به نام خدا

تمرین درس یادگیری آماری

سری پنجم

امیدرضا داودنیا

زمستان 1402

۱- فرض کنید دیتاست زیر برای طبقه بندی در اختیار داریم و می خواهیم از درخت تصمیم گیری برای طبقه بندی استفاده کنیم:

Weight	Eye Color	Num. Eyes	Output
N	A	2	L
N	V	2	L
N	V	2	L
U	V	3	L
U	V	3	L
U	A	4	D
N	A	4	D
N	V	4	D
U	A	3	D
U	A	3	D

درخت تصمیم گیری کامل را برای طبقه بندی این دیتاست رسم کنید.

در مرحله اول باید یکی از متغییر های توصیفی مثل وزن و یا رنگ چشم و یا اینکه تعداد چشم کمتر یا بیشتر از ۲,۵ باشد و یا اینکه تعداد چشم کمتر یا بیشتر از ۳,۵ باشد، باید انتخاب گردد، ازین رو ما برای اینکه بتوانیم پاسخ نهایی خود را با درخت شبه بهینه انتخاب شده برای افراز فضای ما توسط کد و به صورت دستی قابل مقایسه باشد، از معیار ***Gini*** در طول محاسبات استفاده خواهیم کرد.

متغییر های باقیمانده	وزن	رنگ چشم	تعداد چشم کمتر از ۲,۵	تعداد چشم کمتر از ۳,۵

مرحله اول :

$$\text{Gini}_N = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{12}{25} = 0.48$$

$$\text{Gini}_L = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{12}{25} = 0.48$$

$$\text{Gini}_{Weight} = \text{Gini}_N \times \left(\frac{5}{10}\right) + \text{Gini}_L \times \left(\frac{5}{10}\right) = 0.48$$

$$\text{Gini}_A = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = \frac{8}{25} = 0.32$$

$$\text{Gini}_V = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = \frac{8}{25} = 0.32$$

$$\text{Gini}_{Weight} = \text{Gini}_A \times \left(\frac{5}{10}\right) + \text{Gini}_V \times \left(\frac{5}{10}\right) = 0.32$$

همین جا می توان گفت متغییر رنگ چشم از متغییر وزن مناسب تر است.

$$\text{Gini}_{<2.5} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$\text{Gini}_{>2.5} = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{5}{7}\right)^2 = \frac{20}{49}$$

$$\text{Gini}_{2.5} = \text{Gini}_{<2.5} \times \left(\frac{3}{10}\right) + \text{Gini}_{V>2.5} \times \left(\frac{7}{10}\right) = 0.285$$

همین جا می توان گفت بهترین ترمینال اول تا به اینجا تعداد چشم کمتر یا بیشتر از ۲,۵ است.

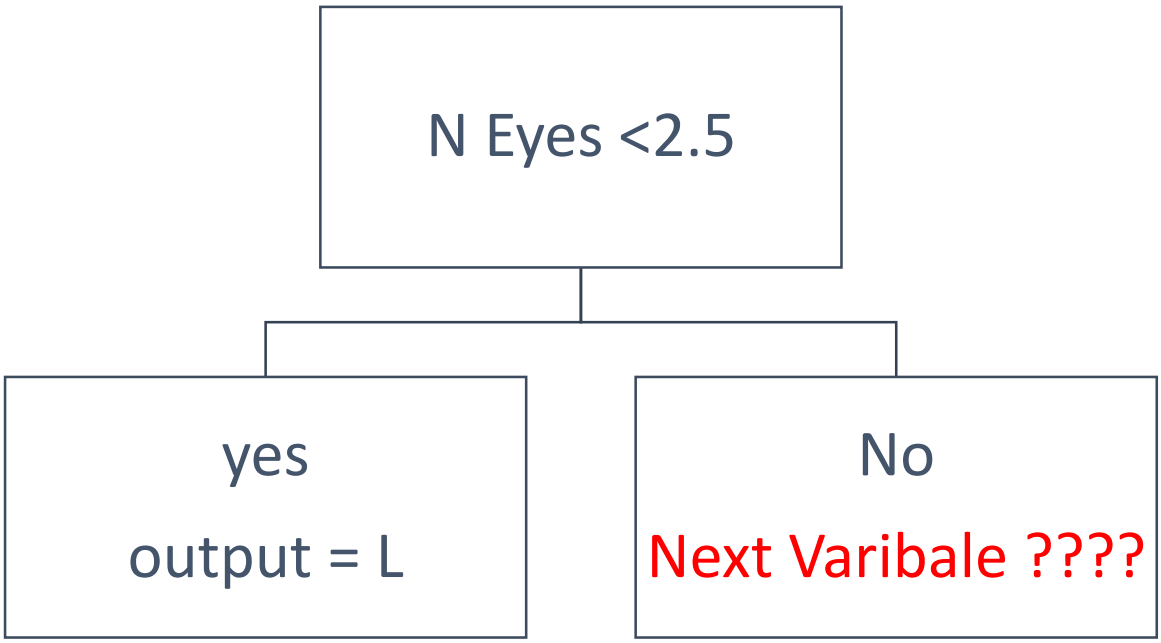
$$\text{Gini}_{<2.5} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$\text{Gini}_{>2.5} = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{5}{7}\right)^2 = \frac{20}{49}$$

$$\text{Gini}_{2.5} = \text{Gini}_{<2.5} \times \left(\frac{3}{10}\right) + \text{Gini}_{>2.5} \times \left(\frac{7}{10}\right) = 0.285$$

همین جا می توان گفت بهترین ترمینال اول تا به اینجا تعداد چشم کمتر یا بیشتر از ۲,۵ است، البته برای داشتن گره در ۳,۵ نیز همین مقدار را داریم ، که برای متابقت با شبیه سازی ما نیز گره ۲,۵ را انتخاب خواهیم کرد.

$$\text{Gini}_{3.5} = \text{Gini}_{<3.5} \times \left(\frac{7}{10}\right) + \text{Gini}_{>2.5} \times \left(\frac{3}{10}\right) = 0.285$$



درخت تابحال یک گره دارد در سمت چپ دیگر گره ای نخواهد داشت و در سمت راست ۷ داده دیگر برای دسته بندی خواهیم داشت پس جدول بدین صورت خواهد بود.

در این مرحله جدول برای شاخه سمت راست بدین صورت در خواهد آمد.

متغییر های باقیمانده	وزن	رنگ چشم	تعداد چشم کمتر از ۳,۵

مرحله دوم :

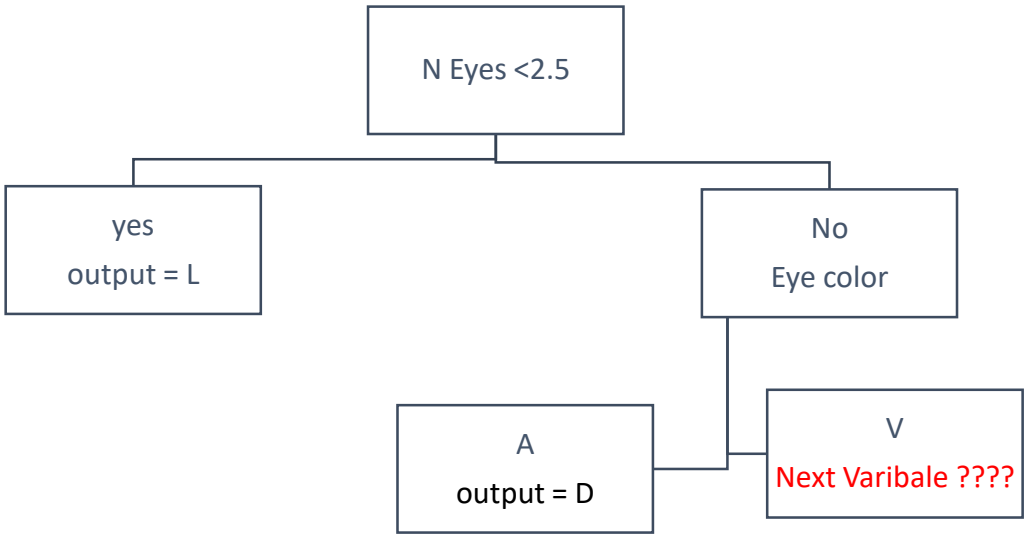
Weight	Eye Color	Number of Eyes	Output
U	V	۳	L
U	V	۳	L
U	A	۳	D
U	A	۳	D
U	A	۴	D
N	A	۴	D
N	V	۴	D

$Gini_{3.5} = 0.285$

$Gini_{Eye\ color} = 0.19$

$Gini_{weight} = 0.285$

پس در این مرحله رنگ چشم انتخاب خواهد شد.



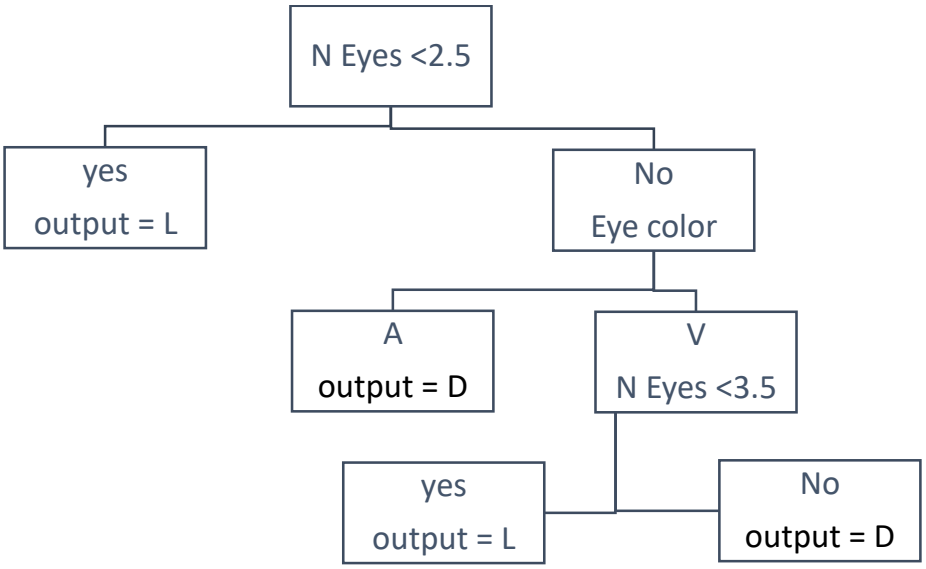
در این مرحله جدول برای شاخه سمت راست بدین صورت در خواهد آمد.

متغییر های باقیمانده	وزن	تعداد چشم کمتر از ۳,۵

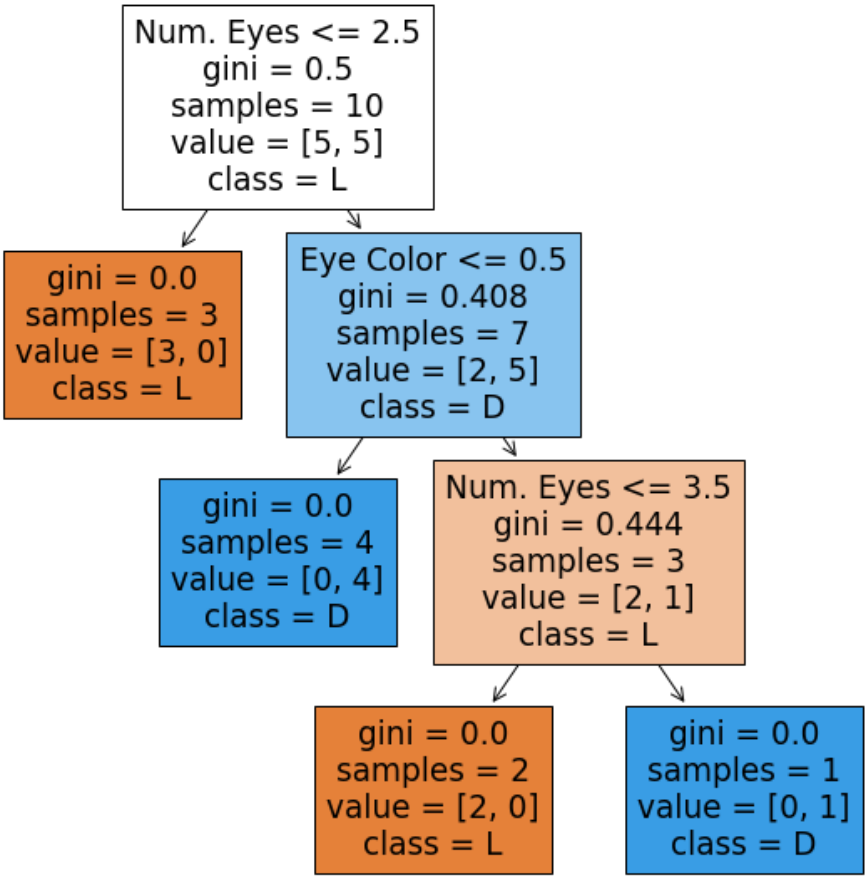
مرحله سوم :

Weight	Number of Eyes	Output
U	۳	L
U	۳	L
N	۴	D

همانطور که شکل جدول نیز واضح است دیگر نیازی به محاسبه نیست و انتخاب هر یک از دو متغییر تعداد چشم کمتر یا بیشتر از ۳,۵ و وزن می تواند دقیقا یک جواب بدهد؛ ما نیز دقیقا برای اینکه با شبیه سازی یک جواب داشته باشیم متغییر ۳,۵ را برای تعداد چشم انتخاب خواهیم کرد.



مراحل طی شده که توسط محاسبات دستی انجام شدند ، توسط کد نیز قابل پیاده سازی است که در ادامه آورده شده است.



۲- با توجه به داده‌های موجود در جدول زیر، اولین ویژگی‌ای که در درخت تصمیم‌گیری انتخاب می‌شود چه خواهد بود؟ چرا؟

x_1	x_2	x_3	x_4	class
0	0	0	0	-
0	0	1	1	+
0	1	0	1	-
0	1	1	0	+
1	0	0	0	+
1	0	1	1	+
1	1	1	0	+
1	1	1	1	+

با نگاه کردن به مقادیر Gini *مربوط به هر پارامتر خواهیم داشت.*

$$Gini_{x_1} = \frac{1}{8}$$

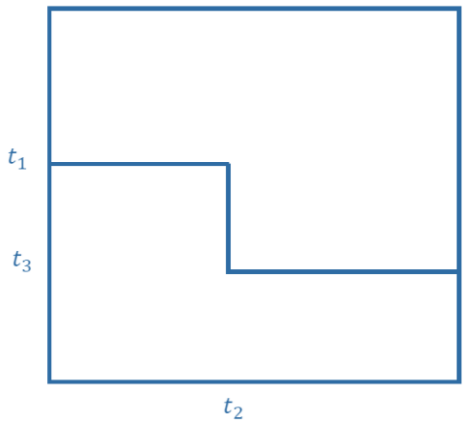
$$Gini_{x_2} = \frac{6}{16}$$

$$Gini_{x_3} = \frac{1}{6}$$

$$Gini_{x_4} = \frac{6}{16}$$

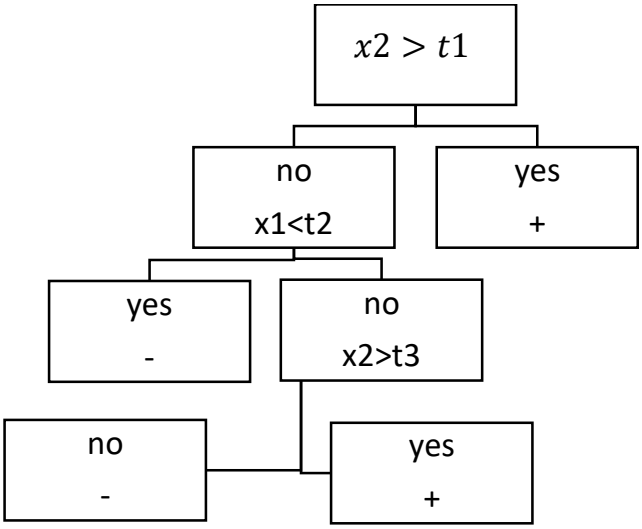
پس برای اولین گره متغییر x_3 *مناسب ترین متغییر خواهد بود.*

۳- الف: یک درخت تصمیم گیری برای مرز نشان داده شده در شکل زیر با حداقل برگ پیدا کنید.



دقیقا شکل مورد نظر با چنین مرز تصمیم گیری توسط درخت قابل ایجاد نخواهد بود ولی مرز زیر را می توان ساخت.

+	+
−	+
−	−



ب: آیا این مرز می تواند در نتیجه ترکیب چند درخت به عمق یک بدست آمده باشد؟ توضیح دهید.(توجه داشته باشید که در ترکیب درخت های تصمیم گیری از رای اکثریت درخت ها برای مشخص کردن برچسب یک ناحیه استفاده می کنیم.)

اگر شکل مورد نظر را بدین صورت در نظر بگیریم.(برای مرز $t1$)

+ و **a**
b و_

همچنین برای مرز $t2$ بدین صورت؛

+	−

و برای مرز $t3$ بدین صورت؛

+
-

اگر بین سه درخت مورد نظر رای گیری انجام گردد، با وزن ۱ برای هر درخت تک نود.

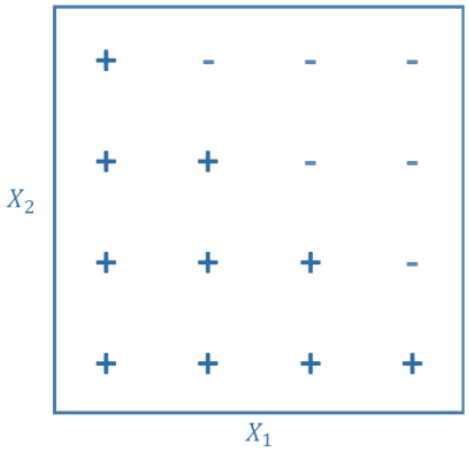
+	+
–	+
–	–

و دقیقا به شکل مورد نظر خواهیم رسید.

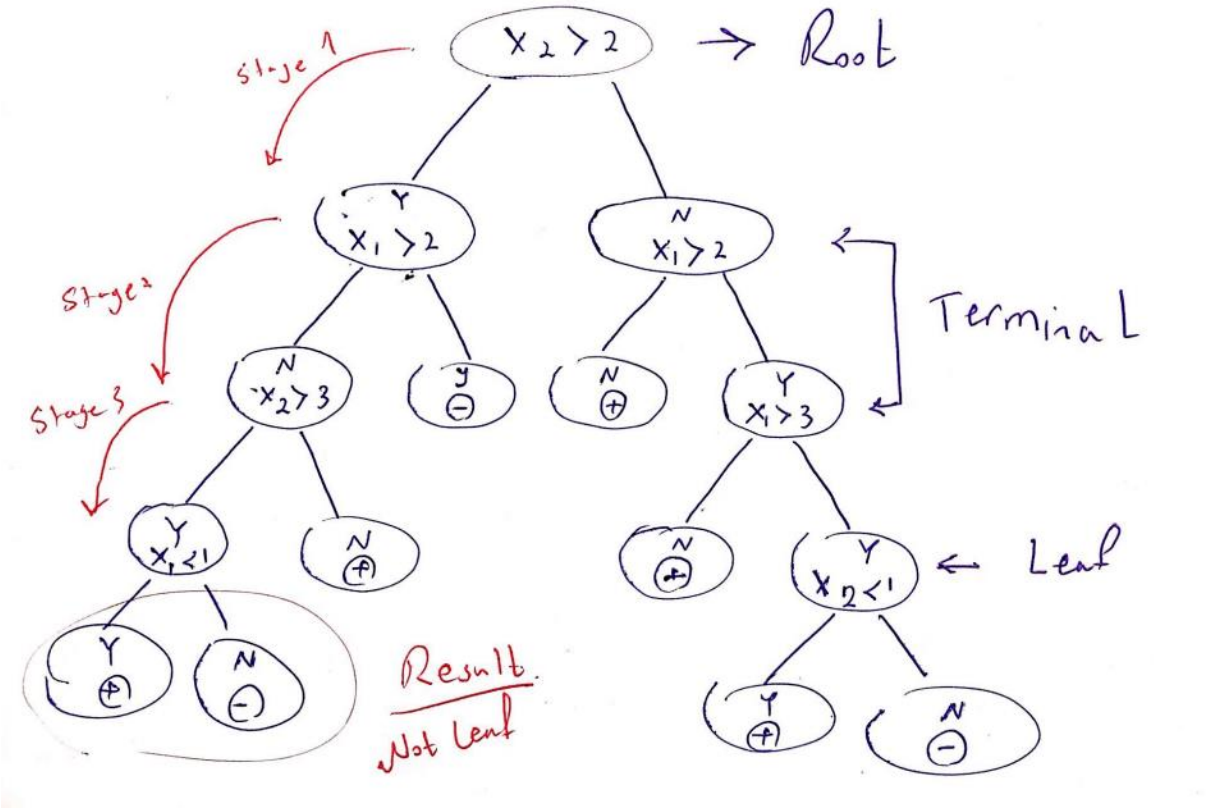
۴- در درخت تصمیم چند متغیره، در هر انشعاب، تابعی از ویژگی‌ها مورد بررسی قرار می گیرند. مثلا ممکن است در یک انشعاب، اگر $3X_2 + 6X_1 + 8 \geq 0$ به شاخه راست و اگر $3X_2 + 6X_1 + 8 < 0$ به شاخه چپ برویم.

داده های شکل زیر را در نظر بگیرید:

الف: درخت تصمیم تک متغیره با خطای آموزش صفر بسازید.



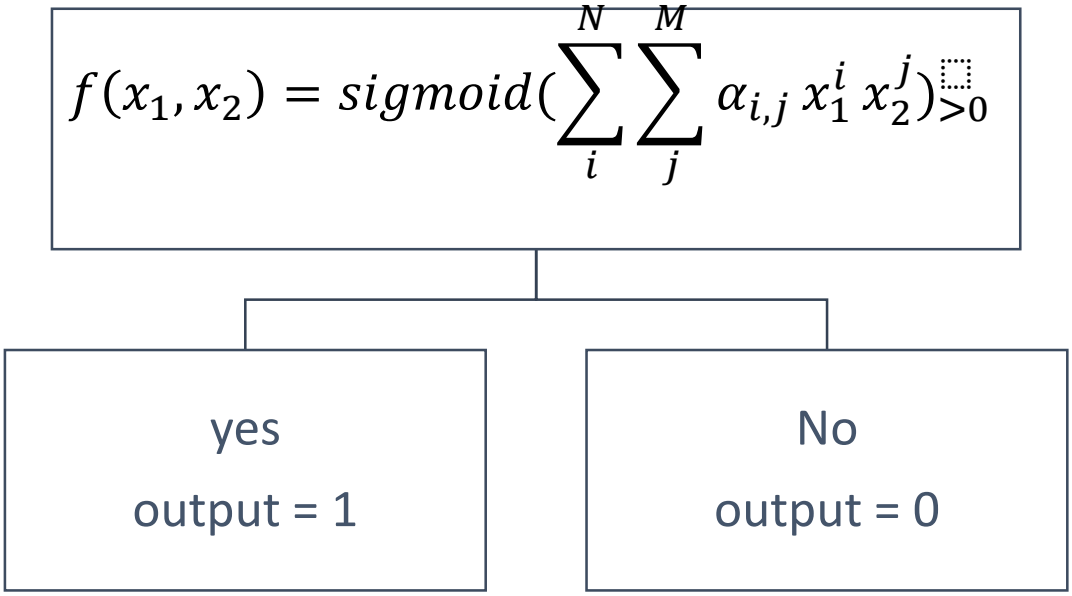
فرض کنید بین هر علامت مثبت یا منفی یک مقدار صحیح است یعنی بعد از مثبت اول ۱ و بعد از مثبت دوم ۲ و بعد از مثبت سوم عدد ۳ است.



ب: نشان دهید درخت تصمیم چند متغیره با عمق یک وجود دارد که خطای آموزش آن صفر است.

اگر برای شناسایی مرز مورد نظر از رگرسیون لوجیستیک یا روش بردار های پشتیبان با متغییر های اضافه مثلا از مرتبه اول تا مرتبه ۳ استفاده گردد می توان مرز مورد نظر را به خوبی ایجاد کرد.

از ضرایب مورد نظر می توان به عنوان ریشه درخت استفاده نمود.



چندجمله ای بالا قابل حصول با استفاده از رگرسیون لوجستیک خواهد بود.

قسمت های علمی به خوبی در فایل نوتبوک پیاده سازی و توضیح داده شده اند.