

# گزارش تکلیف ID3 Algorithm

درس یادگیری ماشین

امیرحسین ابوالحسنی

۴۰۰۴۰۵۰۰۳

## فهرست مطالب

۳	۱	مقدمه
۳	۲	بررسی دیتاست
۳	۱.۲	آشنایی با ویژگی‌ها
۳	۲.۲	مقادیر هیچ مقدار
۴	۳.۲	نمودارها
۹	۴.۲	دسته بندی ویژگی‌ها
۹	۵.۲	حذف دستی برخی ویژگی‌ها
۱۰	۳	انتخاب ویژگی
۱۰	۴	آموزش مدل
۱۰	۱.۴	تقسیم دیتاست
۱۱	۲.۴	آموزش مدل روی همه ویژگی‌ها
۱۱	۳.۴	آموزش مدل روی ۴ ویژگی
۱۲	۴.۴	آموزش مدل روی ۸ ویژگی
۱۲	۵	نتایج
۱۳	۶	نتیجه گیری
۱۳	۷	مصورسازی دیتاست با PCA
۱۳	۸	مراجع

## ۱ مقدمه

درخت تصمیم گیری یک مدل یادگیری نظارت شده است که به طور گسترده ای در مسائل طبقه بندی مورد استفاده قرار می گیرد. الگوریتم ID3 یکی از پرکاربردترین الگوریتم های ساخت درخت تصمیم می باشد. این الگوریتم با استفاده از معیار انتروپی<sup>۱</sup> بهترین ویژگی را برای تقسیم گره انتخاب می کند و به طور بازگشتی این فرایند را تا زمان رسیدن به یکی از شرط های پایه انجام می دهد. در این گزارش، ابتدا به بررسی دیتاست و پیش پردازش های روی آن پرداخته می شود، سپس توضیحی درباره شیوه Feature Selection داده می شود و در نهایت، نتایج هر درخت روی زیرمجموعه ای از ویژگی ها بررسی می گردد.

## ۲ بررسی دیتاست

### ۱.۲ آشنایی با ویژگی ها

در این تکلیف دیتاست با نام Salary مورد استفاده قرار می گیرد. این دیتاست متشکل از ۳۲۵۶۱ نمونه، ۱۵ ویژگی افراد را همراه با کلاس درآمد سالانه شان ثبت کرده است.

نام ویژگی	نوع ویژگی	تعداد مقادیر یکتا	نمونه مقدار
age	عددی		۵۰
workclass	گسسته	۹	Federal-gov
fnlwgt	عددی		۷۷۵۱۶
education	گسسته	۱۶	HS-grad
education-num	گسسته	۱۶	۳
marital-status	گسسته	۷	Married-spouse-absent
occupation	گسسته	۱۵	Tech-support
relationship	گسسته	۶	Wife
race	گسسته	۵	White
sex	گسسته	۲	Male
capital-gain	عددی		۱۰۵۶۶
capital-loss	عددی		۹۷۴
hours-per-week	عددی		۸۸
native-country	گسسته	۲	England
salary	گسسته	۲	<=50K, >50K

جدول ۱: ویژگی های دیتاست salary

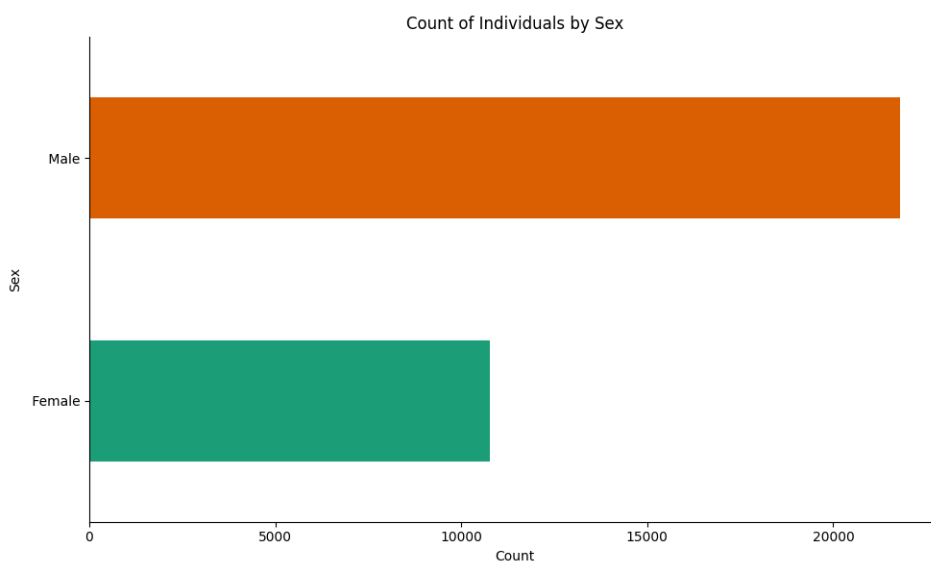
### ۲.۲ مقادیر هیچ مقدار

خوشبختانه این دیتاست دارای هیچ سلول گم شده ای نمی باشد.

<sup>۱</sup>Entropy

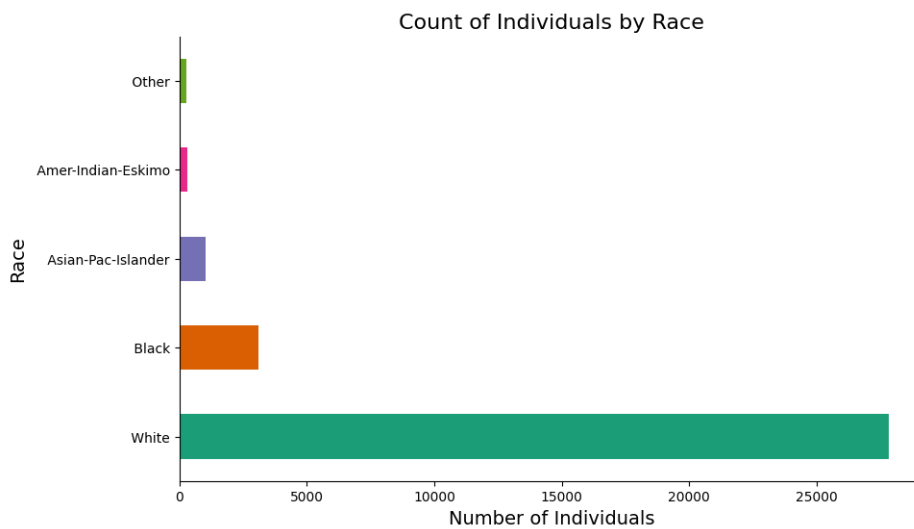
## ۳.۲ نمودارها

توزیع برخی ویژگی‌ها در دیتاست بررسی شده است. همانطور که در نمودار ۱ می‌توان دید، که جمعیت مردان دو برابر جمعیت زنان در این دیتاست می‌باشد.



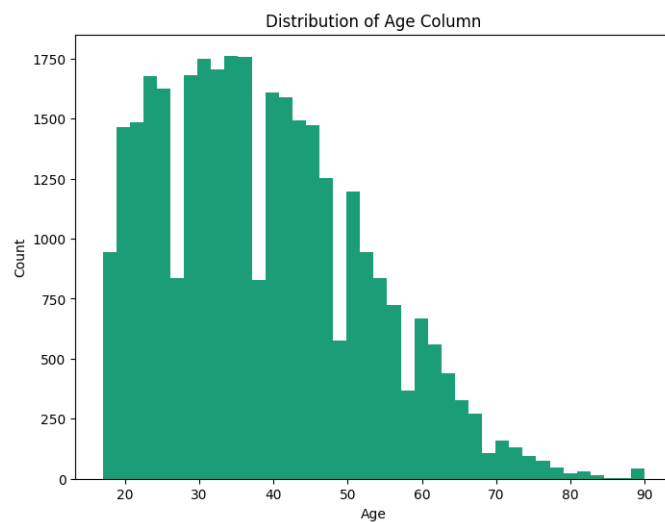
شکل ۱: توزیع ویژگی Sex

یکی از ویژگی‌های دیگر، نژاد هر نمونه در دیتاست می‌باشد، همانطور که در نمودار ۲ مشاهده می‌شود، افراد سفید پوست بیشترین افراد و افراد هندی-اسکیمو کمترین نژاد مشخص در این دیتاست هستند.



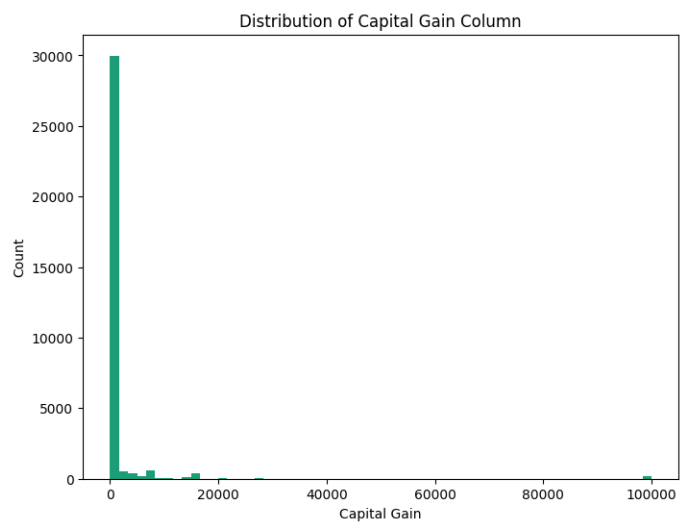
شکل ۲: توزیع ویژگی Race

یکی از مهمترین توزیع‌های این دیتاست، توزیع متغیر Age می‌باشد. همانطور که در نمودار ۳ مشاهده می‌شود، بیشتر نمونه‌ها در ۳۰ تا ۴۰ سالگی خود قرار دارند. و همچنین افراد زیر ۱۰ سال و بالای ۹۰ سال عضویت بسیار کمی در این دیتاست دارند.

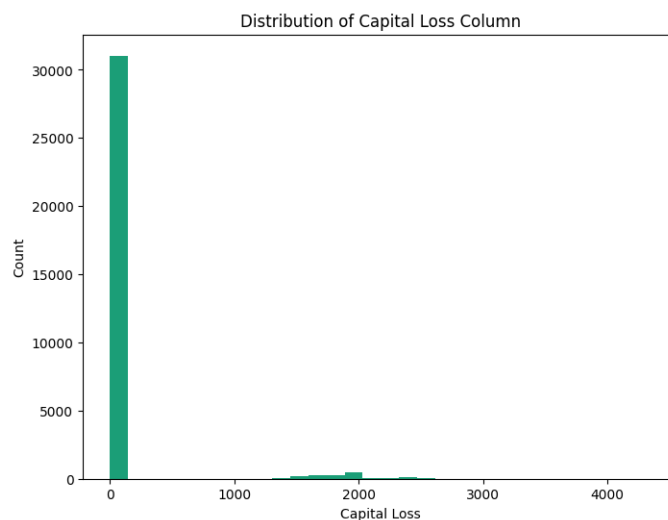


شکل ۳: توزیع ویژگی Age

همچنین توزیع ویژگی‌های افزایش سرمایه و کاهش سرمایه را در نمودارهای ۴ و ۵ می‌توان بررسی کرد. با توجه به ارتباط مالی با موضوع به نظر می‌رسد ویژگی‌های مرتبطی به تارگت باشند.

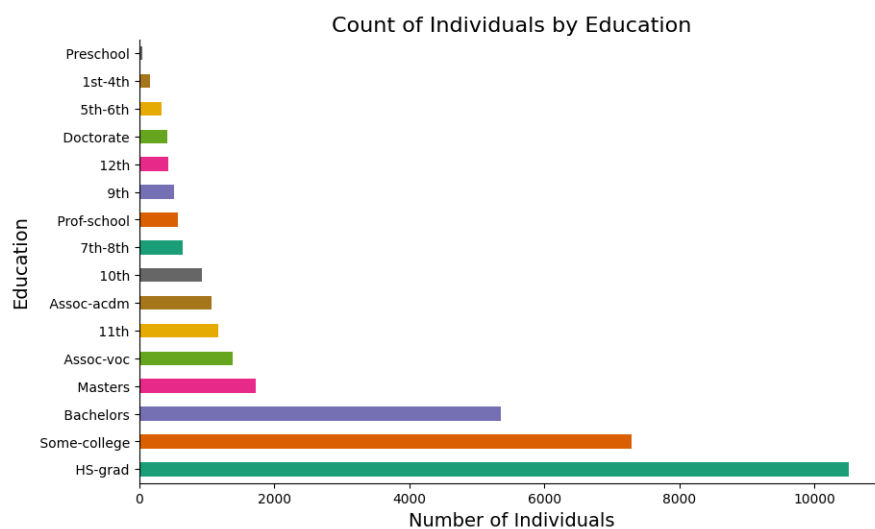


شکل ۴: توزیع ویژگی Capital Gain

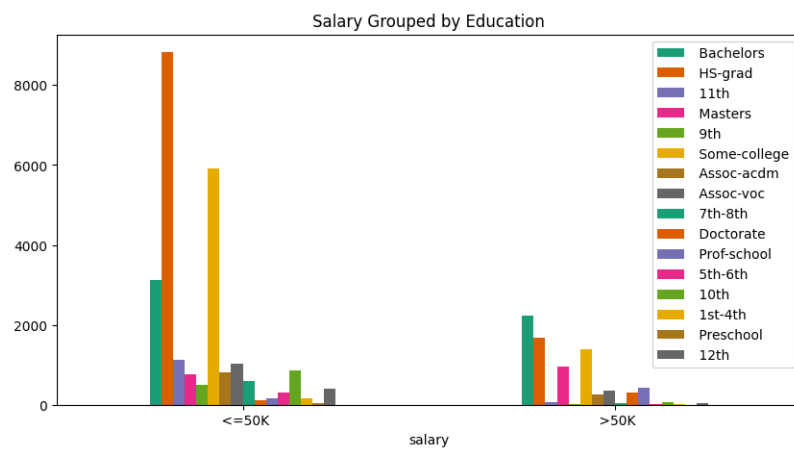


شکل ۵: توزیع ویژگی Capital Loss

یکی دیگر از ویژگی‌های مهم سطح تحصیلات فرد است که در کشورهایی که روابط منطق تا حد قابل قبولی در آن برقرار است!، معمولاً افرادی که سطح بالاتری از تحصیلات را دارا هستند جزو افرادی هستند که درآمد خوبی دارند (نمودار ۷)، هرچند عکس این مورد صحیح نمی‌باشد.

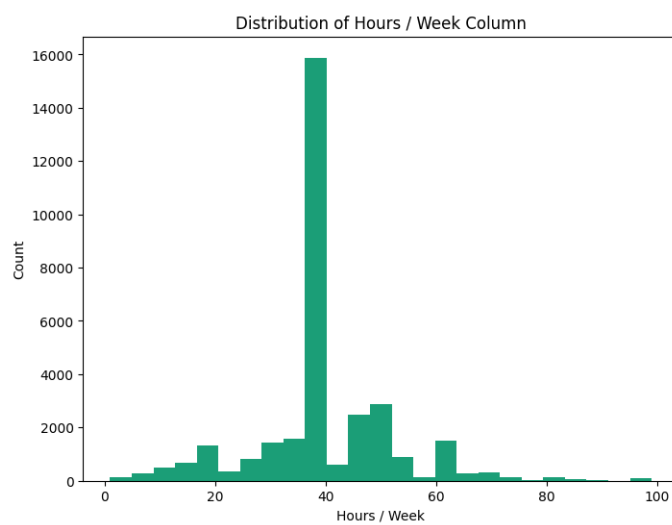


شکل ۶: توزیع ویژگی Education



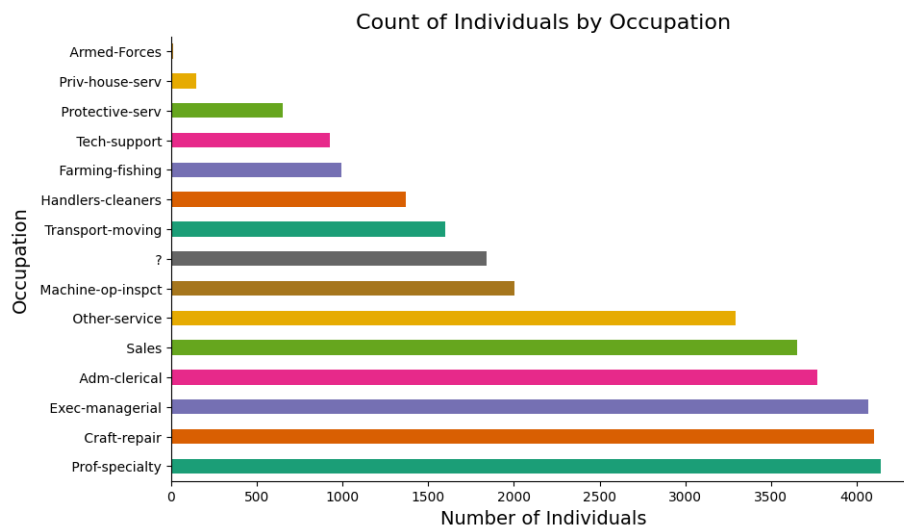
شکل ۷: توزیع ویژگی Salary بر اساس Education

همچنین توزیع ساعت کار روزانه نمونه‌ها در نمودار ۸ نشان داده شده است.



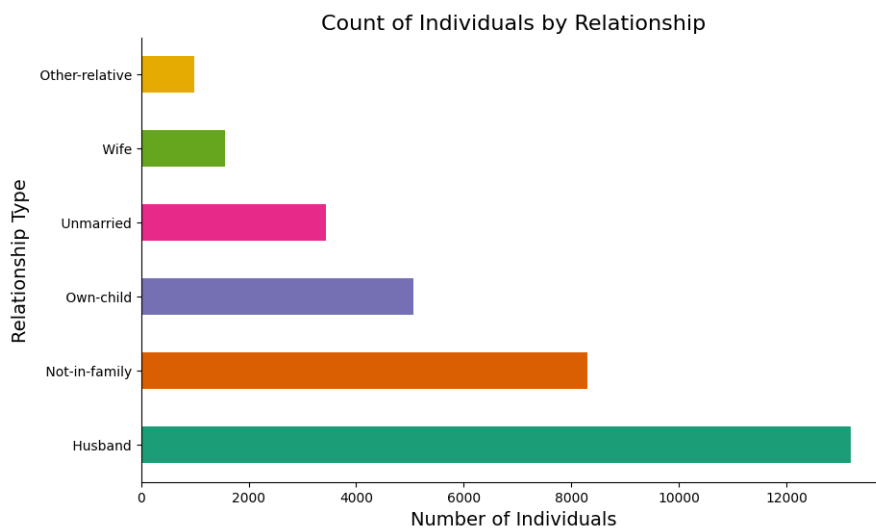
شکل ۸: توزیع ویژگی Hours Per Week

از دیگر ویژگی‌های تقریباً مرتبط می‌توان به نوع شغل افراد اشاره کرد که توزیع آن در نمودار ۹ نشان داده شده است.



شکل ۹: توزیع ویژگی Occupation

یکی از ویژگی‌های کلیدی که بعداً توسط درخت به دست می‌آید، ویژگی Relationship می‌باشد. (نمودار ۱۰)



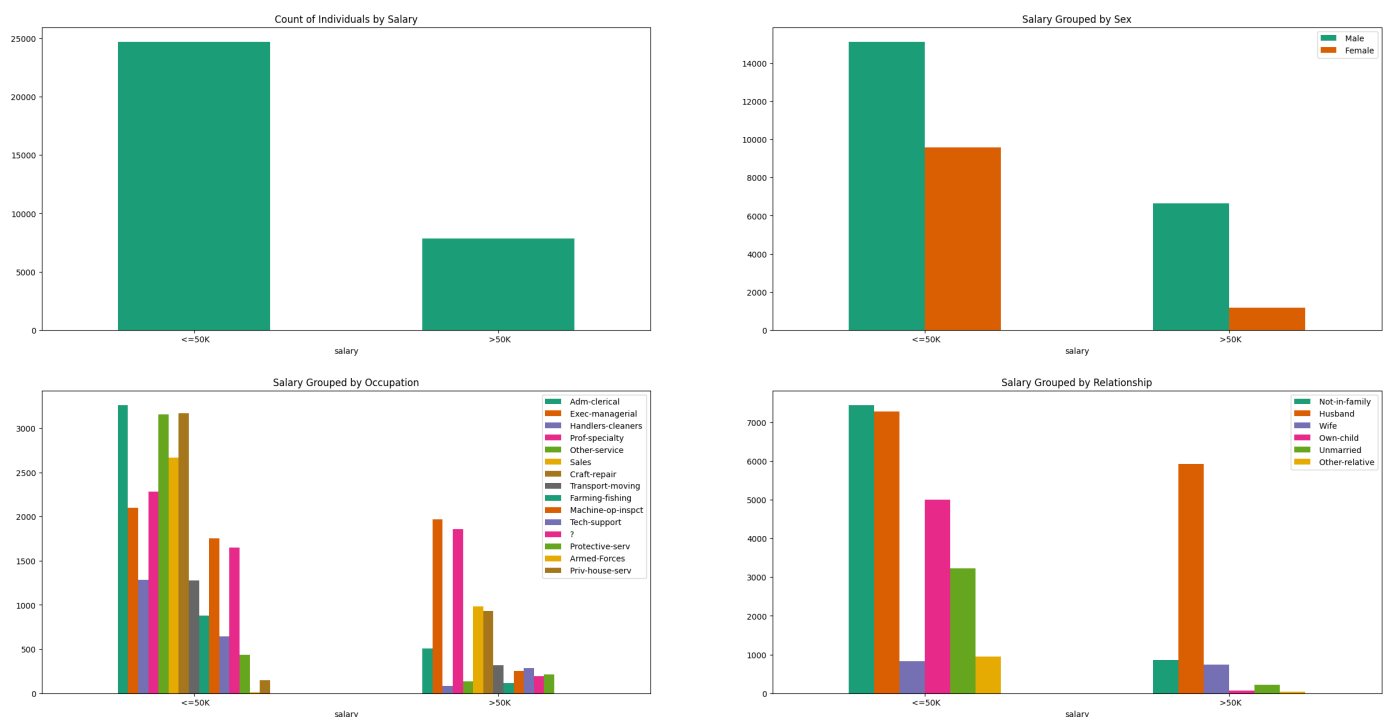
شکل ۱۰: توزیع ویژگی Relationship

در انتها برای جمع بندی نمودارها سعی شده توزیع کلاس‌های ویژگی هدف بررسی شود. همانطور که مشاهده می‌شود، دیتا ست به هیچ وجه بالانس نمی‌باشد و داده‌های کلاس مینور<sup>۱</sup> مربوط به کلاس درآمد بالاتر می‌باشد.

همچنین در نمودار ۱۱ توزیع کلاس هدف با توجه به سه ویژگی نشان داده شده تا درک بهتری از رابطه هر ویژگی با هر کلاس ویژگی هدف به دست بیاید.

<sup>۱</sup>Minor





شکل ۱۱: توزیع ویژگی Salary طبق ویژگی‌های Occupation, Relationship, Sex

## ۴.۲ دسته بندی ویژگی‌ها

برای کار با درخت تصمیم نیاز به این است که داده‌ها گسسته باشند. با تعیین بازه‌هایی، ویژگی‌های Age, Hours per Week, Capital Gain گسسته سازی شدند. در جداول ۲ و ۳ و ۴ مقادیر هر ویژگی و بازه‌های گسسته‌سازی نشان داده شده است.

$(0, 30]$	$(30, 50]$	$(50, \infty)$
۱ - ۳۰	۳۱ - ۵۰	Over 50

جدول ۲: گسسته سازی Age

$(0, 20]$	$(20, 40]$	$(40, 60]$	$(60, \infty)$
Low	Average	High	Very High

جدول ۳: گسسته سازی  $\frac{\text{Hours}}{\text{Week}}$

$(0, 15000]$	$(15000, \infty)$
$\leq 15K$	$> 15K$

جدول ۴: گسسته سازی Capital Gain

## ۵.۲ حذف دستی برخی ویژگی‌ها

در اینجا به علل حذف سه ویژگی fmlwgt و education-num و capital-loss اشاره می‌گردد.

- education-num: این ویژگی بدین علت که با ویژگی Education یکی است. باعث ایجاد افزونگی می‌شود.
- capital-loss: با نگاه به نمودار ۵ می‌توان استنتاج کرد که حجم ضرری که افراد متحمل شدند آنقدر زیاد نیست که در درآمد سالانه آنها تاثیر بگذارد، اما بالعکس، چگم capital gain با توجه به اینکه در یکسری افراد، خیلی بالاست، قابل تاثیر گذاری در درآمد سالانه فرد می‌باشد.

### ۳ انتخاب ویژگی<sup>۱</sup>

برای این بخش، از معیاری به نام آزمون کای-دو<sup>۲</sup> برای انتخاب مجموعه از ویژگی‌ها که بیشترین ارتباط را با متغیر هدف دارند، استفاده شده است. آزمون کای-دو یکی از روش‌های آماری پرکاربرد است که برای تحلیل داده‌های کیفی و بررسی روابط بین متغیرهای گسسته استفاده می‌شود. این آزمون به طور گسترده در حوزه‌های مختلف از جمله یادگیری ماشین، تحلیل داده و تحقیقات علمی مورد استفاده قرار می‌گیرد. در زمینه انتخاب ویژگی در درخت تصمیم و الگوریتم ID3، آزمون کای-دو برای اندازه‌گیری میزان وابستگی بین ویژگی‌ها و متغیر هدف استفاده می‌شود. این آزمون به ما کمک می‌کند تا تشخیص دهیم کدام ویژگی‌ها ارتباط قوی‌تری با متغیر هدف دارند.<sup>۳</sup>

فرمول آزمون کای-دو به صورت زیر است:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

مقدار معیار کای-دو هر ویژگی در جدول ۵ قابل مشاهده است.

$\chi^2$	ویژگی
۳۶۵۹	Relationship
۱۸۶۶	Capital Gain
۱۱۲۳	Marital Status
۱۱۱۳	Age
۵۰۴	Occupation
۵۰۲	Sex
۲۹۷	Education
۱۹۹	Hours per Week
۴۷	Work class
۳۳	Race
۱۳	Native Country

جدول ۵: مقادیر  $\chi^2$  برای هر ویژگی

### ۴ آموزش مدل

قبل از آموزش مدل، با تغییری در کد تابع id3 قابلیت هرس کرن بر اساس یک آستانه برای Information Gain را در تابع ایجاد کردیم. این به مدل کمک می‌کند تا از بیش برآزش<sup>۴</sup> جلوگیری کند. همچنین باعث کم شدن عمق درخت می‌شود که در نهایت به پیچیدگی حافظه درخت نیز کمک خواهد کرد.

#### ۱.۴ تقسیم دیتاست

برای اینکه فاز Evaluation عادلانه باشد، دیتاست به دو بخش (Train ( 80% و Test ( 20% تقسیم می‌شود.

<sup>۱</sup> Feature Selection

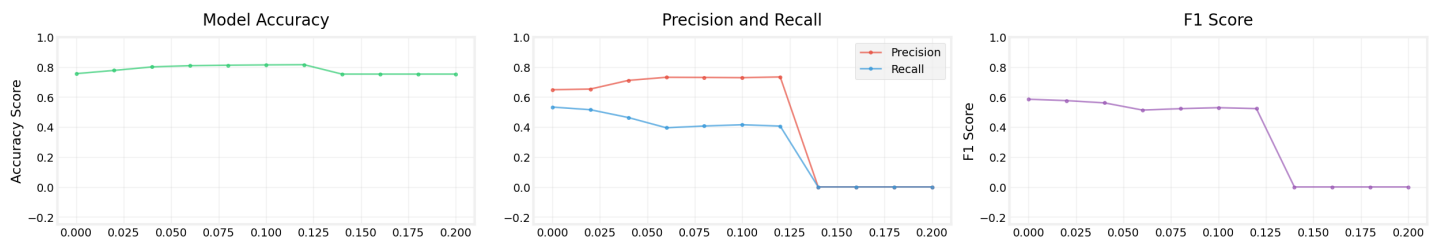
<sup>۲</sup> Chi Square Test

<sup>۳</sup> توضیحات مربوط به آزمون کای-دو توسط مدل Claude نوشته شده است.

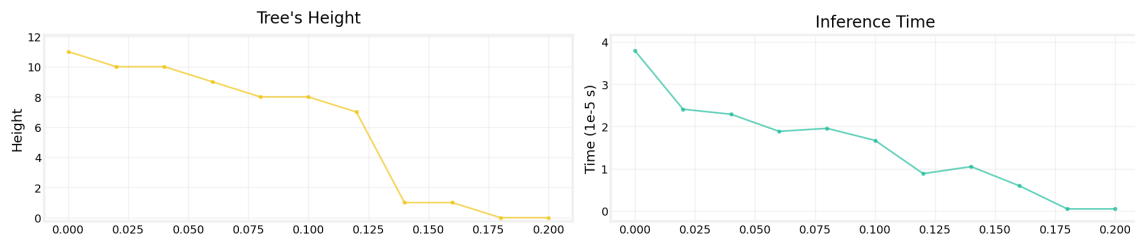
<sup>۴</sup> Overfit

## ۲.۴ آموزش مدل روی همه ویژگی‌ها

برای داشتن یک پایه برای مقایسه، ابتدا درخت را روی همه ویژگی‌ها آموزش می‌دهیم. همچنین برای دیدن تاثیر تغییر آستانه Information Gain، (و به نوعی تاثیر بیش‌برازش در دقت و F1 Score) در هر مرحله این مقدار را از ۰ به ۰.۲ با قدم‌های ۰.۰۲ برده شده است. (شکل ۱۲ و ۱۳)



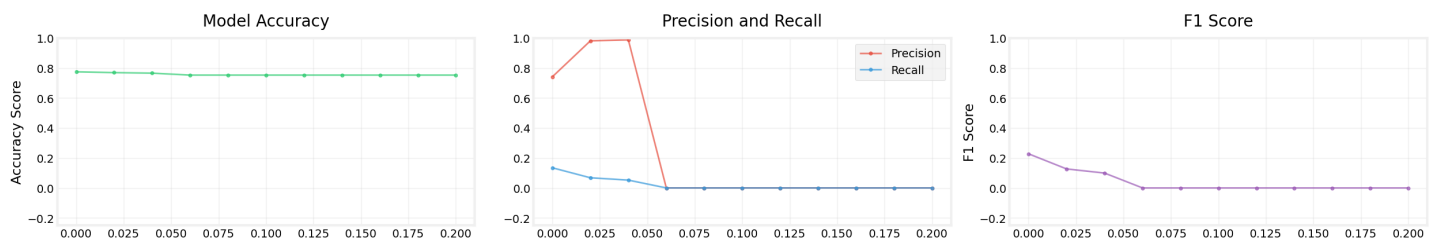
شکل ۱۲: نمودارهای Accuracy, F1 Score, Precision & Recall با توجه به مقدار آستانه Information Gain



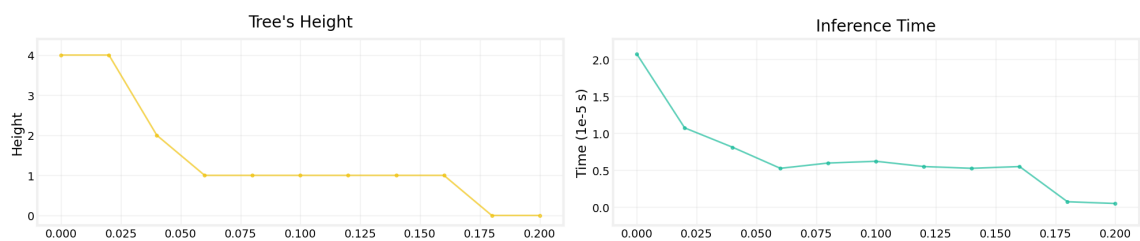
شکل ۱۳: نمودارهای Inference Time و ارتفاع درخت، با توجه به مقدار آستانه Information Gain

## ۳.۴ آموزش مدل روی ۴ ویژگی

پس از اینکه با استفاده از معیار  $\chi^2$  ویژگی‌ها را طبقه‌بندی کردیم. در این بخش چهار تا از بهترین ویژگی‌ها را انتخاب کرده و با آنها درخت تصمیم جدید را تشکیل می‌دهیم. در شکل ۱۴ و ۱۵ نمودارهای مختلف روی مقادیر مختلف آستانه Information Gain نشان داده می‌شود.



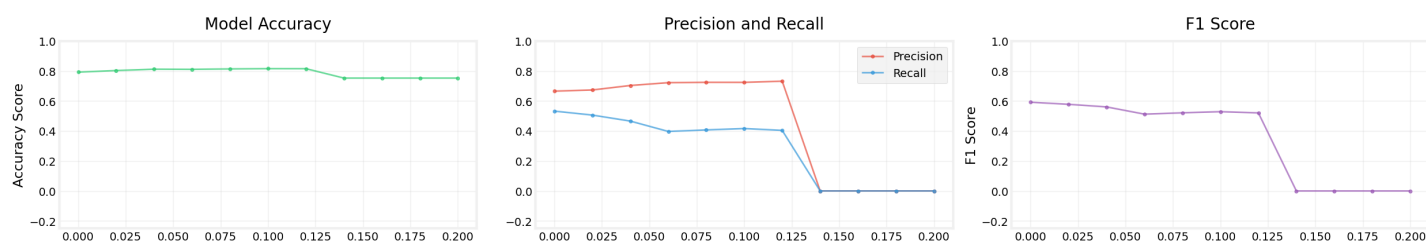
شکل ۱۴: نمودارهای Accuracy, F1 Score, Precision & Recall با توجه به مقدار آستانه Information Gain



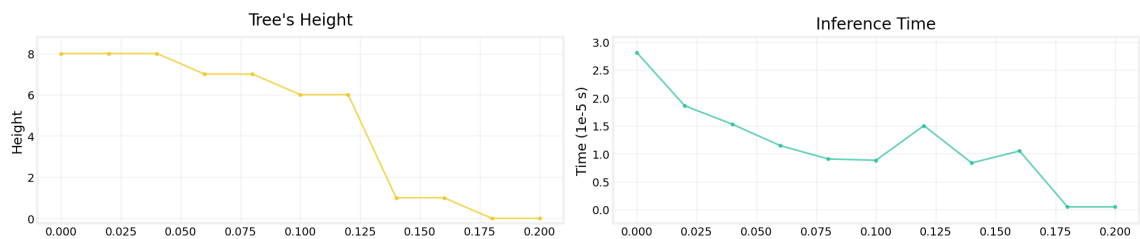
شکل ۱۵: نمودارهای Inference Time و ارتفاع درخت، با توجه به مقدار آستانه Information Gain

## ۴.۴ آموزش مدل روی ۸ ویژگی

اینبار هشت تا از بهترین ویژگی‌ها را انتخاب کرده و با آنها درخت تصمیم جدید را تشکیل می‌دهیم. در شکل ۱۶ و ۱۷ نمودارهای مختلف روی مقادیر مختلف آستانه Information Gain نشان داده می‌شود.



شکل ۱۶: نمودارهای Accuracy, F1 Score, Precision & Recall با توجه به مقدار آستانه Information Gain



شکل ۱۷: نمودارهای Inference Time و ارتفاع درخت، با توجه به مقدار آستانه Information Gain

## ۵ نتایج

نمره‌های بهترین مدل‌های بخش 4 در جدول ۶ جمع آوری شده است.

متد استفاده شده	آستانه IG	ارتفاع	$(\times 10^{-5})$ زمان استنتاج	دقت	Precision	Recall	F1 Score
معمولی	۰.۱۲	۷	۰.۸	۰.۸۱	۰.۷۳	۰.۴۰	۰.۵۲
$\chi^2$ - بهترین ۸ تا	۰.۱	۶	۰.۸۸	۰.۸۱	۰.۷۲	۰.۴۱	۰.۵۳
$\chi^2$ - بهترین ۴ تا	۰.۰	۴	۲	۰.۸۱	۰.۷۳	۰.۱۳	۰.۲۲

جدول ۶: جدول مقایسه نمره‌های بهترین مدل‌ها

۶	نتیجه گیری
۷	مصورسازی دیتاست با PCA <sup>۱</sup>
۸	مراجع