درس یادگیری ماشین

گزارش تکلیف Decision Tree

استاد درس:
دکتر افتخاری

نگارش:
امیرحسین ابوالحسنی
شماره دانشجویی: ۴۰۰۴۰۵۰۰۳

پاییز 1403

فهرست مطالب

۱ مقدمه

درخت تصمیم گیری یک مدل یادگیری نظارت شده است که به طور گسترده ای در مسائل طبقه‌بندی مورد استفاده قرار می‌گیرد. الگوریتم ID3 یکی از پرکاربرد ترین الگوریتم‌های ساخت درخت تصمیم می باشد. این الگوریتم با استفاده از معیار انتروپی [۱] بهترین ویژگی را برای تقسیم گره انتخاب می کند و به طور بازگشتی این فرایند را تا زمان رسیدن به یکی از شرط‌های پایه انجام می‌دهد.
در این گزارش، ابتدا به بررسی دیتاست و پیش پردازش های روی آن پرداخته می‌شود، سپس توضیحی درباره شیوه Feature Selection داده می‌شود و در نهایت، نتایج هر درخت روی زیرمجموعه‌ای از ویژگی‌ها بررسی می گردد.

۲ بررسی دیتاست

۱.۲ آشنایی با ویژگی‌ها

در این تکلیف دیتاست با نام Salary مورد استفاده قرار می‌گیرد. این دیتاست متشکل از ۳۲۵۶۱ نمونه، ۱۵ ویژگی افراد را همراه با کلاس درامد سالانه‌شان 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	۴گسسته	۲	England
salary	۲گسسته	۲	<=50K, >50K

جدول ۱: ویژگی‌های دیتاست salary

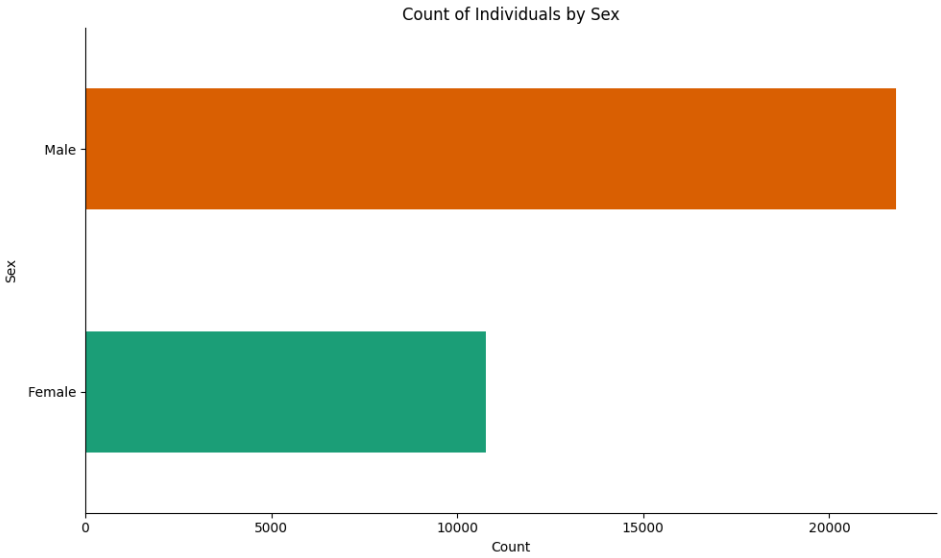
۲.۲ مقادیر هیچ مقدار

خوشبختانه این دیتاست داری هیچ مقدار گم شده‌ای نمی‌باشد.

[۱] Entropy

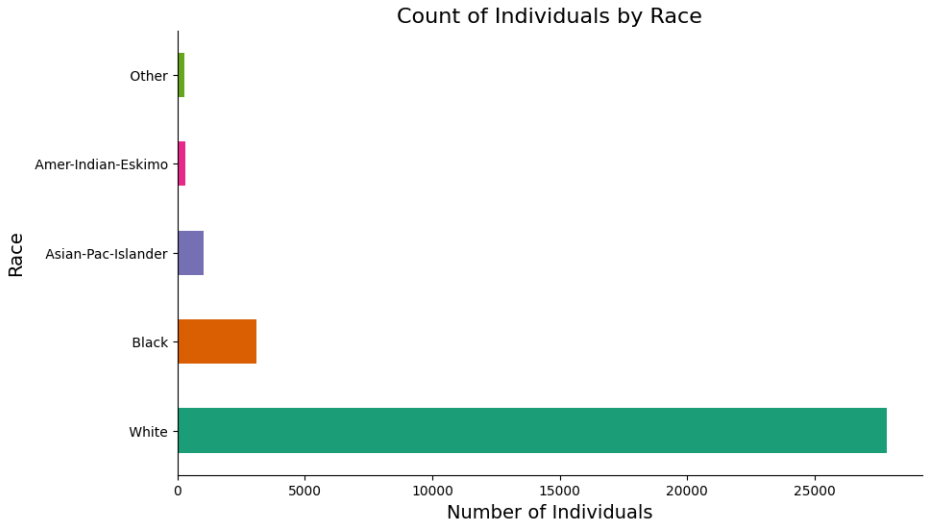
۳.۲ نمودارها

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



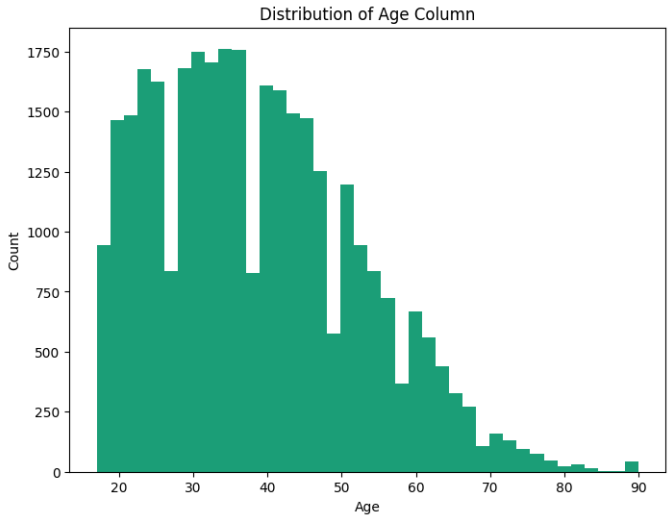
شکل ۱: توزیع ویژگی Sex

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



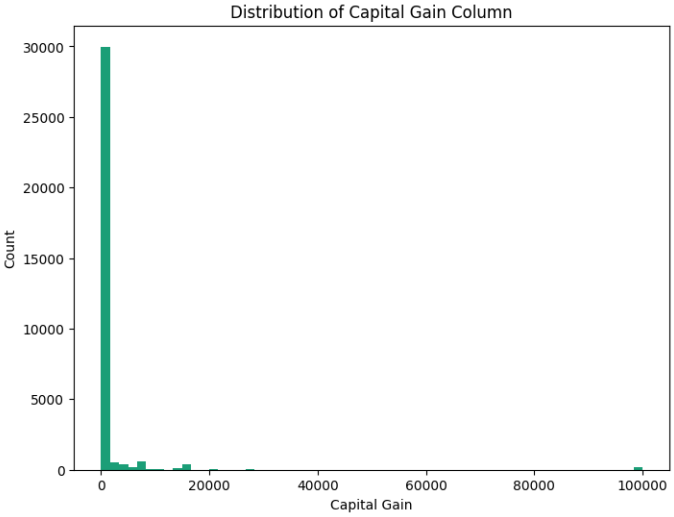
شکل ۲: توزیع ویژگی Race

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.

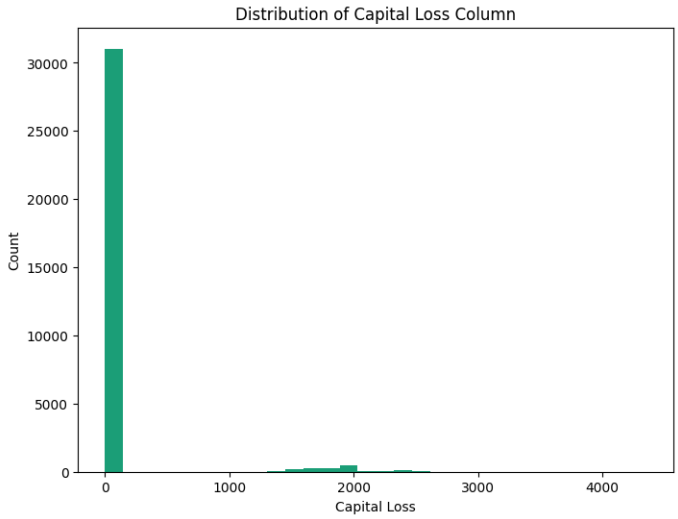


شکل ۳: توزیع ویژگی Age

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.

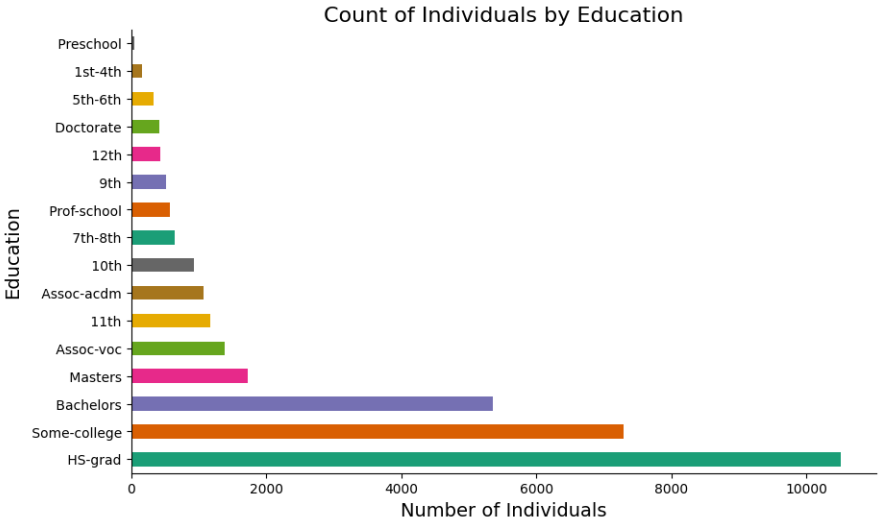


شکل ۴: توزیع ویژگی Capital Gain

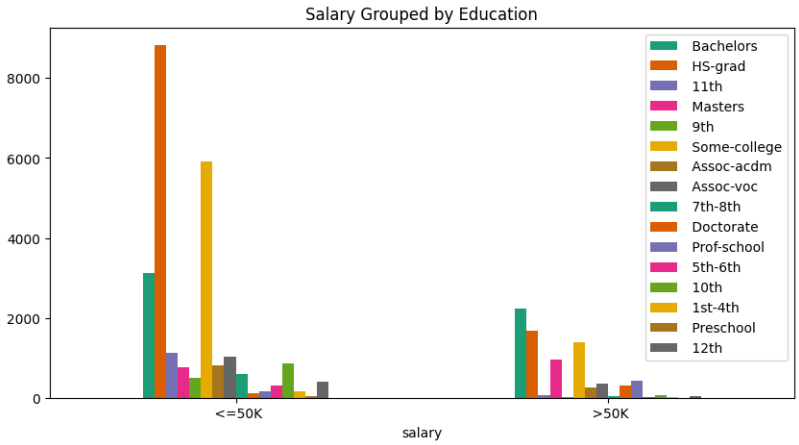


شکل ۵: توزیع ویژگی Capital Loss

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

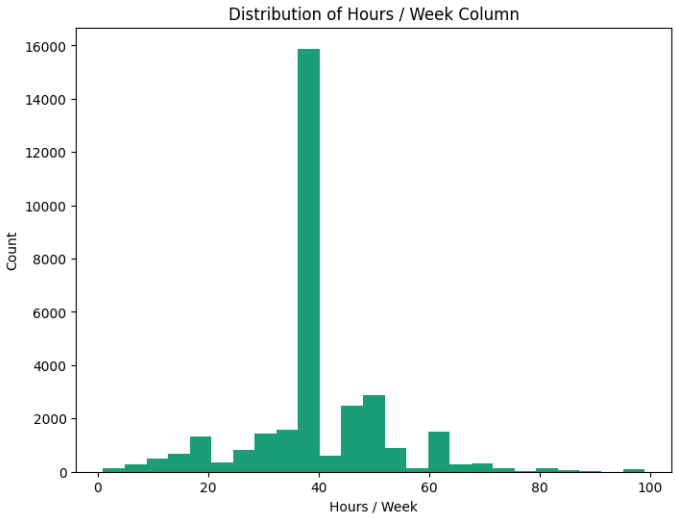


شکل ۶: توزیع ویژگی Education



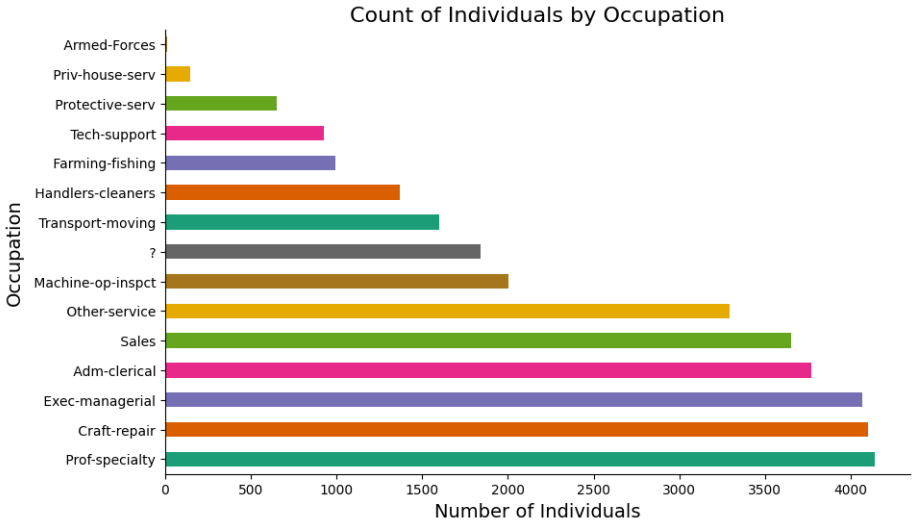
شکل ۷: توزیع ویژگی Salary بر اساس Education

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



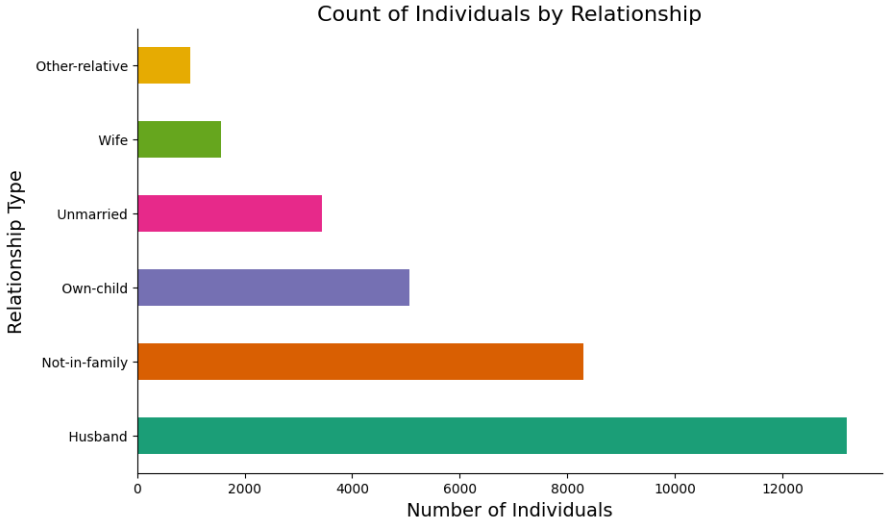
شکل ۸: توزیع ویژگی Hours Per Week

از دیگر ویژگی‌های تقریبا 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 بعدا توسط درخت به دست می‌آید، ویژگی Relationship می باشد.(نمودار ۱۰)

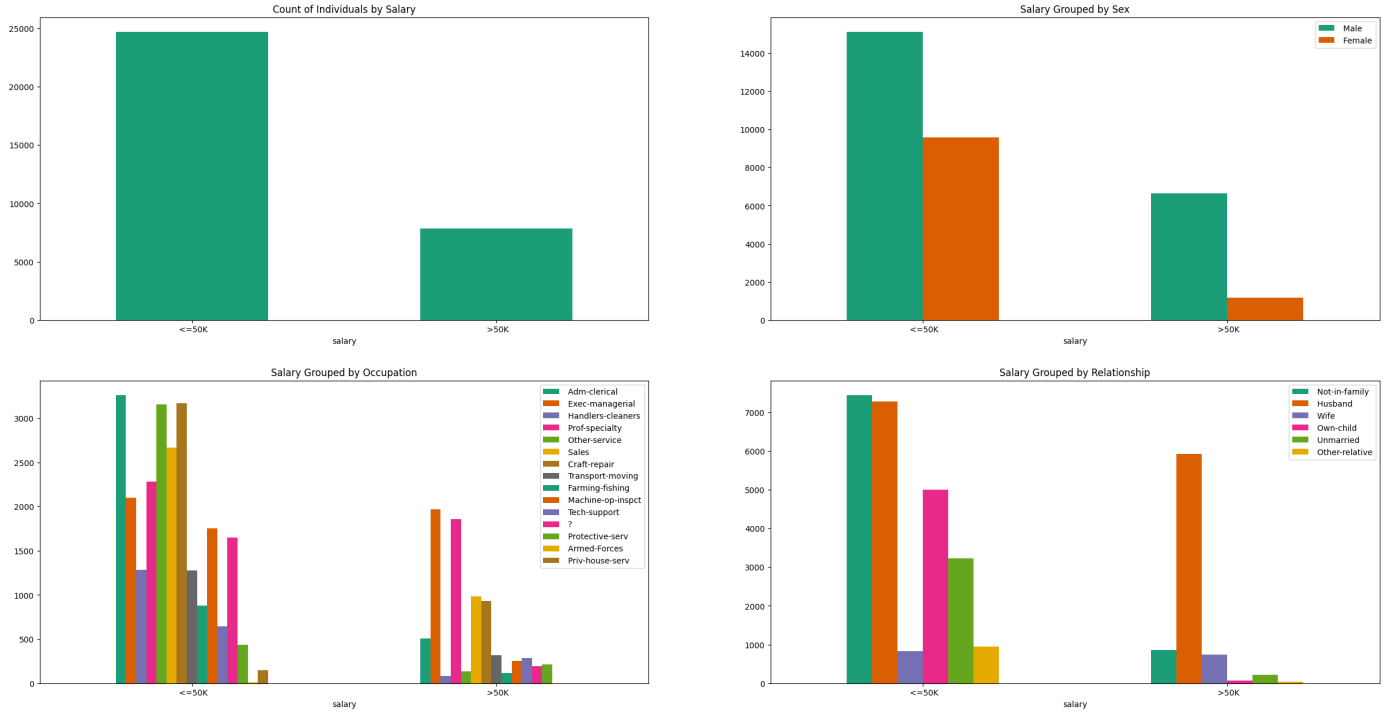


شکل ۱۰: توزیع ویژگی Relationship

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 [۱] مربوط به کلاس درامد بالاتر می‌باشد.

همچمنین 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.

Minor[۱]



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.

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

جدول ۲: گسسته سازی Age

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

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

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

جدول ۴: گسسته سازی Capital Gain

۵.۲ حذف دستی برخی ویژگی‌ها

در اینجا به علل حذف سه ویژگی fnlwgt 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، حجم capital gain با توجه به اینکه در یکسری افراد، خیلی بالاست، قابل تاثیر گذاری در درآمد سالانه فرد می‌باشد.

۳ انتخاب ویژگی [۱]

برای این بخش، از معیاری به نام آزمون کای-دو [۲] برای انتخاب مجموعه از 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. [۳]
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.

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

جدول ۵: مقادیر χ^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 برازش [۴] جلوگیری کند. همچنین باعث کم شدن عمق درخت می‌شود که در نهایت به پیچیدگی حافظه درخت نیز کمک خواهد کرد.

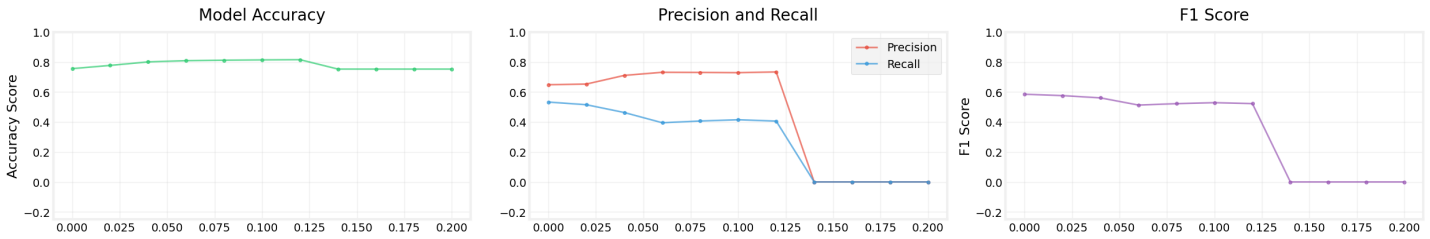
۱.۴ تقسیم دیتاست

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

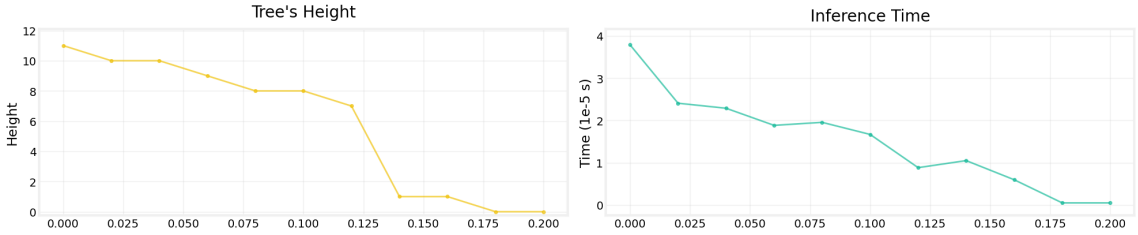
---
[۱] Feature Selection
[۲] Chi Square Test
[۳] توضیحات مربوط به آزمون کای-دو توسط مدل Claude نوشته شده است.
[۴] Overfit

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)



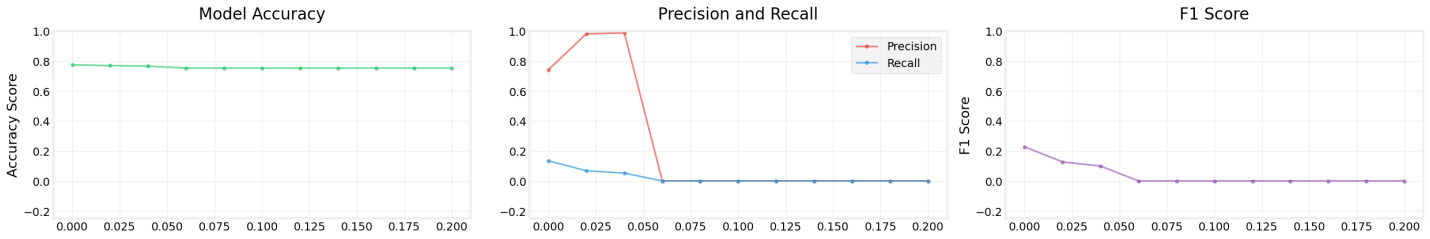
شکل ۱۲: نمودارهای Accuracy, F1 Score, Precision & Recall با توجه به مقدار آستانه Information Gain



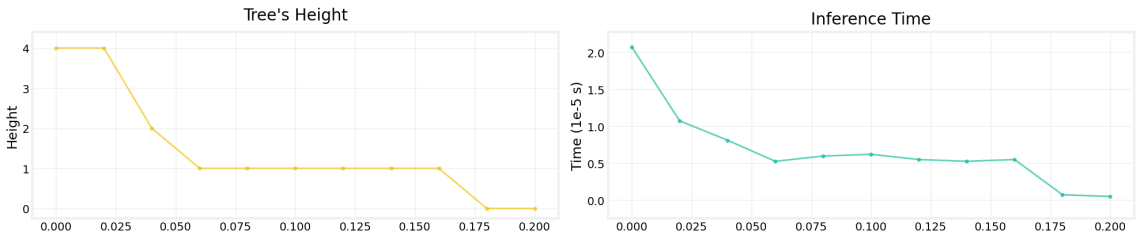
شکل ۱۳: نمودارهای Inference Time و ارتفاع درخت، با توجه به مقدار آستانه Information Gain

۳.۴ آموزش مدل روی ۴ ویژگی

پس از اینکه با استفاده از معیار χ^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.



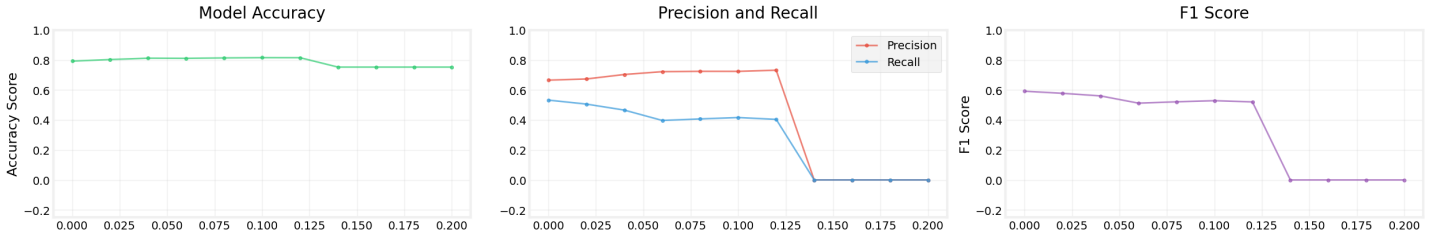
شکل ۱۴: نمودارهای Accuracy, F1 Score, Precision & Recall با توجه به مقدار آستانه Information Gain



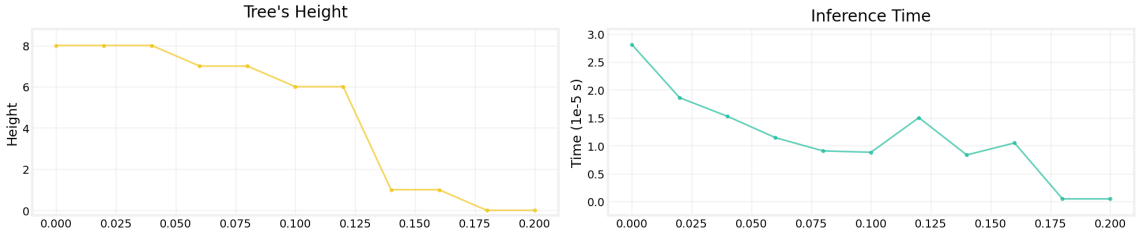
شکل ۱۵: نمودارهای Inference Time و ارتفاع درخت، با توجه به مقدار آستانه Information Gain

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



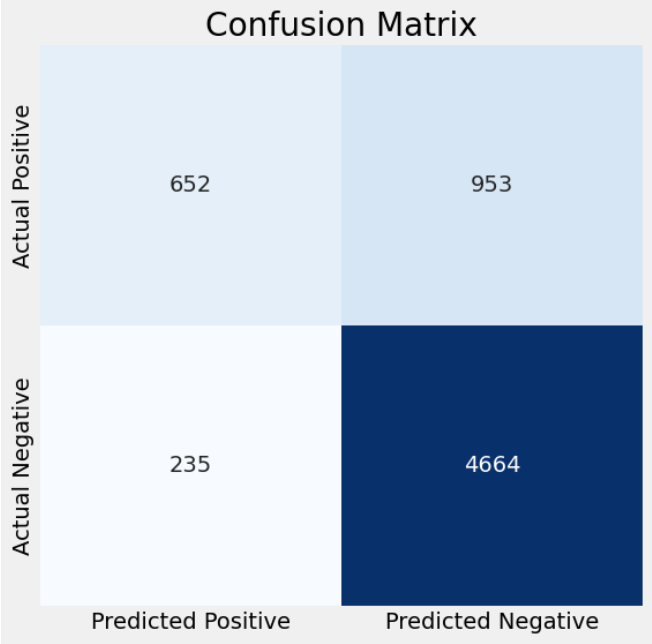
شکل ۱۷: نمودارهای Inference Time و ارتفاع درخت، با توجه به مقدار آستانه Information Gain

۵ نتایج

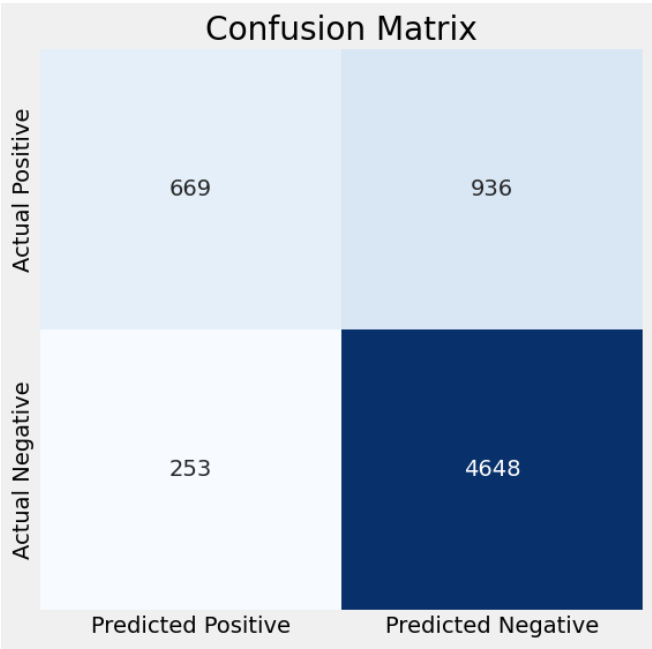
نمره‌های بهترین مدل‌های بخش 4 در جدول ۶ جمع آوری شده است.

متد استفاده شده	آستانه IG	ارتفاع	زمان استنتاج ($\times 10^{-5}$)	دقت	Precision	Recall	F1 Score
-	۰.۱۲	۷	۰.۸	۰.۸۱	۰.۷۳	۰.۴۰	۰.۵۲
بهترین ۸ تا - χ^2	۰.۱	۶	۰.۸۸	۰.۸۱	۰.۷۲	۰.۴۱	۰.۵۳
بهترین ۴ تا - χ^2	۰	۴	۲	۰.۷۷	۰.۷۴	۰.۱۳	۰.۲۲

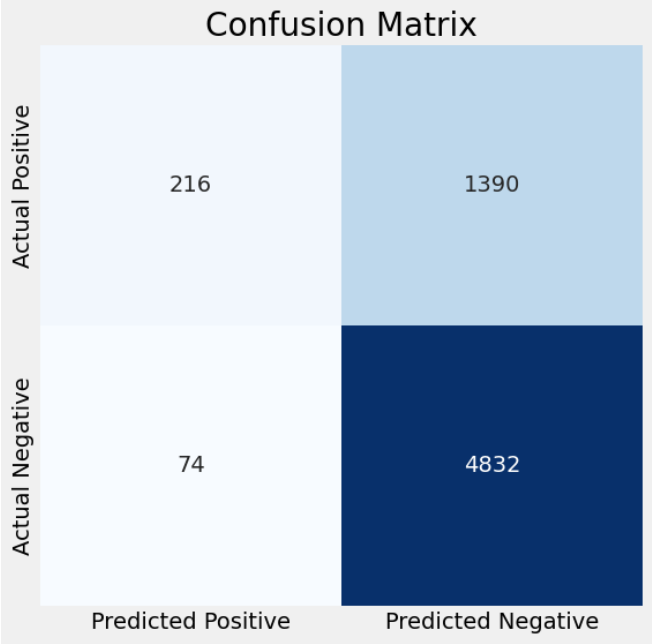
جدول ۶: جدول مقایسه نمره‌های بهترین مدل‌ها



شکل ۱۸: ماترس سردرگمی درخت معمولی



شکل ۱۹: ماترس سردرگمی درخت ساخته شده با ۸ ویژگی



شکل ۲۰: ماترس سردرگمی درخت ساخته شده با ۴ ویژگی

۶ نتیجه گیری

همانطور که در جدول ۶ دیده می‌شود، در این دیتاست دقت با کمتر شدن تعداد فیچرها بعد از ۸ ویژگی افت می‌کند، همچنین F1 Score ابتدا سیر صعودی گرفته اما در ۴ ویژگی به شدت کاهش پیدا می‌کند.
می‌توان از جدول ۶ نتیجه گرفت مدلی که با ۸ ویژگی اول جدول ۵ و با آستانه IG برابر با ۰.۱ آموزش ببیند، بهترین عملکرد را خواهد داشت.
یکی از تفسیرها برای مقدار کم Recall می‌تواند بالانس نبودن داده‌ها باشد. فرمول Recall بدین صورت است:

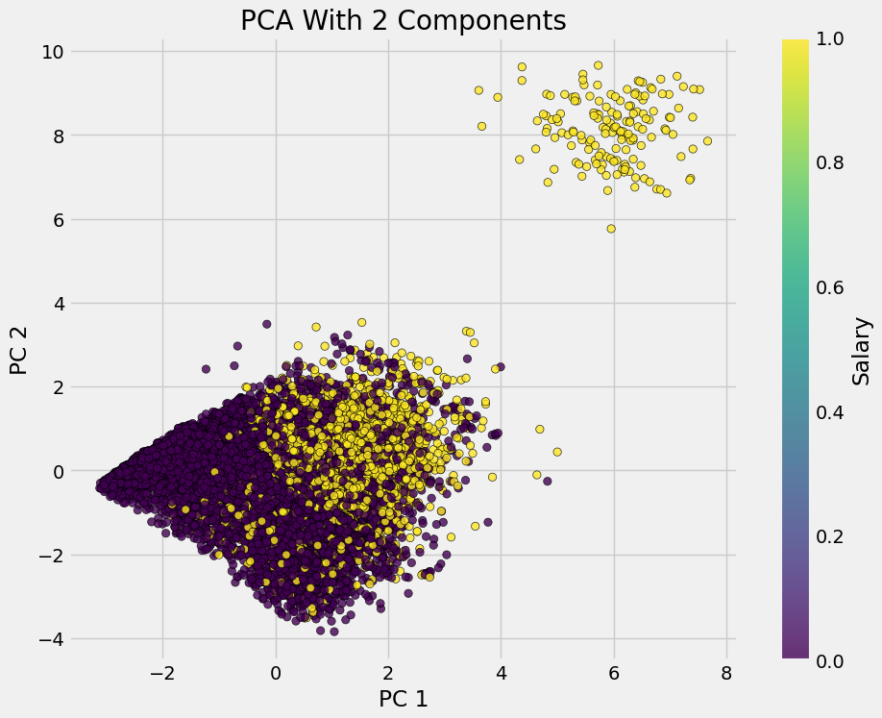
$$Recall = \frac{TP}{TP + FN}$$

زمانی که داده‌ها بالانس نباشد و کلاس مینور کلاس مثبت باشد، بایاس روی مقادیر کلاس منفی اتفاق می افتد و باعث می‌شود خیلی از نمونه‌ها منفی گزارش شوند، حتی اگر متعلق به کلاس مثبت باشند.(شکل ۲۰ و ۱۹ و ۱۸) این باعث می‌شود مقدار $False\ Negative$ بالا برود و با این اتفاق، مقدار Recall کم بشود.

۷ کنجکاوی: مصورسازی دیتاست با PCA [۱]

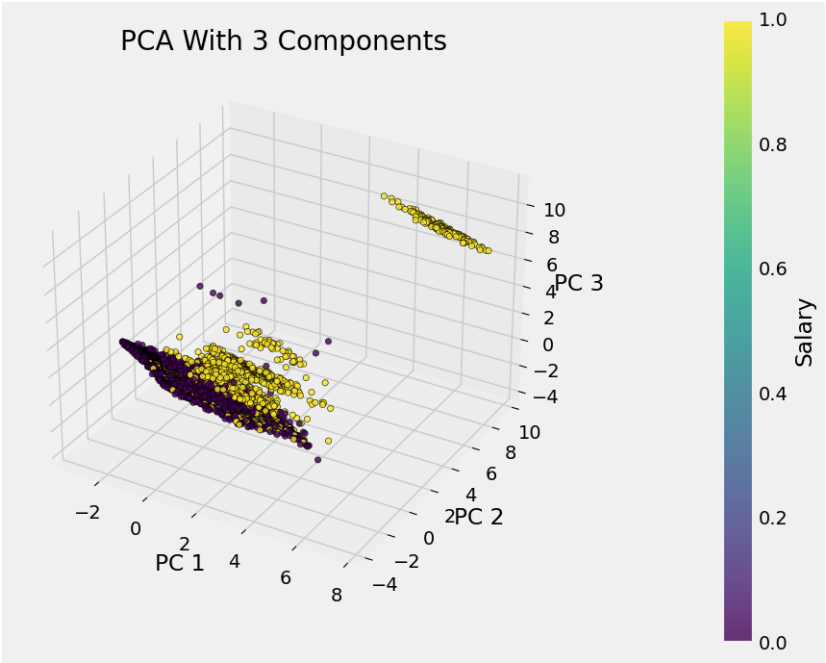
هرچند این مورد در توضیحات تکلیف ذکر نشده است اما حقیقتا مبحث کاهش ابعاد [۲] همیشه باعث شگفتی است!
در شکل ۲۱ دیده می‌شود که اگر PCA با دو مولفه اصلی انجام شود، جز معدودی از داده‌های کلاس مثبت، قابل جداسازی از داده‌های کلاس منفی نیستند.

---
[۱] Principle Component Analysis
[۲] Dimensionality Reduction



شکل ۲۱: نمودار پراکندگی داده‌ها پس از انجام PCA با دو مولفه

اما با ۳ مولفه، می‌توان در شکل ۲۲ دید که جداپذیری داده‌های هر کلاس بالاتر میرود، انگار که وجه دیگری از داده به نمایش گذاشته می‌شود.



شکل ۲۲: نمودار پراکندگی داده‌ها پس از انجام PCA با سه مولفه