توضیحات پروژه

داده‌های موجود حاصل گزارش‌های موسسات تحصیلی مختلف به آژانس فدرال است که با توجه به حجم زیاد آن و تخصصی بودن بعضی از قسمت‌های آن و با توجه به توضیحاتی که در داکیونت داده‌ها(https://collegescorecard.ed.gov/data/documentation/) نوشته شده بود تصمیم بر آن شد که قسمتی از داده‌ها را انتخاب کرده و تحلیل داده را روی آن قسمت انجام دهیم. این داده‌ها که درباره موسسات تحصیلی از قبیل کالج‌ها است شامل اطلاعات زیر است:

شهر: شهری که موسسه مورد نظر یا شعب این موسسه در آن واقع شده‌اند.

گرایش مذهبی: برخی از موسسات فقط پذیرای قشر خاصی از دانش آموزان هستند، که برای این موسسات یک گرایش مذهبی خاص در نظر گرفته شده است که لیست این گرایشات به شرح زیر است:

HBCU: Historically Black Colleges and Universities -
PBI: Predominantly Black Institutions -
ANNHI: Alaska Native-/Native Hawaiian-serving Institutions -
TRIBAL: Tribal Colleges and Universities -
AANAPII: Asian American-/Native American-Pacific Islander-serving Institutions -
HSI: Hispanic-serving Institutions -
NANTI: Native American Non-Tribal Institutions -

که هر یک از گرایشات فوق به صورت یک ستون در فایل اکسل با مقادیر صفر و یک است. مقادیر فوق به شکل یک بولین است که کار با آن برای ما سخت بود و تصمیم گرفتیم که بجای داشتن چند ستون با مقادیر بولین یک ستون به صورت اسمی داشته باشیم که محتوای آن گرایش هر موسسه باشد ولی با توجه به این که موسساتی که دارای گرایش خاصی نیستند به ازای هر یک از گرایشات فوق مقدار صفر می‌گیرند، و بخش مربوط به گرایش آن‌ها خالی می‌ماند یک گرایش با نام NORMAL به وجود آوردیم تا موسسات بدون گرایش خاص با این نام مقداردهی شوند.

میزان درآمد خانواده‌های دانش آموزان: این بخش از داده شامل پنج ستون به شرح زیر است:

INC_PCT_LO = $0 - $30,000 -
INC_PCT_M1 = $30,001 - $48,000 -
INC_PCT_M2 = $48,001 - $75,000 -
INC_PCT_H1 = $75,001 - $110,000 -
- INC_PCT_H2 = $110,001 -

که هر یک از دسته‌بندی‌های فوق به صورت یک ستون در فایل اکسل است که شامل درصد دانش آموزانی است که در هر یک از دسته ها قرار می‌گیرد است، برای مثال بیست درصد از دانش آموزان یک موسسه در دسته اول یعنی کم درآمدها و سی درصد دانش آموزان در دسته آخر یعنی پر درآمد ها قرار می‌گیرند و سایر دانش آموزان هم در سایر دسته ها پخش می‌شوند پس با توجه به توضیحات فوق مجموع مقادیر این پنج ستون برای هر موسسه(هر سطر از داده) برابر یک یا صد درصد است.

میزان کار نیمه وقت: چند درصد از دانش آموزان هر موسسه در حال کار پاره وقت هستند.

طبقه بندی برنامه های آموزشی(The Classification of Instructional Programs (CIP)): این بخش شامل ۵۴ ستون است که هر ستون معرف یک طبقه خاص از یک طبقه‌بندی با نام CIP است که برای مثال CIP43 معرف علوم و فنون قضایی است.

اطلاعات تکمیلی از این بخش در لینک https://nces.ed.gov/ipeds/cipcode/Default.aspx?y=55 وجود دارد. اطلاعات هر ستون از این ۵۴ ستون شامل درصاد دانش آموزانی است که در این گرایش تحصیلی در سال ۲۰۱۶ فارغ التحصیل شده اند.

روند کار

ابتدا توسط یک اسکریپت قسمتی از داده‌ها که مورد نیاز است را برداشته و تمیز کردیم، که انجام این کار توسط اسکریپت reduce_data.py انجام شده است که به عنوان ورودی دو فایل، یکی فایل داده‌ها و دیگری فایل اطلاعات مورد نیاز، را می‌گیرد و یک فایل csv خروجی می‌دهد که شامل داده‌های مورد نظر است.

درخت تصمیم:

در این بخش از چهار مورد اول از موارد بالا استفاده کردیم و یک درخت تصمیم بوجود آوردیم که در آن با توجه به شهر و گرایش مذهبی و درآمد خانواده به میزان کار نیمه وقت یک موسسه رسیدیم.
ابتدا برای درست کردن درخت تصمیم مدل داده‌ها را عوض کردیم.

پنج ستون مربوط به درآمد خانواده را به یک ستون با دو مقدار low و high تقلیل دادیم. با توجه به اینکه در یک میانگین‌گیری بیشتر موسسات دارای میانگین بیش از پنجاه درصد خانواده کم درآمد هستند، ما مقادیر درصدهای متوسط یک، متوسط دو، زیاد یک و زیاد دو را با همدیگر جمع زده و به یک مقدار پر درآمد تبدیل کردیم و مقدار درصد کم را تبدیل به کم درآمد کردیم. به صورت خلاصه:
- پر درآمد = INC_PCT_H2 + INC_PCT_H1 + INC_PCT_M2 + INC_PCT_M1
- کم درآمد = INC_PCT_LO
با این کار کل داده پنج ستونی فوق تبدیل به یک داده اسمی شده که کار ما را برای درخت تصمیم زدن راحتتر می‌کند.
ستون‌های مربوط به گرایش مذهبی را نیز همانطور که در بالا اشاره شد به یک ستون با مقادیر اسمی تبدیل کردیم و بعد از آن ستون مربوط به میزان کار نیمه وقت را نیز با محاسبه سهک و میانگین و بدست آوردن یک معیار خوب برای تقسیم‌بندی به یک متغیر اسمی با سه مقدار low و medium و high تبدیل کردیم.

حال با استفاده از الگوریتم C4.5 یک درخت تصمیم برای داده‌ها بدست آوردیم که در نهایت برای نگه داری و به نمایش در آوردن درخت به دست آمده از قوانین IF ... THEN استفاده شده است که با اجرای فایل decision_tree.py روی داده‌های انتخابی این قوانین بر روی ترمینال به نمایش در می‌آید در زیر مثالی از این قوانین آورده شده است:

IF city = Chicago, income = low, affiliation = NORMAL, THEN low_working = 26.666666666666668 - medium_working = 60.0 high_working = 13.333333333333334

IF city = Kettering, income = low, THEN low_working = 0 medium_working = 100.0 high_working = -0

مثلا قانون شماره دوم بدین معناست که اگر شهر Kettering باشد و درآمد خانواده موسسه به طور میانگین پایین باشد با احتمال ۱۰۰ درصد نیمی از دانشجویان در حال کار پاره وقت هستند.

کشف اقلام مکرر

ستون‌های مربوط به درصد دانشجویان مربوط به یک رشته را با استفاده از اسکریپت reduce_data.py از داده‌های اصلی جدا می‌کنیم و هر سطر را به یک وکتور بولی، تبدیل می‌کنیم به صورتی که بفهمیم در هر دانشگاه، چه گروه رشته‌هایی وجود دارد.

در این بخش با استفاده از الگوریتم Apriori می‌توان قواعد تکراری را بدست آورد.

تابع Apriori با گرفتن، لیست رشته‌های هر دانشگاه، و حداقل ساپورت مورد نیاز، قوانین مکرر را تولید، و برای استفاده از قانون Apriori و هرس حالت‌های موجود از تابع item_with_min_sup استفاده می‌کند.

ورودی این تابع حداقل ساپورت، لیست رشته‌های هر دانشگاه و مجموعه‌ای از مجموعه آیتم هاست و خروجی این تابع، مجموعه‌ای از مجموعه‌هایی است که در لیست رشته‌ها به صورت مکرر دیده می‌شود.

با اجرا اسکریپت فوق، و با ورودی فایل و حداقل ساپورت، مجموعه‌های مکرر و ساپورت‌های آن در خروجی نمایش داده می‌شود.