

گزارش تکلیف ID3 Algorithm

درس یادگیری ماشین

امیرحسین ابوالحسنی

۴۰۰۴۰۵۰۰۳

فهرست مطالب

۱	مقدمه	۳
۲	بررسی دیتاست	۳
۱.۲	آشنایی با ویژگی‌ها	۳
۲.۲	مقادیر هیچ مقدار	۳
۳.۲	نمودارها	۴
۳	Feature Selection	۹
۴	آموزش مدل	۹
۵	نتایج	۹
۶	نتیجه گیری	۹
۷	مراجع	۹

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^۱ 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
<i>age</i>	عددی		۵۰
<i>workclass</i>	گسسته	۹	Federal-gov
<i>fnlwgt</i>	عددی		۷۷۵۱۶
<i>education</i>	گسسته	۱۶	HS-grad
<i>education-num</i>	گسسته	۱۶	۳
<i>marital-status</i>	گسسته	۷	Married-spouse-absent
<i>occupation</i>	گسسته	۱۵	Tech-support
<i>relationship</i>	گسسته	۶	Wife
<i>race</i>	گسسته	۵	White
<i>sex</i>	گسسته	۲	Male
<i>capital-gain</i>	عددی		۱۰۵۶۶
<i>capital-loss</i>	عددی		۹۷۴
<i>hours-per-week</i>	عددی		۸۸
<i>native-country</i>	گسسته	۲	England
<i>salary</i>	گسسته	۲	<=50K, >50K

جدول ۱: ویژگی های دیتاست salary

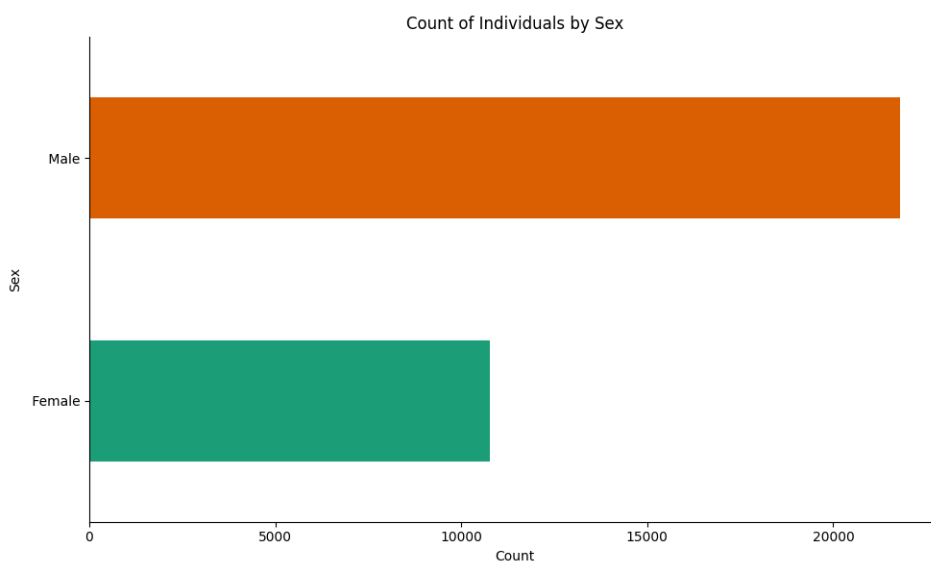
۲.۲ مقادیر هیچ مقدار

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

^۱Entropy

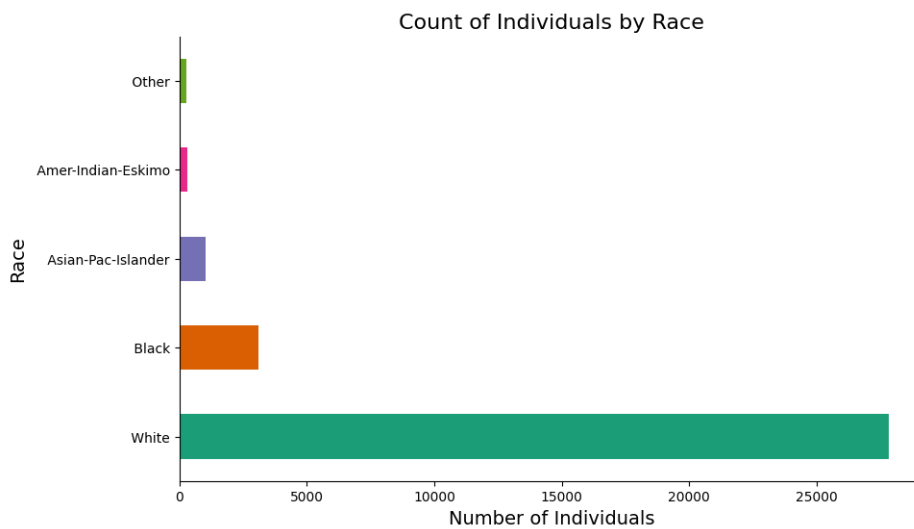
۳.۲ نمودارها

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.



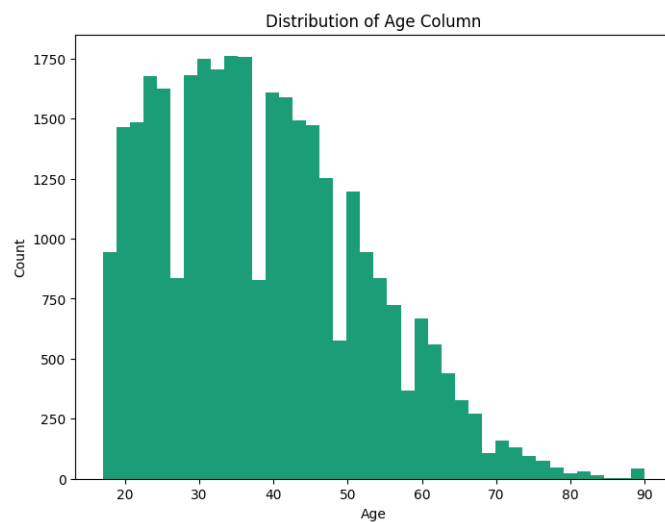
شکل ۱: توزیع ویژگی Sex

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.



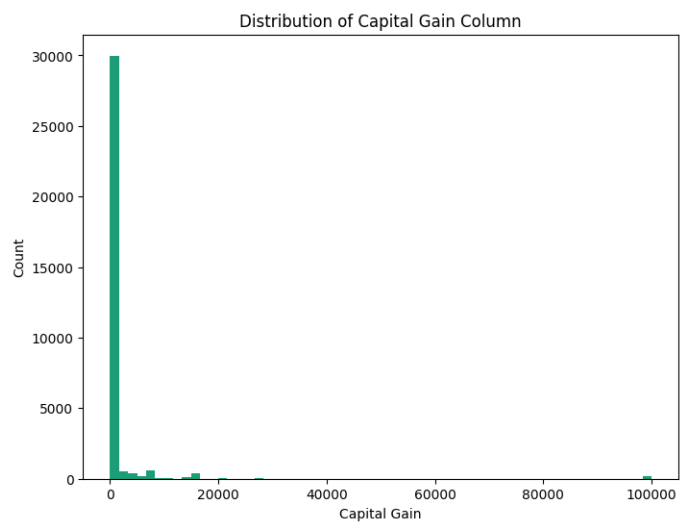
شکل ۲: توزیع ویژگی Race

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.

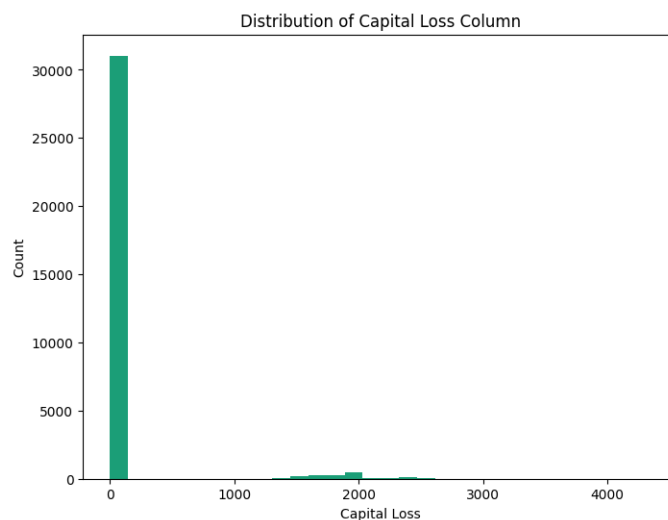


شکل ۳: توزیع ویژگی Age

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.

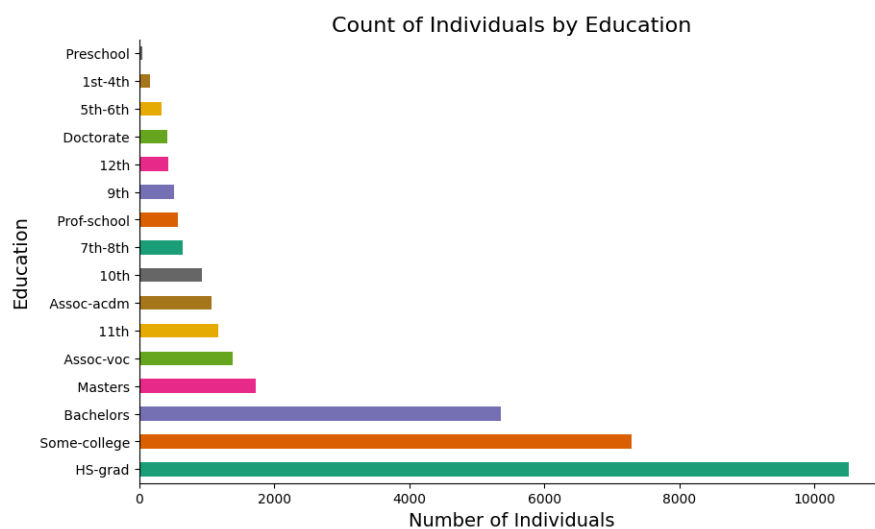


شکل ۴: توزیع ویژگی Capital Gain

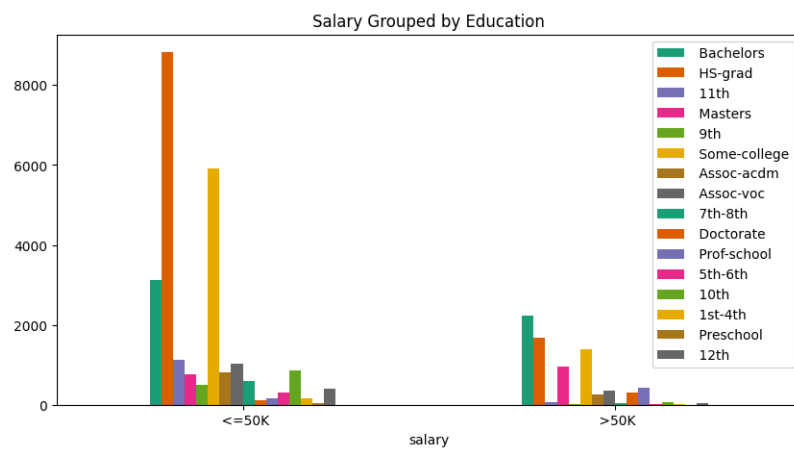


شکل ۵: توزیع ویژگی Capital Loss

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.

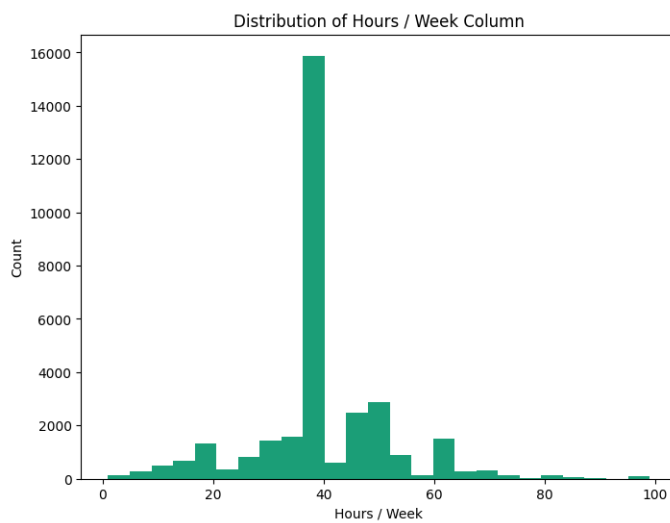


شکل ۶: توزیع ویژگی Education



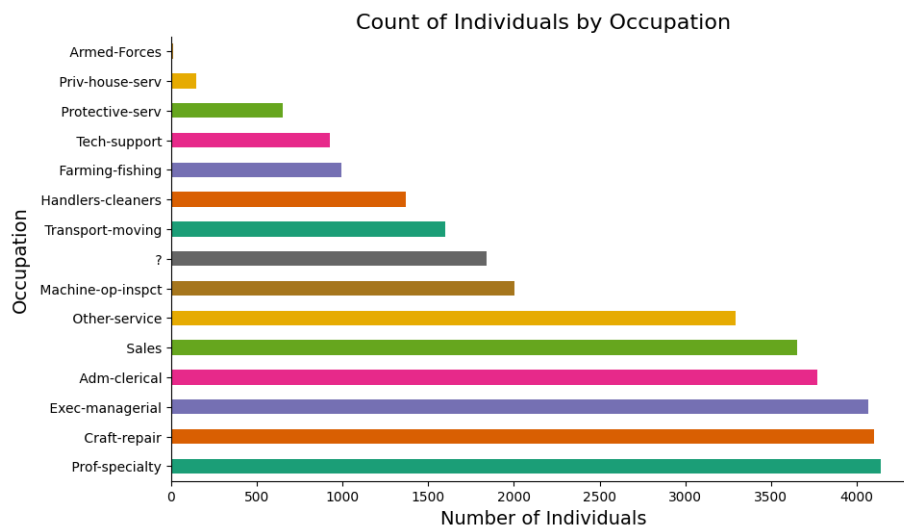
شکل ۷: توزیع ویژگی Salary بر اساس Education

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



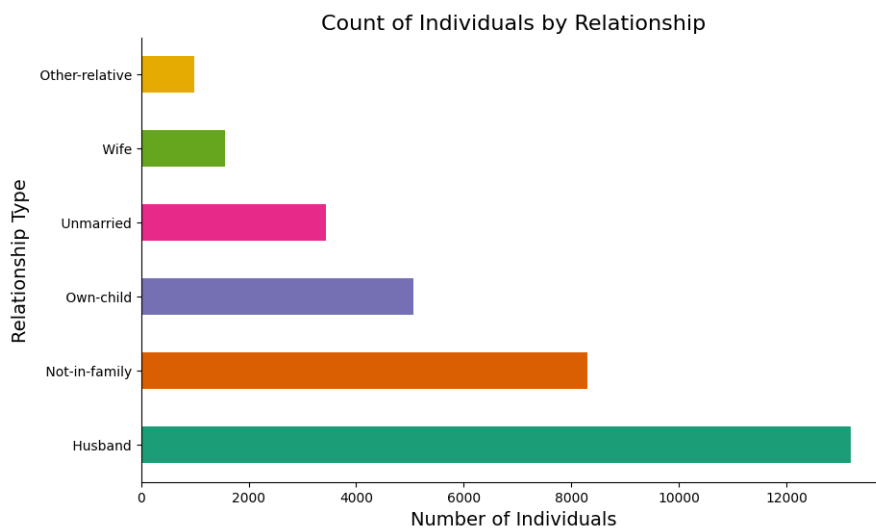
شکل ۸: توزیع ویژگی Hours Per Week

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)

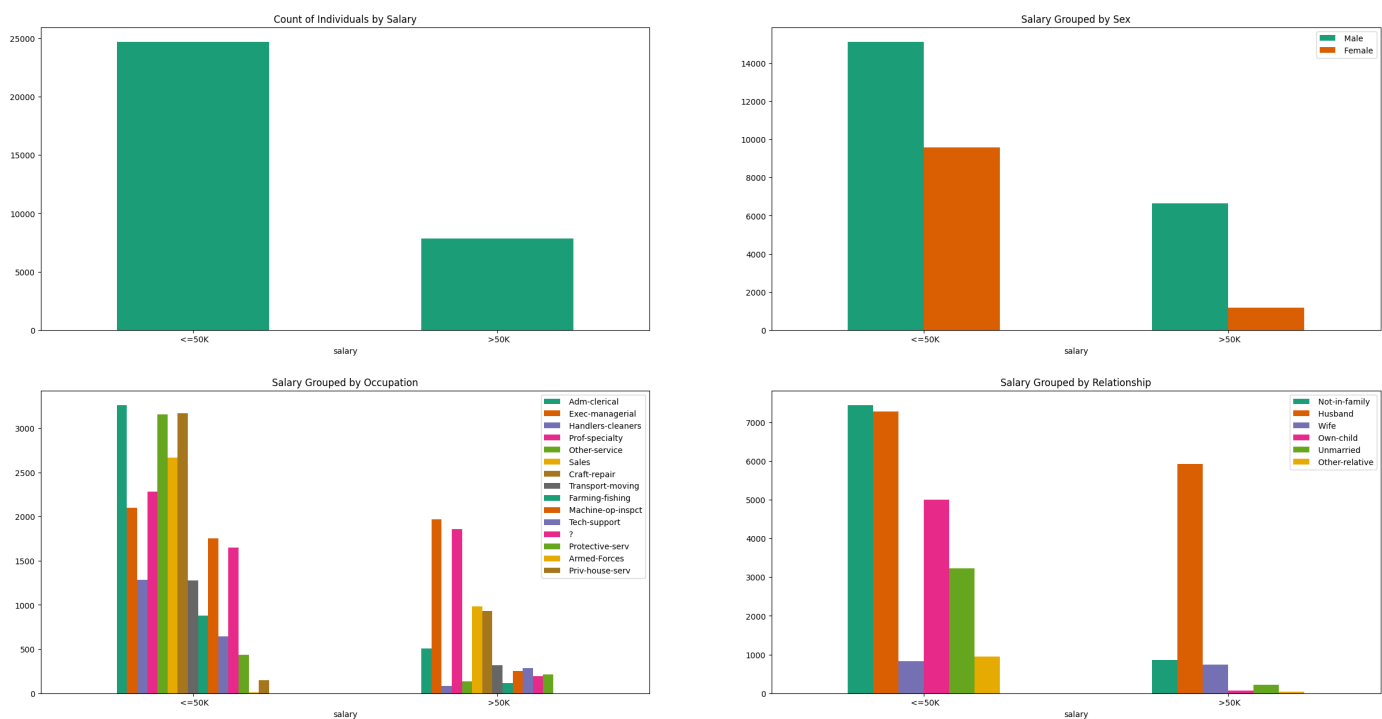


شکل ۱۰: توزیع ویژگی Relationship

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^۱ 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.

^۱Minor



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

۲ Feature Selection

سلام

۴ آموزش مدل

۵ نتایج

۶ نتیجه گیری

۷ مراجع