

# بنام خدا

## Dependent Variable

diagnosis	تشخیص تومور
-----------	-------------

خوشخیم = 0

بدخیم = 1

## Independent Variables

radius_worst	شعاع بزرگترین نمونه تومور
concave_points_worst	نقاط فرو رفته شدید در لبه تومور (بدترین حالت)
area_worst	مساحت بزرگترین نمونه تومور
texture_worst	بافت بدترین نمونه
perimeter_worst	محیط بزرگترین نمونه تومور

## Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-12.18105	8.82679	-1.380	0.1676
radius_worst	-0.89496	1.16552	-0.768	0.4426
concave_points_worst	63.10974	11.40665	5.533	3.15e-08
area_worst	0.02769	0.01161	2.384	0.0171
texture_worst	0.27597	0.05510	5.009	5.48e-07
perimeter_worst	-0.10730	0.08777	-1.223	0.2215

براساس پی مقدار سه متغیر نقاط فرورفته شدید در لبه تومور و مساحت بزرگترین نمونه تومور و بافت بدترین نمونه که مقدار کمتر از 0.05 را دارند معنادار هستند.

نقاط فرورفته شدید در لبه تومور : افزایش این متغیر احتمال بدخیمی تومور را به شدت افزایش می‌دهد

بافت بدترین نمونه : بافت نامنظمتر (مقادیر بالاتر) با احتمال بیشتر بدخیمی همراه است

مساحت بزرگترین نمونه تومور : تومورهای با مساحت بزرگتر احتمال بدخیمی بیشتری دارند

### Dependent Variable

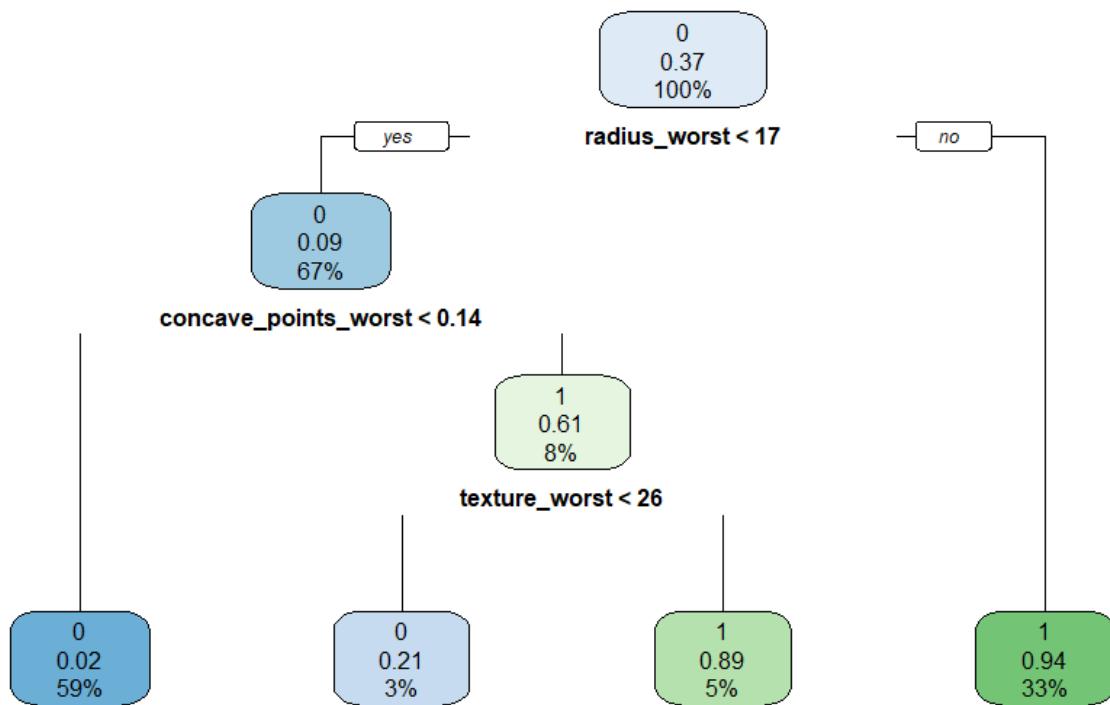
diagnosis	تشخیص نومور
-----------	-------------

خوشخیم = 0

بدخیم = 1

### Independent Variables

radius_worst	شعاع بزرگترین نمونه تومور
concave_points_worst	نقاط فرو رفته شدید در لبه تومور (بدترین حالت)
texture_worst	بافت بدترین نمونه



شرط : 17 < شعاع بزرگترین نمونه تومور

چپ → (yes) اگر بله

راست → (no) اگر خیر

شرط : 14 < نقاط فرو رفته شدید در لبه تومور

اگر بله : تقریباً همه نمونه‌ها خوش‌خیم هستند (تشخیص = 0)

اگر خیر : شرط بعدی بررسی می‌شود

> 26 بافت بدترین نمونه

اگر بله: تشخیص = 0 (خوش‌خیم)

اگر خیر: تشخیص = 1 (بدخیم)

=> شعاع بزرگترین نمونه تومور 17

مستقیماً به تشخیص = 1 می‌رسد، یعنی بدخیم

### Dependent Variable

diagnosis	تشخیص تومور
-----------	-------------

خوشخیم = 0

بدخیم = 1

### Independent Variables

radius_worst	شعاع بزرگترین نمونه تومور
concave_points_worst	نقاط فرو رفته شدید در لبه تومور (بترین حالت)
area_worst	مساحت بزرگترین نمونه تومور
texture_worst	بافت بترین نمونه
perimeter_worst	محیط بزرگترین نمونه تومور

Confusion matrix:

		0	1	class.error
0		347	10	0.02801120 (خطا = 2.8)
1		13	199	0.06132075 (خطا = 6.1)

خوشخیم‌ها با دقت بالا (97.2%) درست تشخیص داده شدند

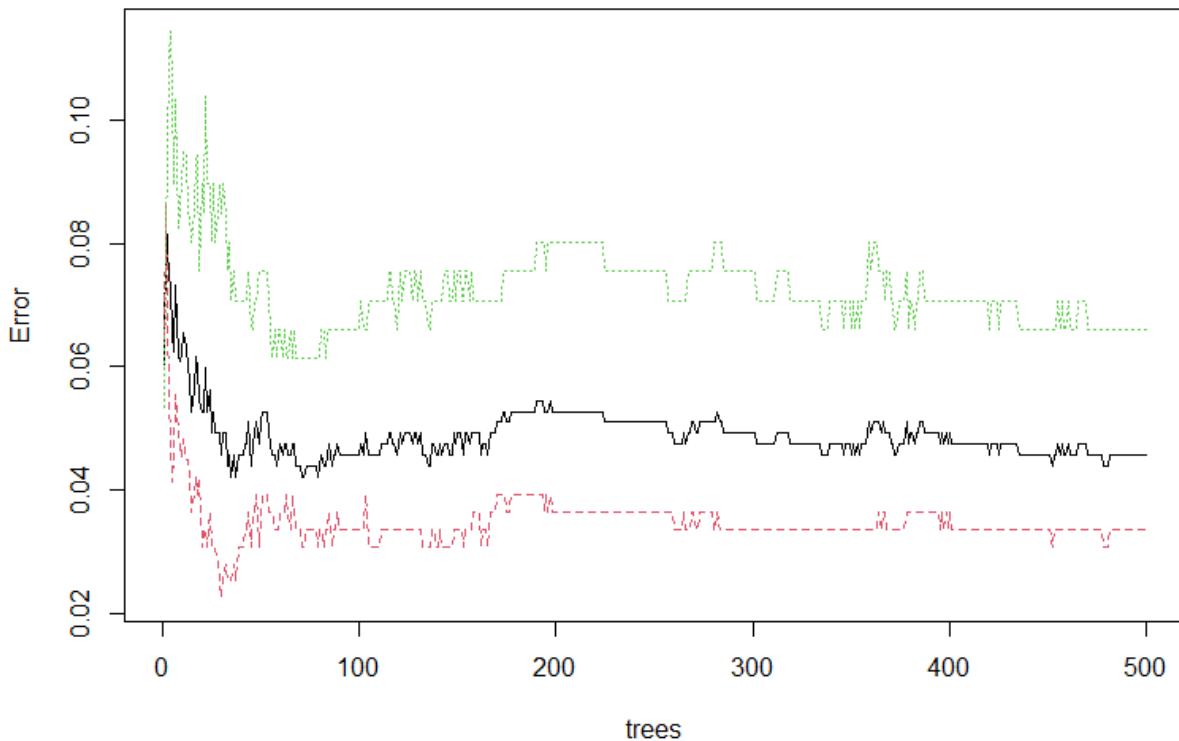
بدخیم‌ها نیز با دقت نسبتاً بالا (93.9%) درست تشخیص داده شدند

			MeanDecreaseAccuracy	MeanDecreaseGini
	0	1		
radius_worst	19.32329	11.76038	21.62009	55.80773
concave_points_worst	40.88478	23.47408	46.50367	63.82972
area_worst	23.39122	14.98019	27.13969	63.73982
texture_worst	21.16218	24.01854	31.71456	15.44424
perimeter_worst	16.92538	13.64748	22.14493	67.01804

متغیر	تأثیر در دقت	Gini تأثیر در
نقاط فرو رفته شدید در لبه تومور	46.50	63.83
مساحت بزرگترین نمونه تومور	27.14	63.74
بافت بدترین نمونه	31.71	15.44
شعاع بزرگترین نمونه تومور	21.62	55.81
محیط بزرگترین نمونه تومور	22.14	67.02

مهمترین متغیرها در پیش‌بینی :  
 محیط بزرگترین نمونه تومور , مساحت بزرگترین نمونه تومور , نقاط فرو رفته شدید در لبه تومور  
 نقاط فرو رفته شدید در لبه تومور بیشترین تأثیر را در دقت و خلوص تصمیم‌گیری دارد.

### نمودار خطای مدل جنگل تصادفی



رنگ خط	معنی	تفسیر
سیاه	خطای کلی مدل	(OOB) خطای کلی بر اساس پیش‌بینی‌های خارج از نمونه
قرمز	خطای کلاس خوش‌خیم	چقدر مدل در تشخیص تومور خوش‌خیم اشتباه کرده
سبز	خطای کلاس بدخیم	چقدر مدل در تشخیص تومور بدخیم اشتباه کرده

x: محور افقی) تعداد درختها

افزایش در تعداد درخت‌ها، یعنی مدل پیچیده‌تر و دقیق‌تر می‌شود تا جایی که خط تثبیت شود.

خطای کلی (خط مشکی)

- تا ۰.۰۵، قرار دارد درخت به بعد، خطابه صورت پایدار در حدود ۰.۰۴ از حدود ۰.۰۵ است یعنی دقت مدل حدود ۹۶ درصد.

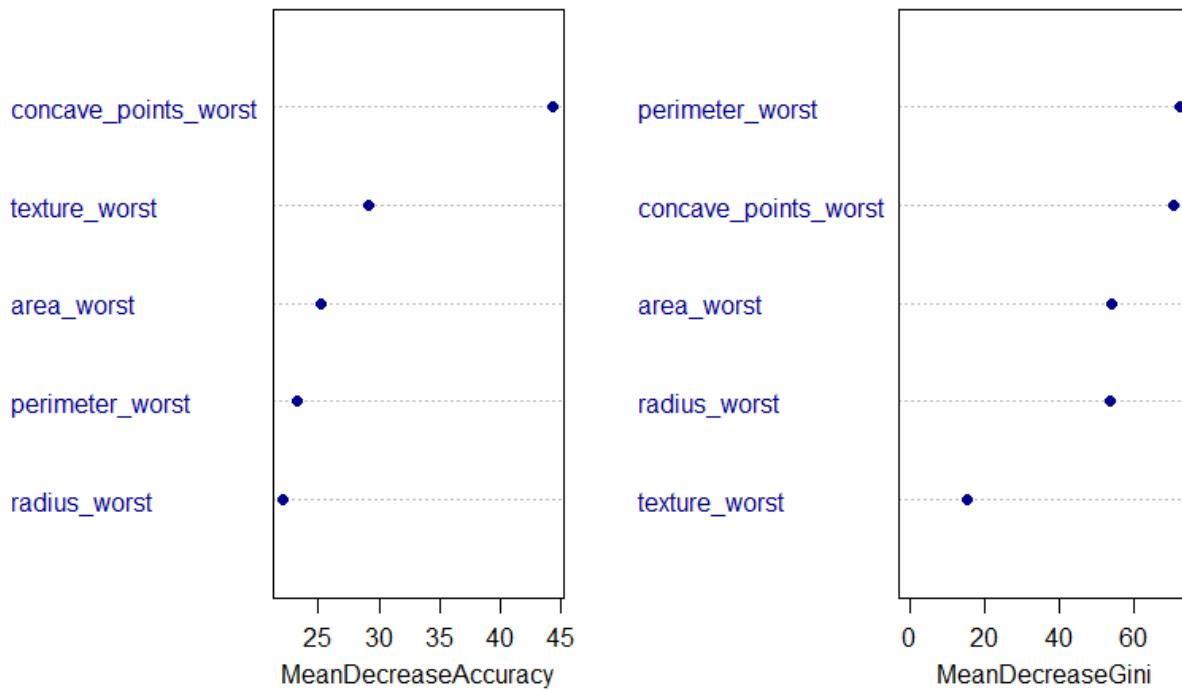
## خط قرمز (کلاس خوشخیم - ۰)

- از همان ابتدا خیلی پایین است (زیر ۳%)
- مدل در تشخیص تومورهای خوشخیم بسیار موفق بوده

## خط سبز (کلاس بدخیم - ۱)

- در ابتدای ساخت مدل خطای بالایی دارد (تا ۱۰%)
- اما پس از حدود ۱۰۰ درخت، کاهش می‌یابد و به حدود ۶% یا کمتر می‌رسد
- این یعنی تشخیص تومورهای بدخیم سخت‌تر از خوشخیم بوده، اما عملکرد قابل قبولی دارد

### اهمیت ویژگی‌ها در مدل جنگل تصادفی



### نمودار سمت چپ

یعنی تأثیر متغیر بر روی این معیار نشان می‌دهد که اگر یک متغیر را از مدل حذف کنیم، چقدر دقت مدل کاهش می‌یابد  
دقت کلی مدل

## نمودار سمت راست

در گره‌های درخت است. هر چه این عدد بیشتر باشد، متغیر (Gini Impurity) این معیار بر اساس کاهش معیار جینی در تفکیک کلاس‌ها نقش بیشتری داشته.

SVM

		Actual	
Predicted		0	1
0	0	58	4
	1	3	49

**True Negatives (TN)** — 58 : تشخیص درست کلاس 0

**True Positives (TP)** — 49 : تشخیص درست کلاس 1

**False Negatives (FN)** — 4 : پیش‌بینی 0 ولی واقعاً 1 بوده

**False Positives (FP)** — 3 : پیش‌بینی 1 ولی واقعاً 0 بوده

Accuracy: 0.9385965

یعنی مدل در حدود 94 درصد موقع درست عمل کرده.

	Mean	SD
svm.sens.train	0.9165194	0.014026787
svm.spec.train	0.9701891	0.011275455
svm.ppv.train	0.9486581	0.017609458
svm.npv.train	0.9517619	0.006836554
svm.acc.train	0.9503769	0.006331962
svm.sens.test	0.9146802	0.033001194
svm.spec.test	0.9662585	0.017973483
svm.ppv.test	0.9417632	0.031176460
svm.npv.test	0.9497712	0.020115013
svm.acc.test	0.9467251	0.014225164

داده‌های آموزش (Train):

حساسیت (Sensitivity/Recall):  $0.9165 \pm 0.0140$

ویژگی (Specificity):  $0.9701 \pm 0.0112$

PPV/Precision: مقدار پیش‌بینی مثبت  $0.9486 \pm 0.0176$

NPV: مقدار پیش‌بینی منفی  $0.9517 \pm 0.0068$

دقت کلی (Accuracy):  $0.9503 \pm 0.0063$

داده‌های آزمون (Test):

حساسیت:  $0.9146 \pm 0.0330$

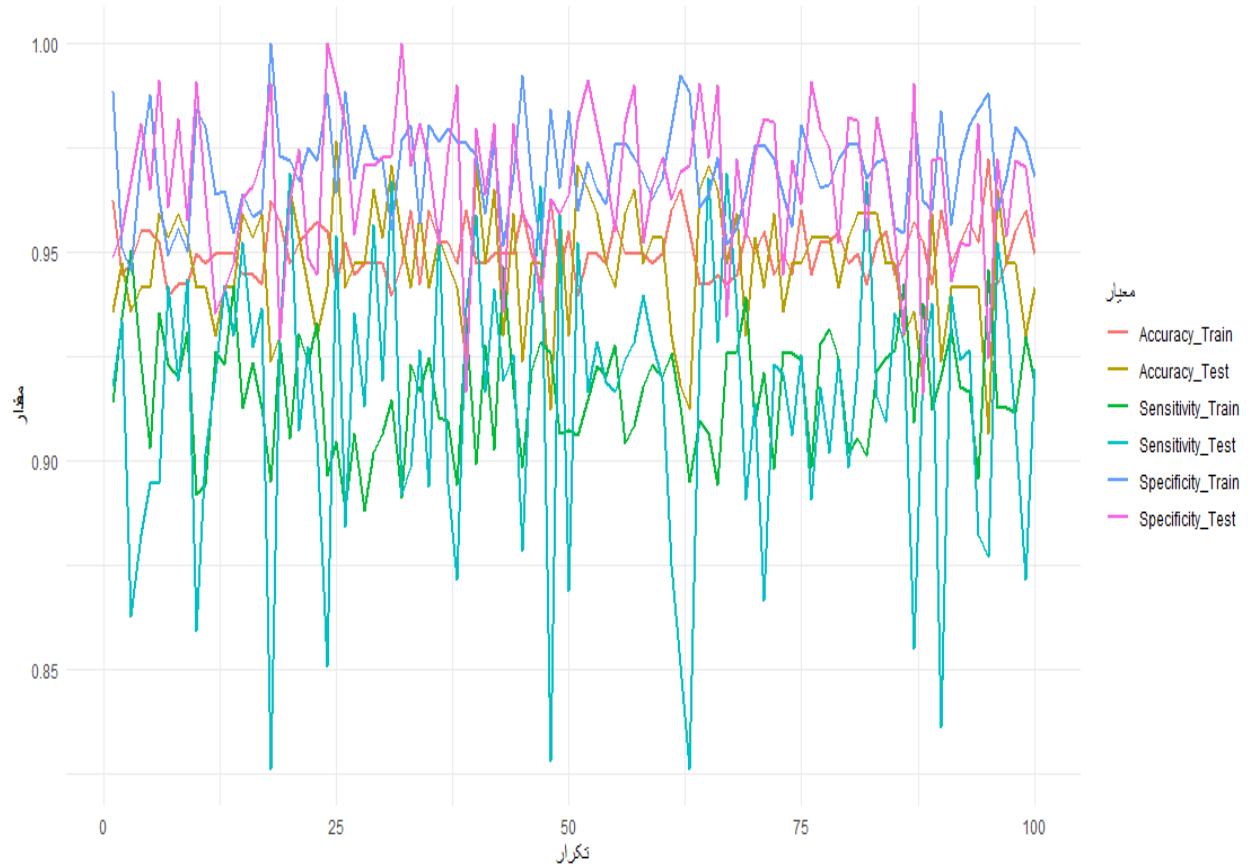
ویژگی:  $0.9662 \pm 0.0179$

PPV:  $0.9417 \pm 0.0311$

NPV:  $0.9497 \pm 0.0201$

دقت کلی:  $0.9467 \pm 0.0142$

نمودار عملکرد مدل در تکرارهای مختلف اس وی ام



Svm مدل پس از حدود 30 تکرار به همگرایی رسیده

عملکرد مدل در داده های آموزش و آزمون مشابه هست

مدل در شناسایی کلاس منفی (Specificity) کمی بهتر از شناسایی کلاس مثبت (Sensitivity) عمل میکند

این نتایج تأیید میکنند که پیاده سازی شده برای داده ها انتخاب مناسبی بوده و به خوبی آموزش دیده است.