«به نام خدا»

تکلیف دوم – سوال ششم – مرضیه علیدادی – 9631983

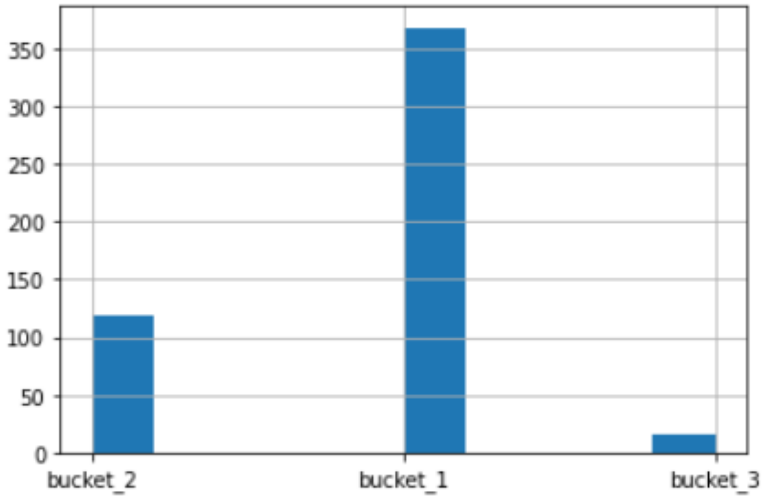
(کد های مربوط، در دو فرمت py. و ipynb. ضمیمه شده اند. – دیتاست اصلاح شده Diabetes_cleared.csv نیز ضمیمه شده است.)

.5

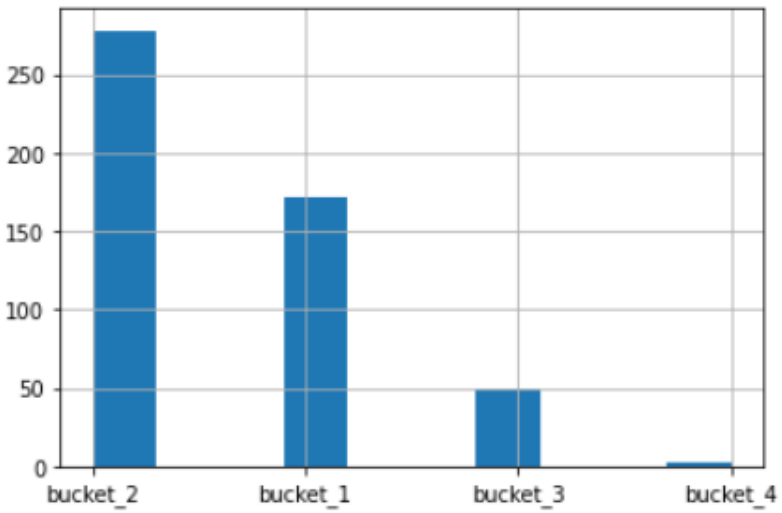
a) تحلیل ها در سوال 3 صورت گرفته است.

b)

- دسته بندی Pregnancies :



- دسته بندی Age :



c) روش دیگر علاوه بر cut و qcut، استفاده از jenkspy است. این روش مبتنی بر clustering است. عمل clustering، خودش یکی از روش های داده کاوی است؛ اما این جا از آن به عنوان پایه ای برای دسته بندی و binning نیز استفاده می‌شود. در این روش، طوری دسته ها تشکیل می شوند، که آن داده هایی که شبیه به هم هستند، یک سبد را تشکیل دهند. و پس از این هم اگر داده ی جدیدی به دیتاست وارد شود، بررسی می شود که به کدام یک از دسته ها شبیه تر است، تا در آن دسته قرار گیرد. این روش، نسبت به دو روش دیگر بهتر است و مشکلات کمتری دارد.

روش دیگر، روش مبتنی بر predictive value است. یعنی در این روش، به متغیر هدف و اینکه قصد ما از داده کاوی این دیتاست چیست، توجه می شود. فرضا می خواهیم نمره ی درس x را با استفاده از نمره ی درس y پیشبینی کنیم. در این صورت مرز هایی از داده های درس y را در نظر می گیریم که قبل و بعد از آن مرز ها، تاثیر و تفاوت قابل توجه ای روی نمره ی درس x مشاهده می شود. یعنی درواقع مرز ها را آن جاهایی قرار می دهیم که تغییر ایجاد می‌شود. و در نهایت، روی همین مرز ها، دسته بندی را انجام می دهیم.