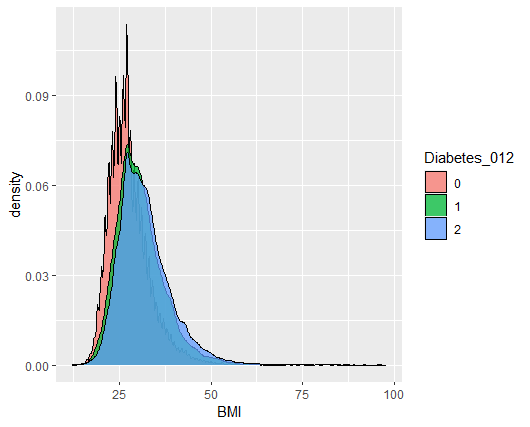
به نام او

سید شاهد رضوی زاده – 99104627

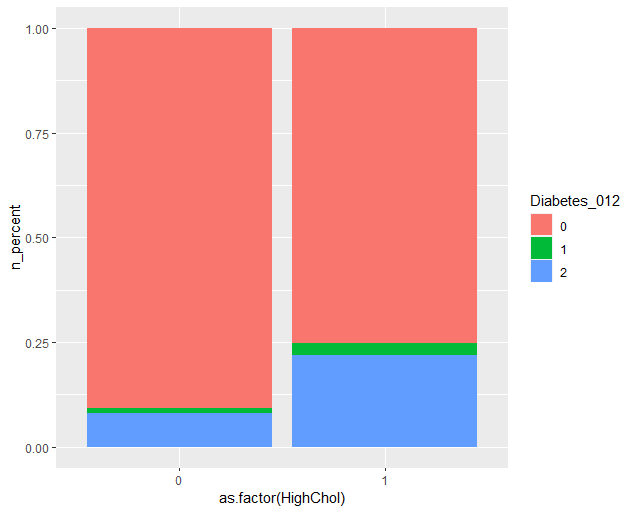
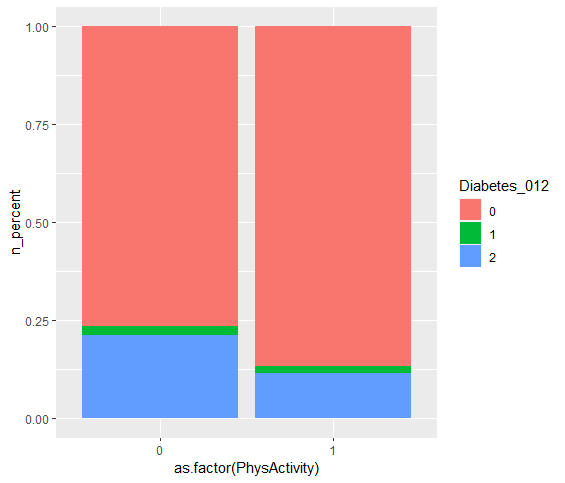
تمرین رگرسیون سری دوم

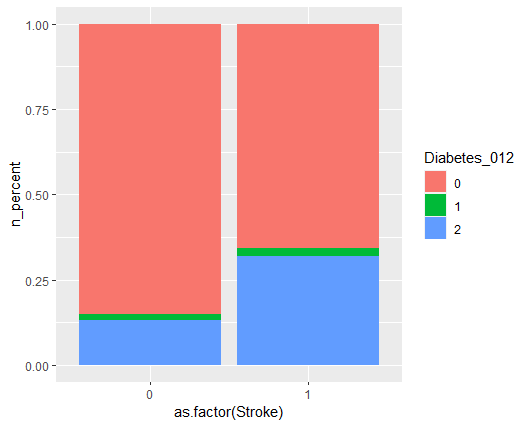
سوال 1) در این بخش نخست بررسی می‌کنیم که آیا تغییرات موجود در متغیرهای توضیحی بین داده‌هایمان با توجه به کلاس متغیر پاسخشان وجود دارد یا خیر، یعنی در واقع چقدر از تغییرات ناشی از کلاس‌ها، قابل برداشت از این متغیرهای توضیحی می‌باشد. اگر بررسی کنیم، متوجه می‌شویم که برای اکثر متغیرهای توضیحی که داریم، همچین اتفاقی می‌افتد و بین افرادی که دیابت ندارند یا دارند (یا پیش دیابت دارند) فرق محسوسی به طور میانگین در متغیرهای توضیحی‌شان (چه به صورت پیوسته و چه به صورت کلاس کلاس) وجود دارد.

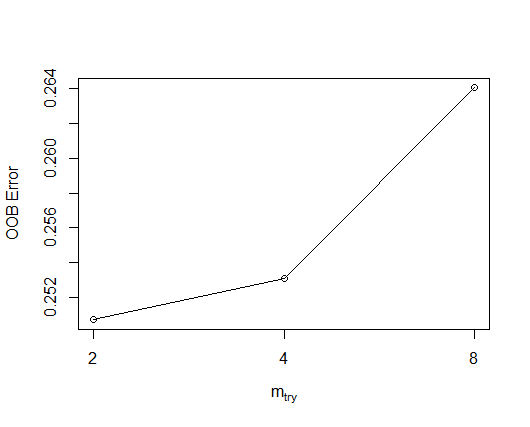
برای مثال شکل زیر تفاوت BMI بین افراد این سه کلاس را نشان می‌دهد.

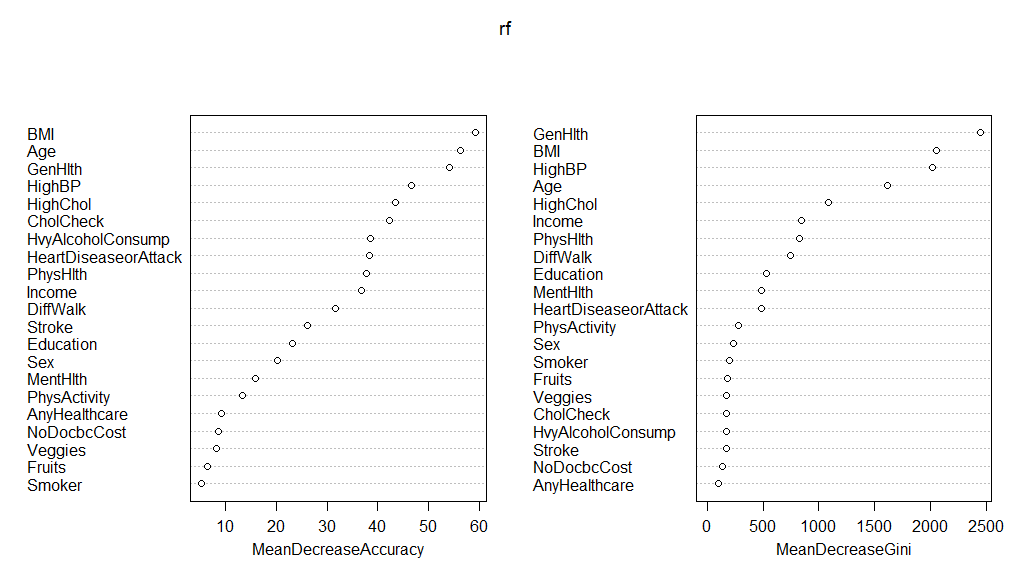


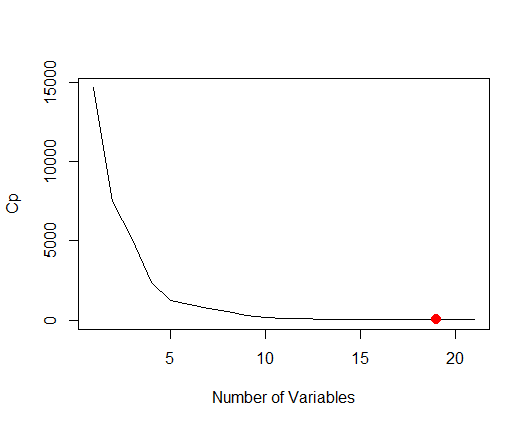
همچین عکس‌های دوم و سوم و چهارم نیز به ترتیب مربوط به کلسترول بالا داشتن، تحرک فیزیکی داشتن، و سکته داشتن در این سه کلاس (با دیابت و پیش دیابت و بدون دیابت) می‌باشد. و همان طور که می‌بینید مطابق انتظار نتایج حاصل شده است و تفاوتی معنادار بین کلاس‌های متغیر هدف وجود دارد. پس می‌توان گفت که با استفاده از این متغیرهای توضیحی می‌توان به نتایج خوبی رسید و از طرفی چون تعداد مشاهدات نیز بسیار بالاست و تعداد متغیرهای توضیحی موثر نیز کمتر از تعداد متغیرهای نظرسنجی می‌باشد که استفاده خواهد شد، پس مشکل overfitting را هم تجربه نخواهیم کرد.



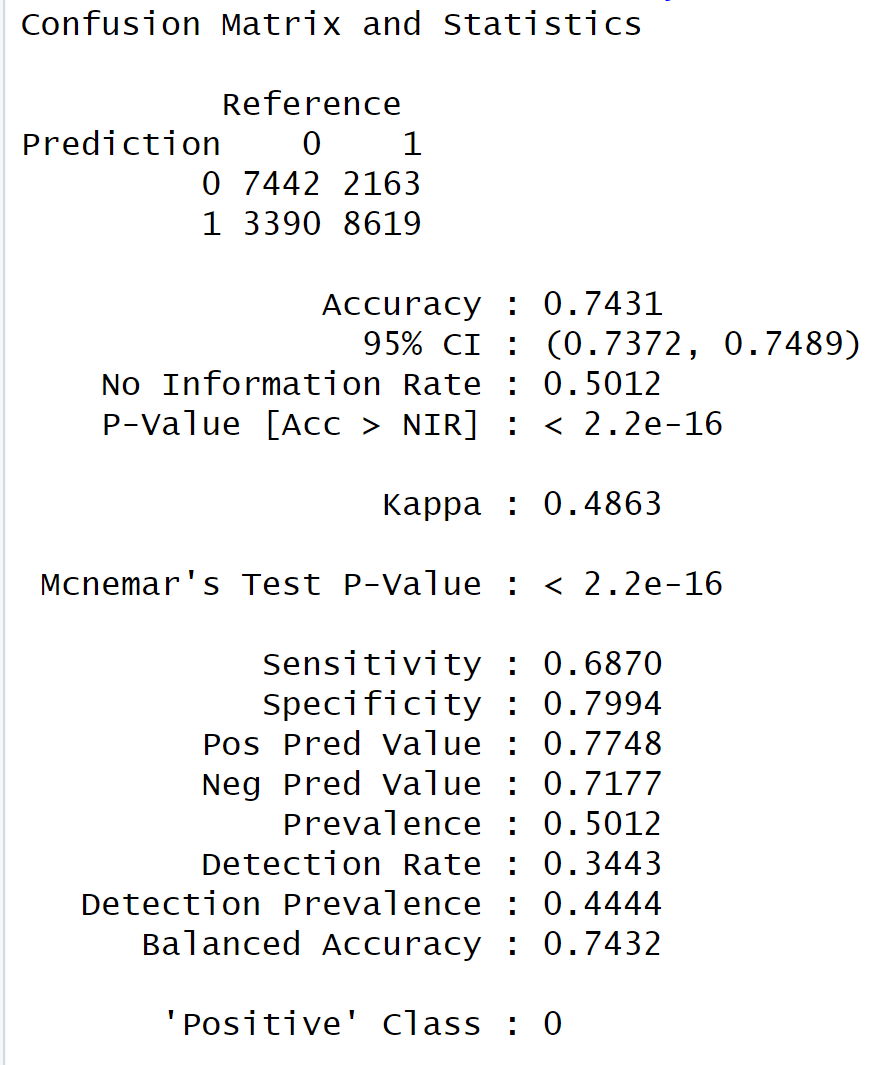


سوال 2) در این بخش ما از روش رندوم فارست که با آن توانستیم اهمیت متغیرها (variable importance) را بیابیم استفاده می‌کنیم، که در درس توضیح داده شد. در این جا نخست ما در متغیر best.m بهترین مقدار برای تعداد متغیر در هر اسپلیت (Split) که با آن درخت تصادفی ساخته می‌شود را پیدا می‌کنیم و سپس به کمک این مقدار بهینه رندوم فارست را پیدا می‌کنیم.

در نهایت نیز با کمک مدل درخت تصادفی، بهترین متغیرها را بر اساس این که با حذفشان چقدر از دقت مدل کاسته می‌شود یا ضریب جینی کاهش می‌یابد، مشخص می‌شود. 

سوال 3) در این بخش از روش و مفهوم best subset selection استفاده کردیم که نخست به ازای تعداد متغیر توضیحی ثابت هر بار بهترین مدل را پیدا کردیم و از آن جا که این مقادیر و جواب‌های بهینه نیاز به روش‌هایی مناسب برای مقایسه داشتند، برای مثال از Cp استفاده می‌کنیم. که در این جا به ازای 19 متغیر (نشان داده شده در شکل روبرو) این مقدار بهینه بوده است.

سوال 4) در این بخش نخست داده دیتاست را به دو بخش تست و ترین با نسبت 3 به 7 تقسیم کردم و سپس بر روی بخش آموزش، مدل رندوم فارست را ساختم (از این جهت که در بخش‌های قبلی ما حالت بهینه را پیدا کردیم دیگر نیازی به ولیدیشن برای پیدا کردن یک هایپرپارامتر بهینه در مدل جنگل تصادفی نیستیم) و نتیجه این مدل در داده های تست به این صورت شد که در شکال پایین قابل مشاهده می‌باشد. همان طور که می‌بینید، نتایج و درصد AUC و میزان sensitivity و دقت آن به این صورت می‌باشد:



سوال 5) در این بخش می‌توان گفت که وقتی مدل ساخته می‌شود، در نهایت به دنبال مدلی هستیم که با هزینه پردازشی سبک و کم در زمان کوتاه بتواند در سمت کلاینت یک ورودی را دریافت کند و خروجی و تخمین خود را به سرعت محاسبه کرده و خروجی دهد. برای همین منظور مشخصا مثلا روشی مثل k-nearest neighbor یک روش بسیار خواهد بود، چون مخصوصا در این داده‌ها که تعدادشان نیز بسیار بالا است باید فاصله هر نقطه دیتاست را با نقطه ورودی مورد نظر بسنجد که بسیار زمان گیر است و روشی مثل روش درخت تصمیم (برای مثال حنگل تصادفی) روش بسیار خیلی خوبی خواهد بود، چون تنها به دنبال این هستیم که به یک سری سوال، جواب بله/خیر داده شود و در نهایت پیش بینی صورت بگیرد. روشی مانند رگرسیون لاجیستیک هم از k-nearest neighbor اما کماکان با گرفتن ورودی، نیاز به قرار دادن آن در تابعی برای انجام یک محاسبه است که با توجه به تعداد متغیرهای توضیحی باز هم وقت خواهد گرفت و به اندازه درخت تصمیم پرسرعت نخواهد بود.