Diabetes Dataset

این پروژه به چهار بخش تقسیم می‌شود: ۱- مدیریت داده‌های گم شده. ۲- حذف کردن داده‌های پرت. ۳- انتخاب ویژگی‌ها. ۴- ارائه و آموزش مدل

در ادامه، به شرح کارهای انجام گرفته در هر کدام از این سه بخش می‌پردازیم.

- **مدیریت داده‌های گم شده (missing values)**

در این قسمت بر اساس شرایط هر ویژگی برای داده‌های گم شده آن تصمیم‌گیری می‌شود.
برای ویژگی "race"، پر تکرارترین مقدار را جایگزین داده‌های گم شده می‌کنیم.
برای ویژگی "gender" ، به دلیل تعداد کم داده‌های گم شده، ردیف‌هایی که این ویژگی آن‌های گم شده است را حذف می‌کنیم.
برای ویژگی "weight"، ابتدا مقدار میانه بازه‌ی گفته شده را جایگزین آن بازه می‌کنیم. (به عنوان مثال، به جای تمام مقادیر [0-10] (10 عدد ۵ را قرار می‌دهیم.) سپس مقدار میانگین سنین گفته شده را جایگزین داده‌های گم شده می‌نماییم.
برای ویژگی‌های diag_1 ، diag_2 و diag_3 ، به دلیل تعداد کم داده‌های گم شده، ردیف‌های متناظر با آن‌ها را حذف می‌نماییم.
برای بقیه ویژگی‌هایی که داده‌های گم شده دارند، به دلیل تعداد بالای داده‌های گم شده و یا کم اهمیت بودن ویژگی‌ها در پیش‌بینی نهایی، آن‌ها را به همان صورت باقی می‌گذاریم.

- **حذف کردن داده‌های پرت (outlier elimination)**

در این بخش، ردیف‌هایی که احتمال وقوع مقدار یکی از ویژگی‌های آن‌ها از عددی معین (outlier threshold) کم‌تر باشد را حذف می‌کنیم.

- **انتخاب ویژگی‌ها (feature selection):**

در این مرحله بر اساس مقدار Mutual Information بین ویژگی مورد نظر و ویژگی هدف، ویژگی-هایی را انتخاب می‌کنیم. ابتدا مقدار Mutual Information تمام ویژگی‌ها را با برچسب داده‌ها (readmitted) حساب کرده و آن‌ها را از بزرگ به کوچک مرتب می‌کنیم. هر ویژگی‌ای که MI بیش-تری با برچسب داده‌ها داشته باشد، مناسب‌تر است. بنابراین از بین ویژگی‌ها، چند ویژگی‌ای که مقدار

MI بیش‌تری دارند را انتخاب می‌کنیم. ویژگی‌های نهایی انتخاب شده در فایل preprosecced_data.csv قابل مشاهده هستند. این فایل بعد از انجام تمام مراحل پیش پردازش داده‌ها که در بخش‌های قبلی به آن‌ها اشاره شد بدست آمده است.

- **ارائه و آموزش مدل**

ابتدا برای استخراج ویژگی‌های آموزنده‌تر از داده‌های عددی، یک شبکه عصبی سه لایه‌ای تعریف کرده و ویژگی‌های عددی را به آن ورودی می‌دهیم. این شبکه عصبی دارای ۷ ورودی، دو لایه مخفی با ۲۱ گره و لایه خروجی با سه گره است. این شبکه عصبی، داده‌ها را به سه گروه تقسیم‌بندی می‌کند. سپس بر روی هر کدام از آن سه گروه یک درخت تصمیم آموزش می‌دهیم. درخت تصمیم باینری خواهد بود و بر اساس معیار GINI انتخاب می‌کنیم که کدام ویژگی را گسترش دهیم. بنابر آزمایشات انجام شده، دقت این مدل برابر ۷۵ درصد است.