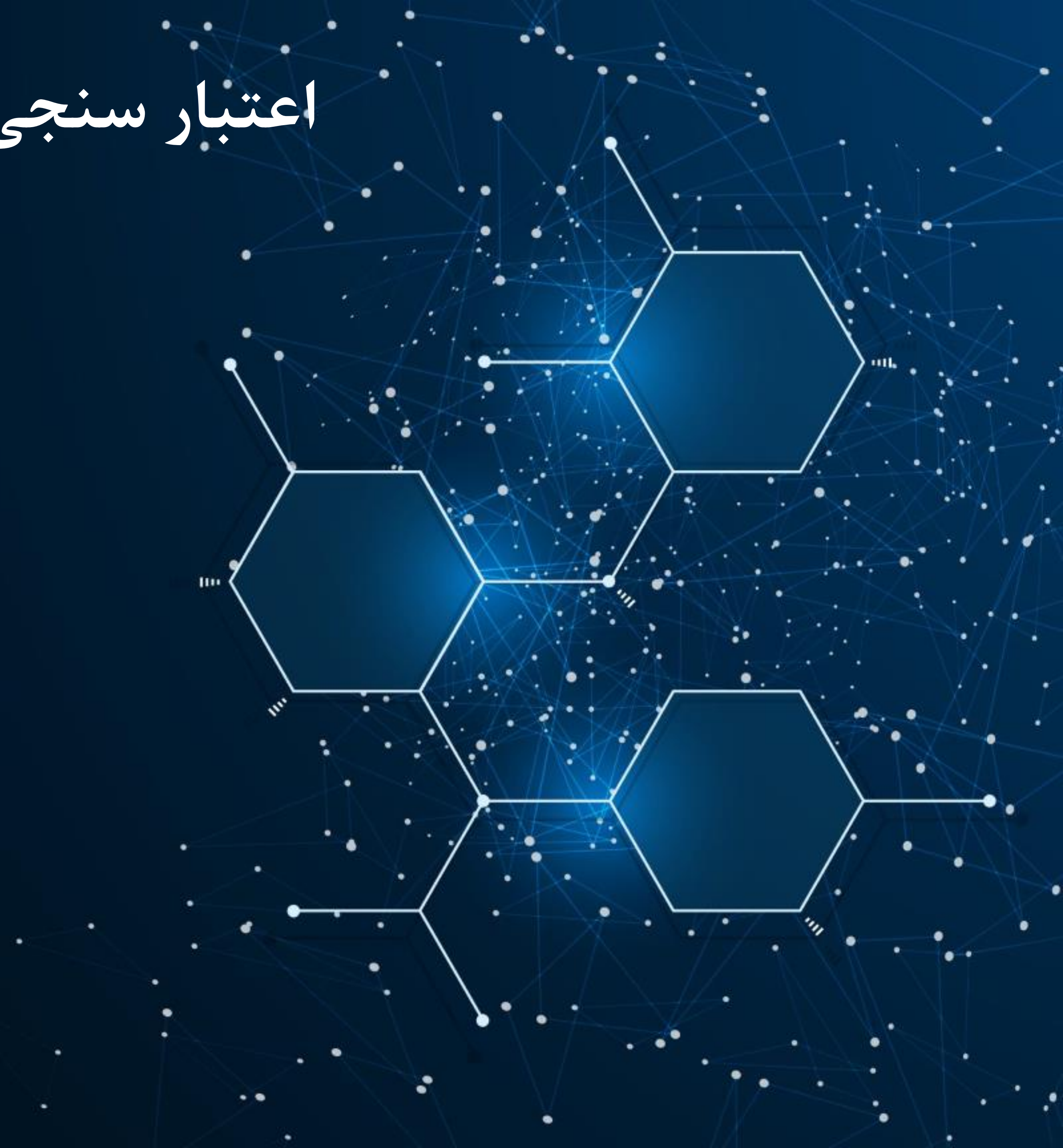
random forest :

$$\begin{bmatrix} 104 & 2 \\ 2 & 30 \end{bmatrix}$$

- **Accuracy**

decision tree :0.94

random forest : 0.97



اعتبارسنجی الگوریتم

- **Precision**

  decision tree : 0.9375

  random forest : 0.9375

- **Recall**

- **F1 score**

  decision tree : 0.89

  random forest : 0.90

check overfitting

accuraccy on train set for decision tree : 0.97

accuraccy on train set for random forest : 0.98

accuraccy on validation set for decision tree : 0.81

accuraccy on validation set for random forest : 0.93

THANK YOU

Authors:

Omid
Nazanin
Sahar eskandari