علی آراسته ۹۶۱۰۱۱۶۵

مقدمه‌ای بر یادگیری ماشین – تمرین عملی سری دوم

۱. آماده‌سازی داده‌ها

ابتدا با استفاده از دستور pandas.read_csv()، دیتاست داده شده را می‌خوانیم سپس قسمتی از آن شامل ۱۲ ویژگی و خروجی quality را جدا می‌کنیم.

برای آماده‌سازی دیتاست ابتدا باید داده‌های توصیفی را به داده‌های عددی تبدیل کنیم. برای این کار ابتدا یک دیکشنری به عنوان map ایجاد می‌کنیم و سپس با استفاده از دستور replace()، داده‌های توصیفی را مطابق متغیر map با داده‌های عددی جایگزین می‌کنیم.

در ادامه نوبت به داده‌های از دست رفته می‌رسد. برای انجام این کار از دو روش متفاوت استفاده می‌کنیم.

در روش اول با استفاده از دستور dropna()، تمام ردیف‌هایی را که حداقل یک داده از دست رفته دارند، حذف می‌کنیم. مشکل اساسی این روش از بین رفتن بخش مهمی از دیتاست است. به طور مشخص در این دیتاست، تنها ۳۳۵۸ نمونه از ۶۴۹۷ نمونه اولیه باقی می‌ماند.

در روش دوم با استفاده از دستورات fillna() و mean()، داده‌های از دست رفته را با میانگین داده‌های ستون مربوط به آن جایگزین می‌کنیم. شایان ذکر است در این روش مقدار میانگین برای داده‌هایی که در ابتدا توصیفی بوده اند، مقدار معتبری نیست؛ اما از آنجایی که برای یادگیری از الگوریتم‌های خطی استفاده می‌کنیم، این امر مشکلی ایجاد نخواهد کرد.

۲. طبقه‌بند Logistic regression

در این طبقه‌بند از به جای MSE از تابع Cross-Entropy به عنوان تابع loss استفاده می‌شود. علت این است که MSE به ازای تابع sigmoid به عنوان تابع طبقه‌بند، دیگر تابعی محدب نیست و در نتیجه برای پیدا کردن کمینه آن، نمی‌توان از روش‌های بهینه‌سازی محدب استفاده کرد.

شایان ذکر است تابع Cross-Entropy با استفاده از بیشینه کردن تابع Log-Likelihood بدست می‌آید و نسبت به بردار W محدب است؛ در نتیجه برای کمینه کردن آن، می‌توان از روش‌های بهینه‌سازی محدب استفاده کرد.

الگوریتم Logistic regression را مطابق با دستور کار طراحی می‌کنیم و سپس عملکرد آن را برای حالت‌های مختلف بررسی می‌کنیم.

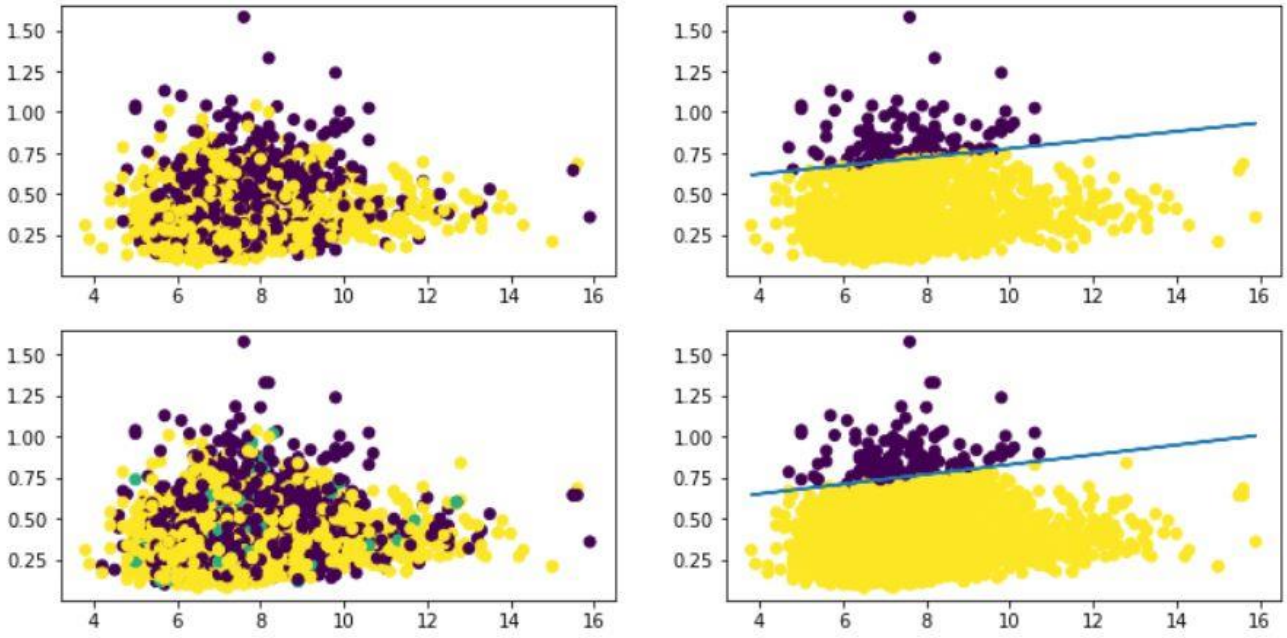
Learning rate = 0.01:

Number of iterations = 10000:

Without removing outlier data:

Model accuracy for first cleaning method = 64.74091721262656

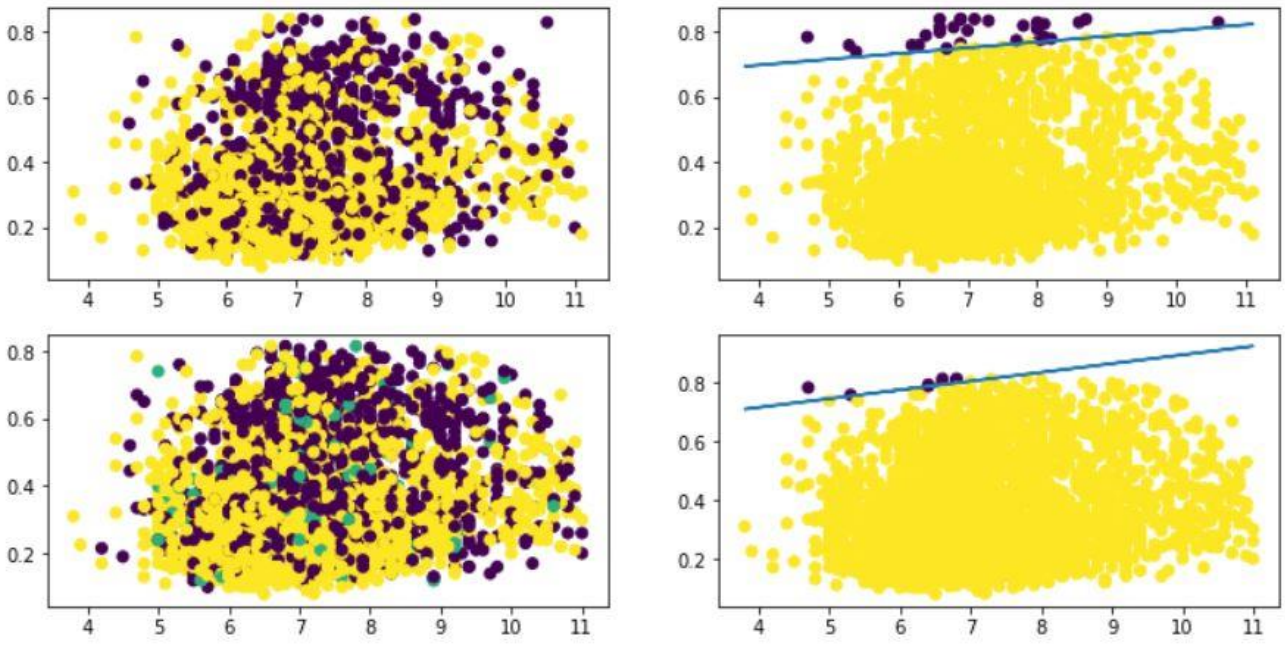
Model accuracy for second cleaning method = 60.9819916884716



With removing outlier data:

Model accuracy for first cleaning method = 63.74305126621371

Model accuracy for second cleaning method = 60.53597064922636

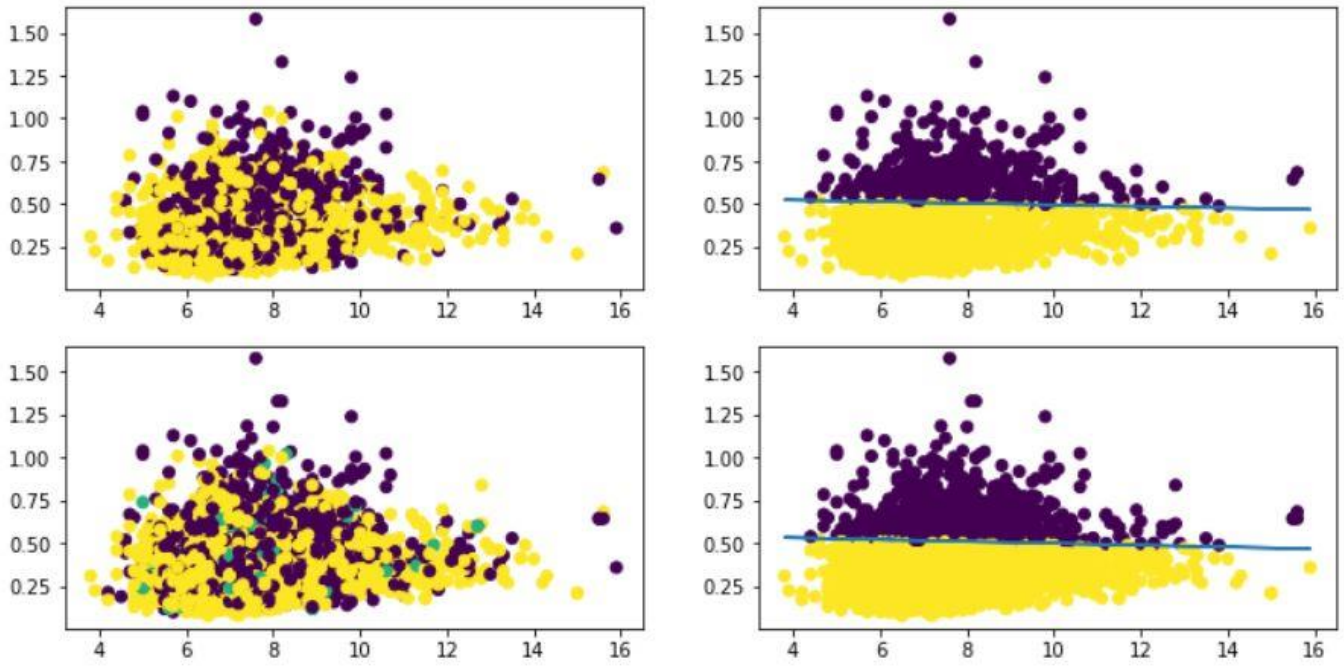


Number of iterations = 100000:

Without removing outlier data:

Model accuracy for first cleaning method = 66.31923764145324

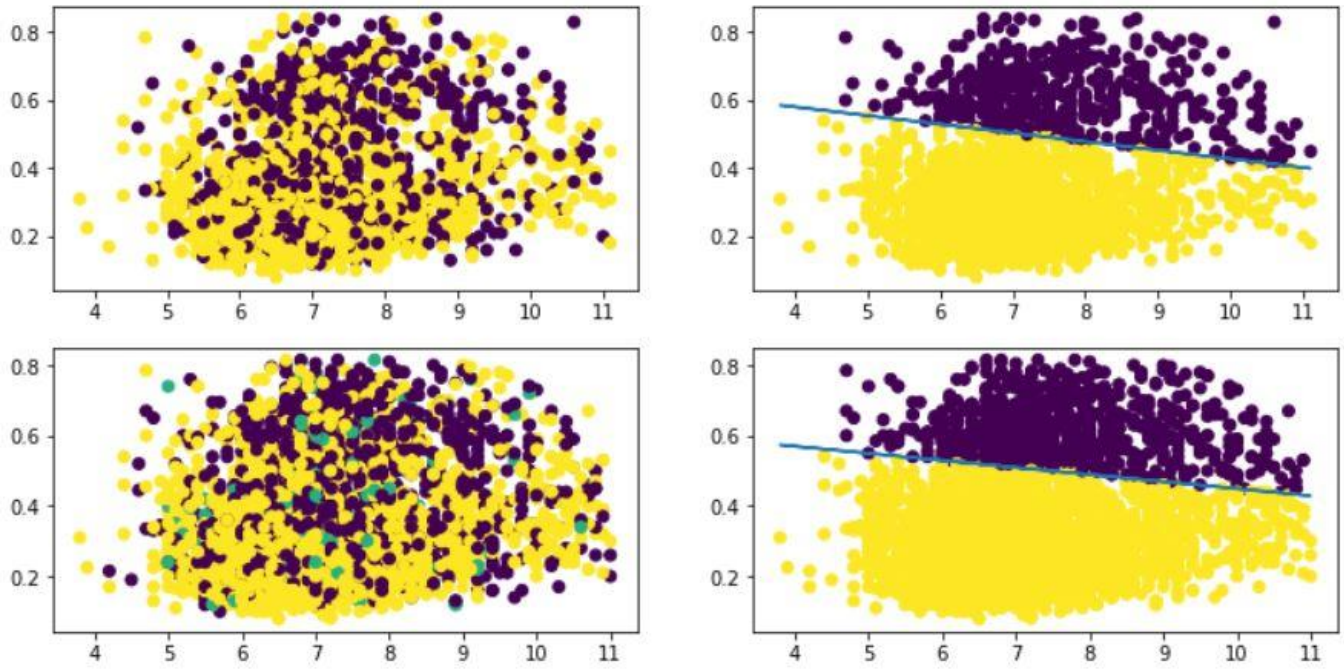
Model accuracy for second cleaning method = 62.92134831460674



With removing outlier data:

Model accuracy for first cleaning method = 66.49166151945646

Model accuracy for second cleaning method = 62.67347264316477



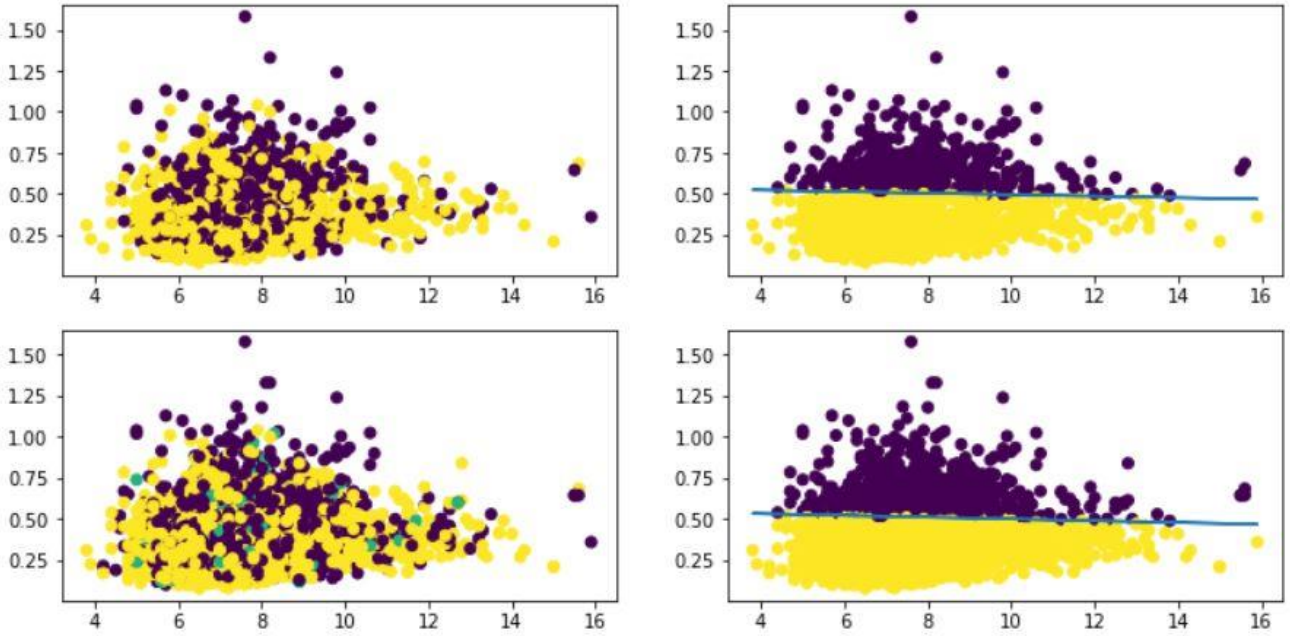
Learning rate = 0.1:

Number of iterations = 10000:

Without removing outlier data:

Model accuracy for first cleaning method = 66.31923764145324

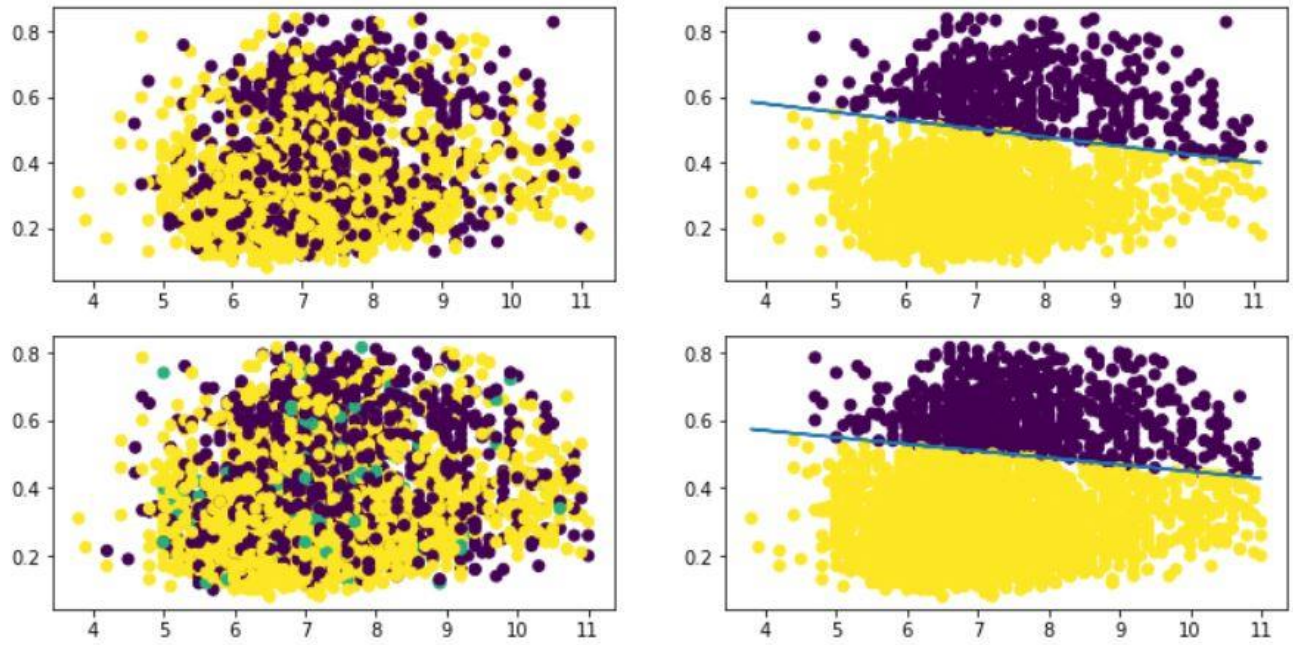
Model accuracy for second cleaning method = 62.92134831460674



With removing outlier data:

Model accuracy for first cleaning method = 66.49166151945646

Model accuracy for second cleaning method = 62.67347264316477

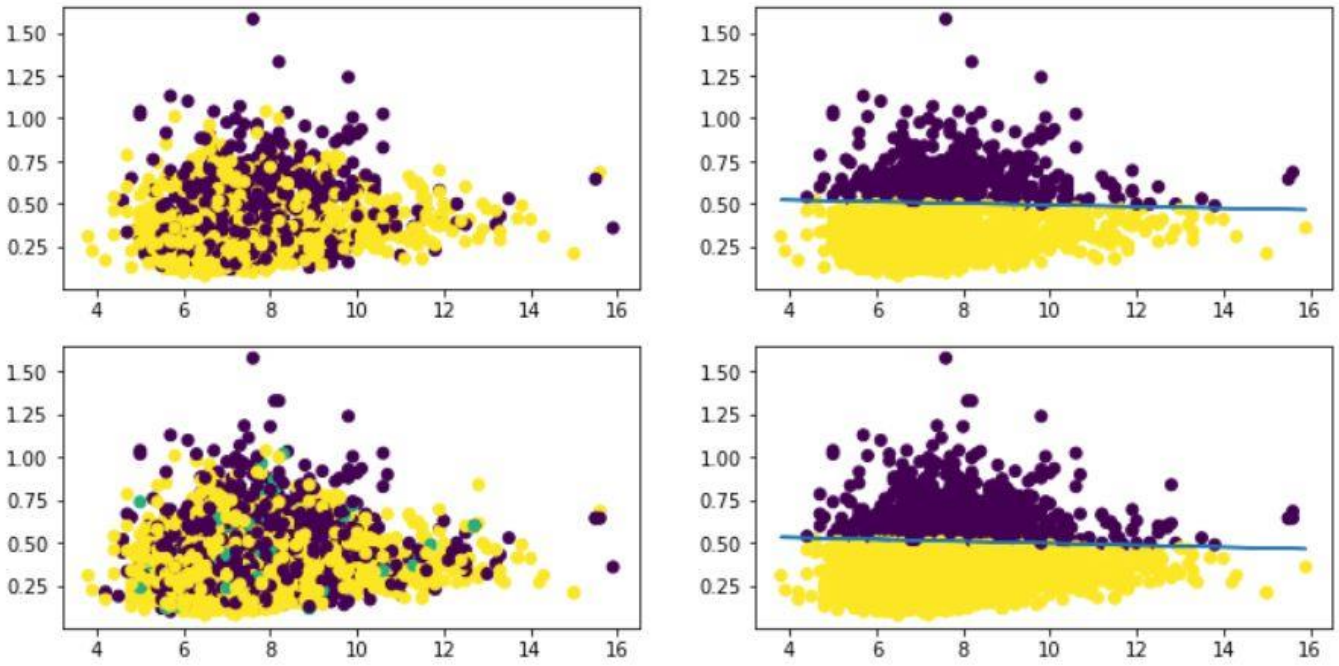


Number of iterations = 100000:

Without removing outlier data:

Model accuracy for first cleaning method = 66.25967837998809

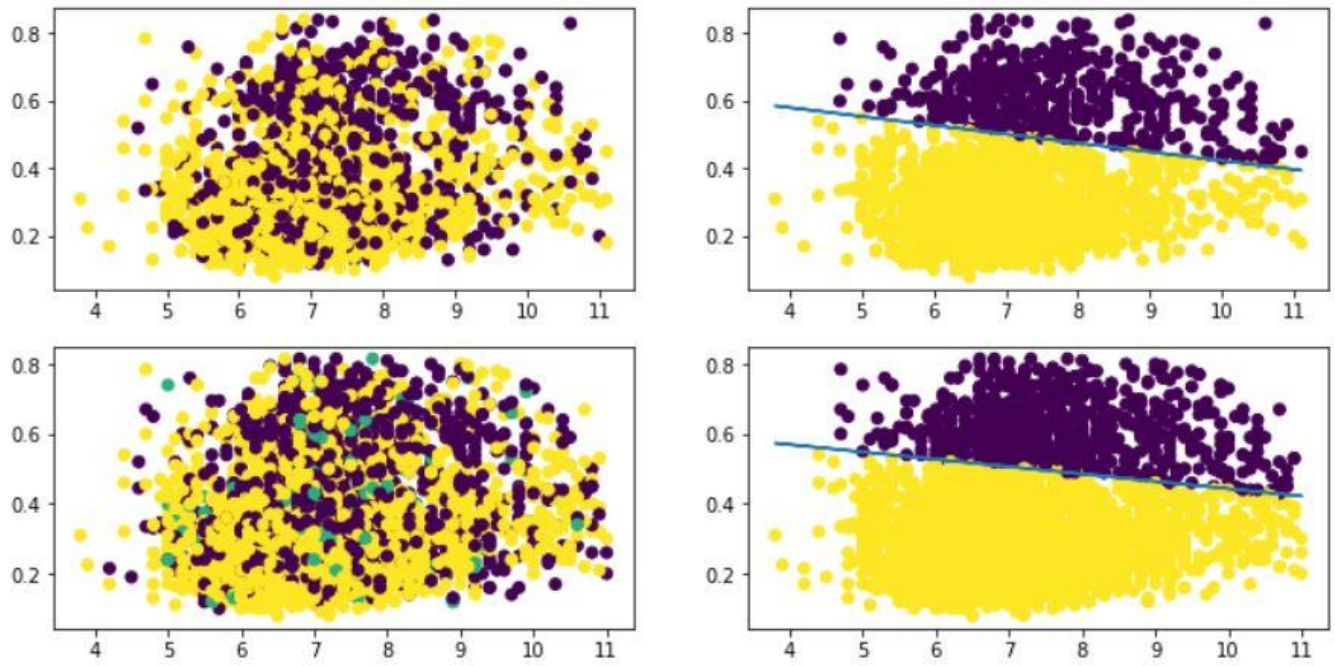
Model accuracy for second cleaning method = 62.92134831460674



With removing outlier data:

Model accuracy for first cleaning method = 66.52254478072884

Model accuracy for second cleaning method = 62.80108470250438



Learning rate = 1:

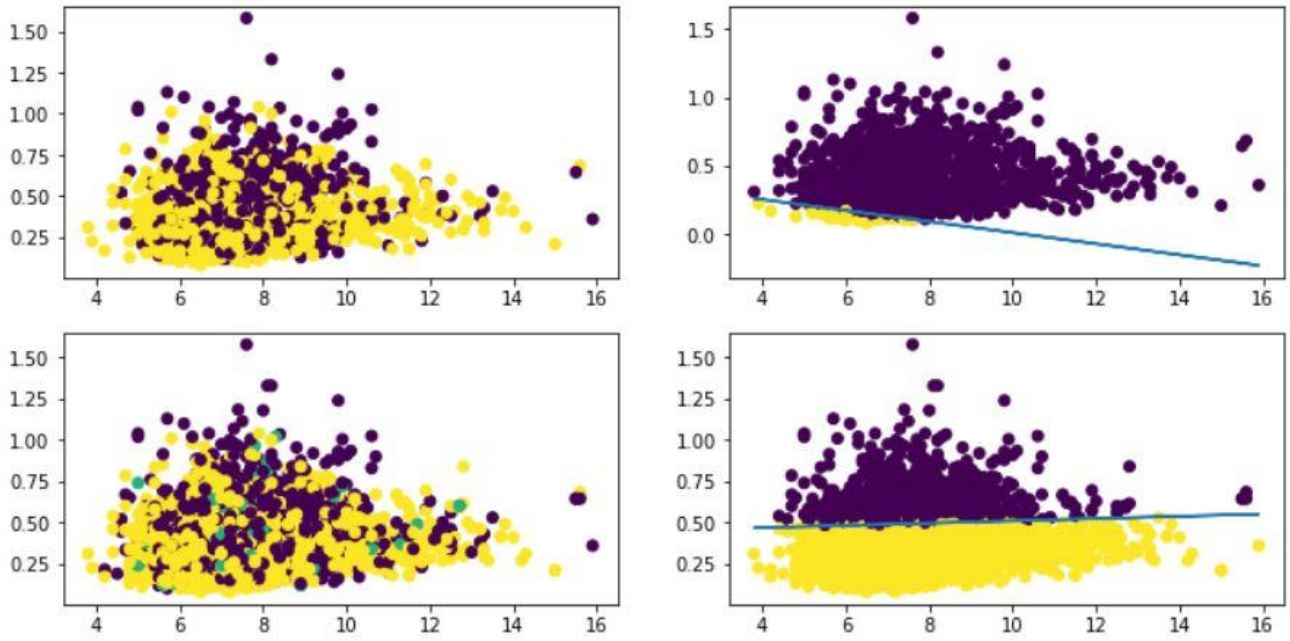
Number of iterations = 10000:

Without removing outlier data:

Model accuracy for first cleaning method = 39.24955330553901

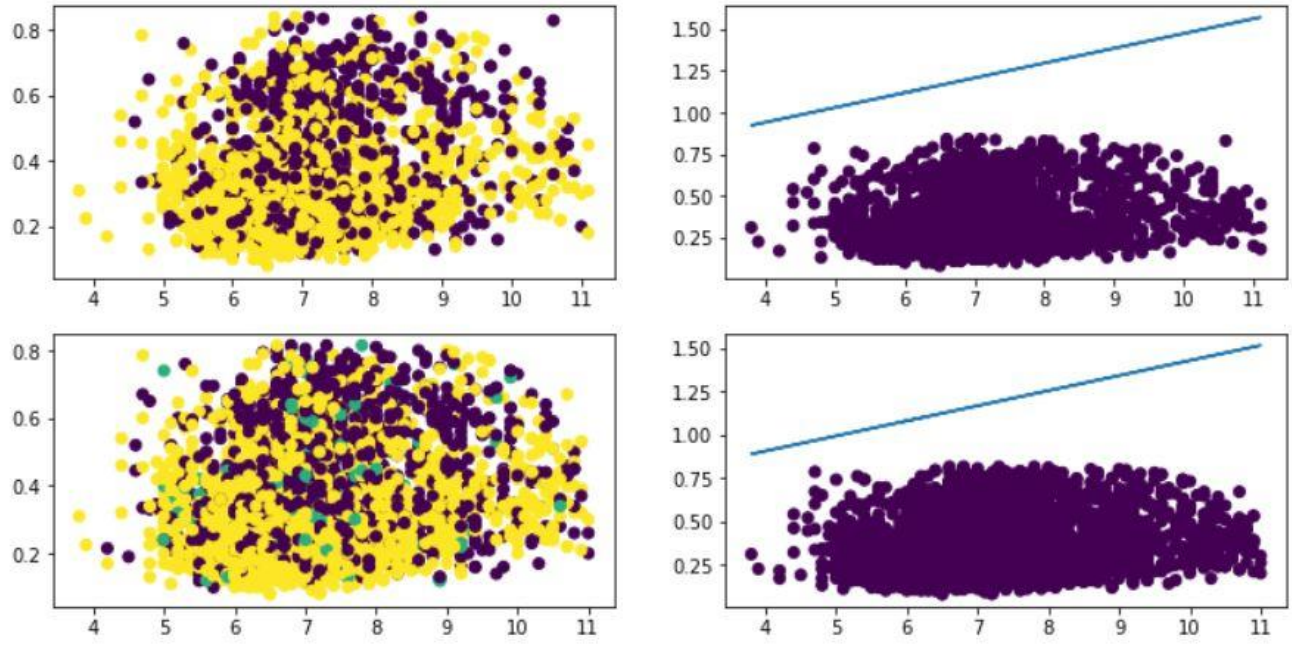
Model accuracy for second cleaning method = 62.70586424503617

With removing outlier data:



Model accuracy for first cleaning method = 63.650401482396546

Model accuracy for second cleaning method = 60.5200191418089

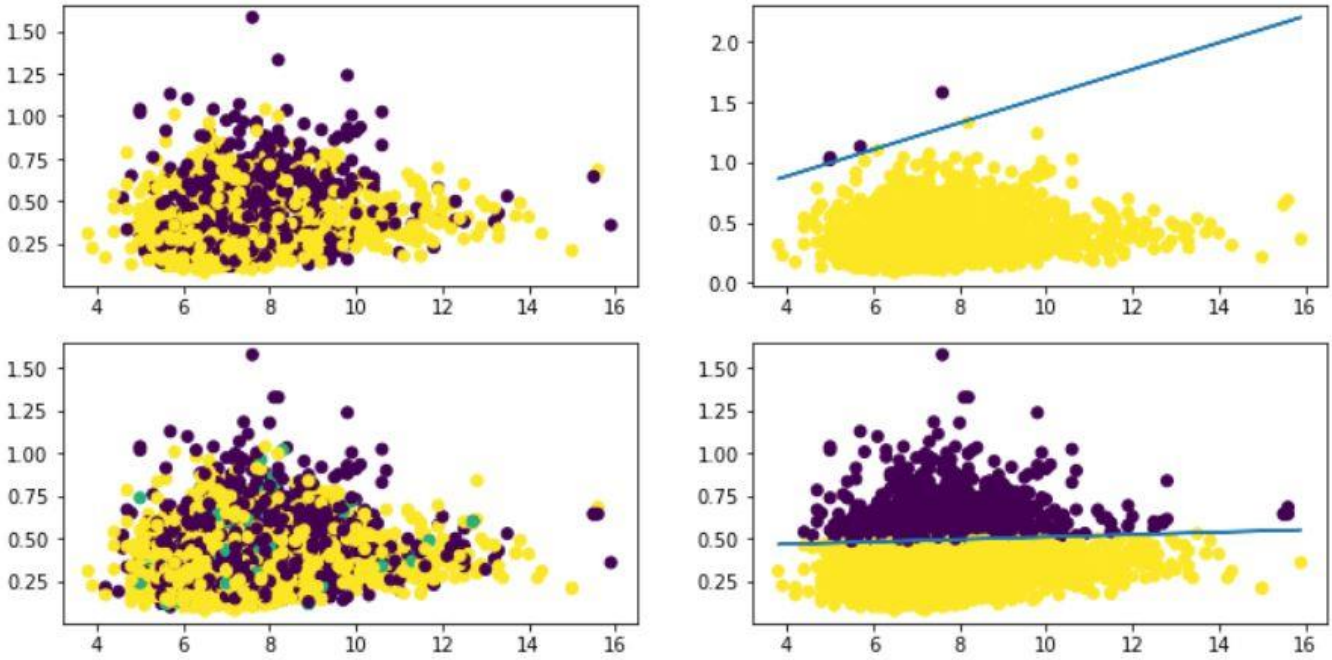


Number of iterations = 100000:

Without removing outlier data:

Model accuracy for first cleaning method = 63.43061346039309

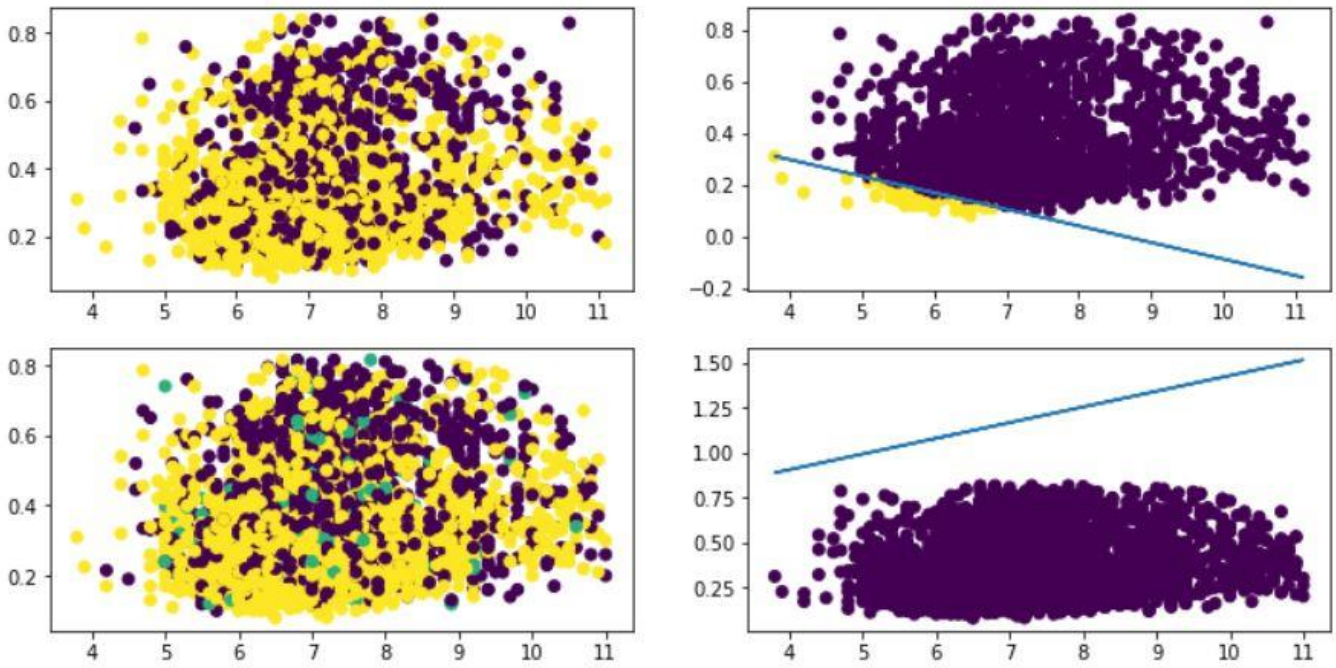
Model accuracy for second cleaning method = 62.70586424503617



With removing outlier data:

Model accuracy for first cleaning method = 38.20259419394688

Model accuracy for second cleaning method = 60.5200191418089



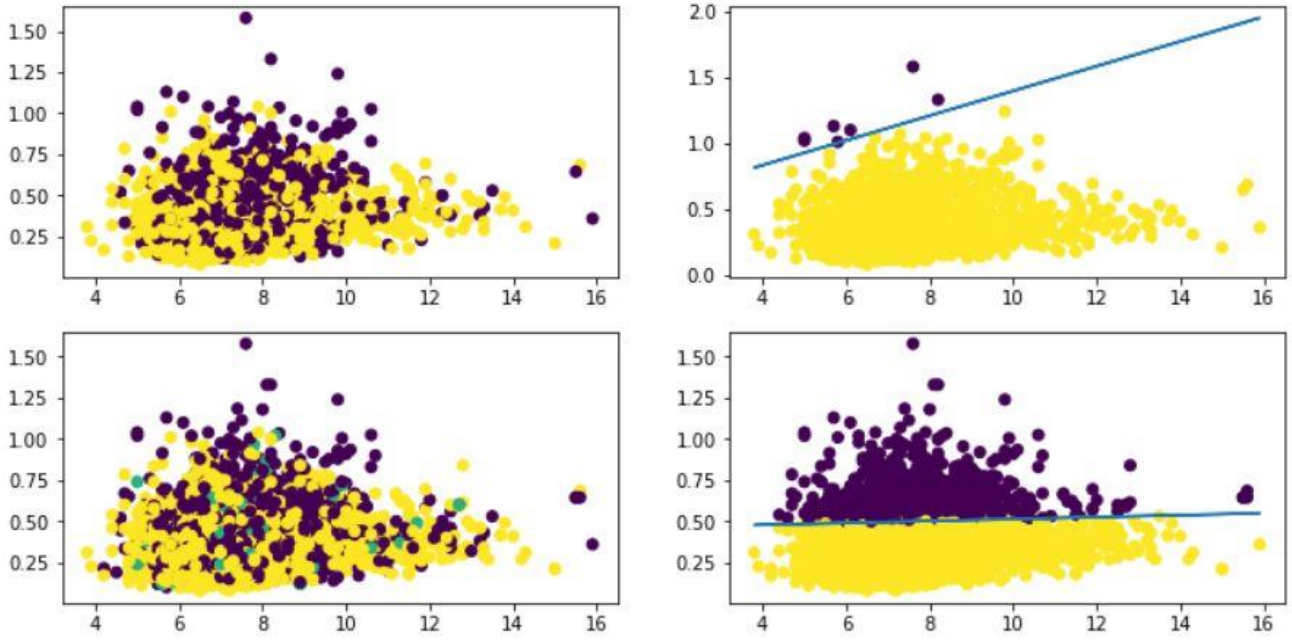
Learning rate = 10:

Number of iterations = 10000:

Without removing outlier data:

Model accuracy for first cleaning method = 63.46039309112567

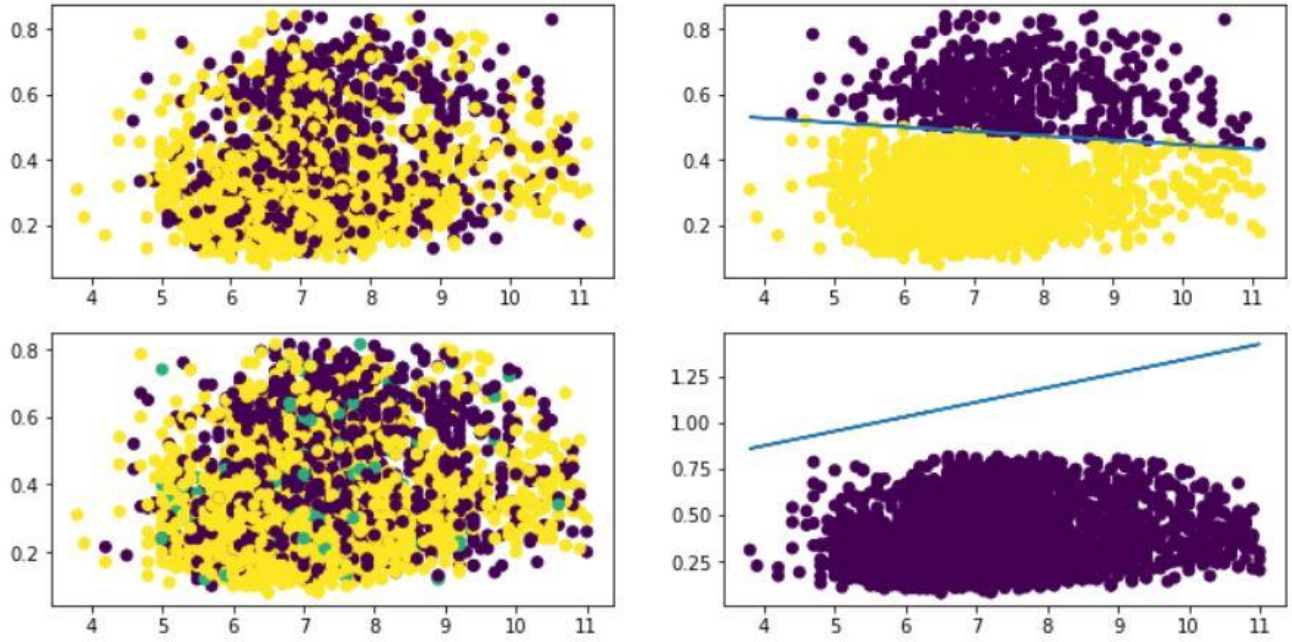
Model accuracy for second cleaning method = 62.70586424503617



With removing outlier data:

Model accuracy for first cleaning method = 66.46077825818406

Model accuracy for second cleaning method = 60.5200191418089

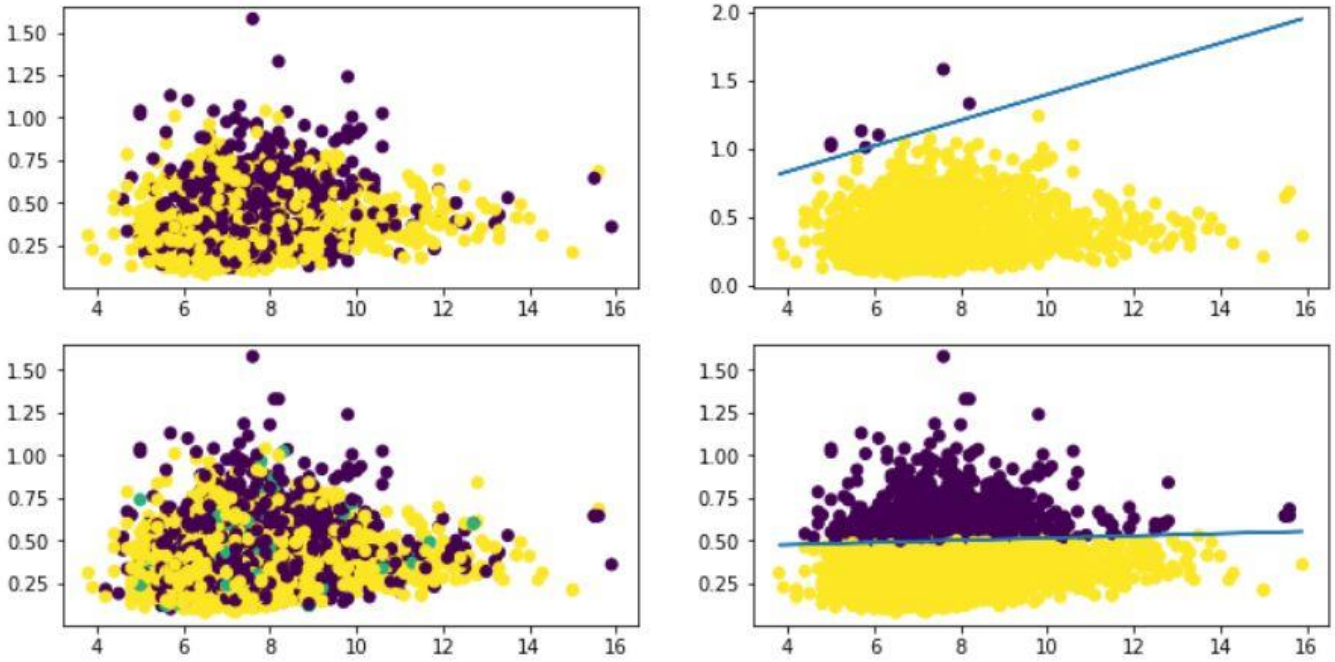


Number of iterations = 100000:

Without removing outlier data:

Model accuracy for first cleaning method = 63.46039309112567

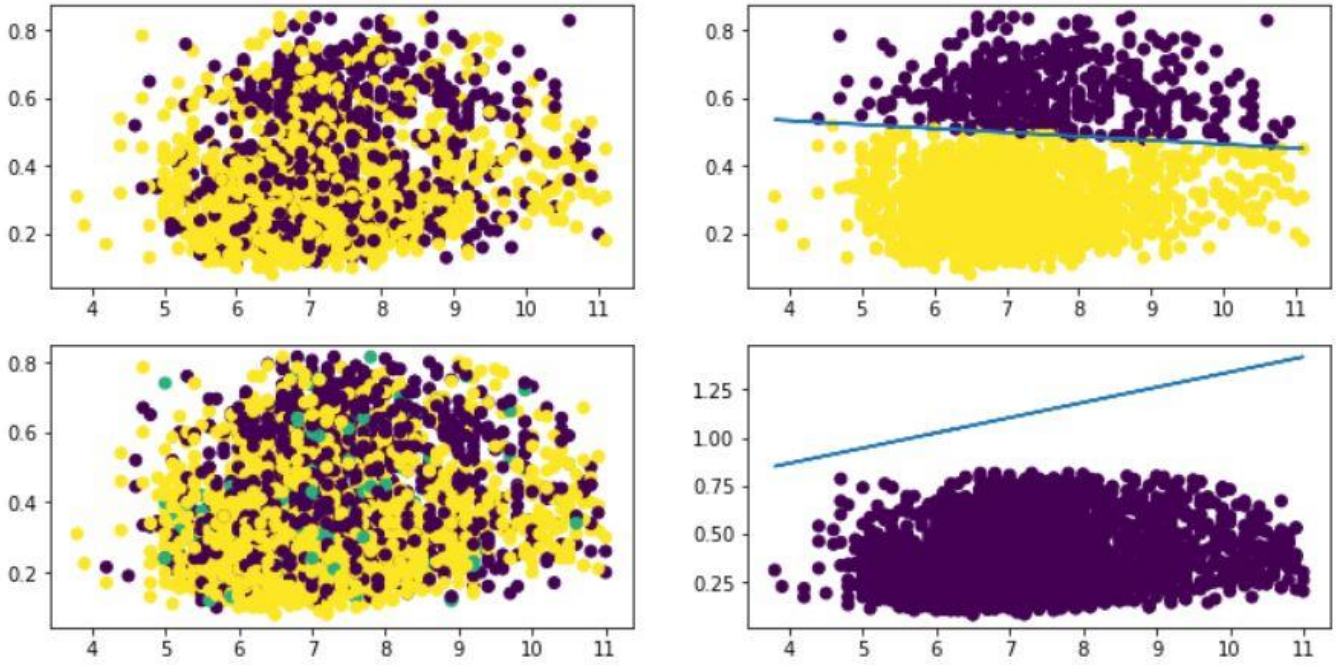
Model accuracy for second cleaning method = 62.690472525781125



With removing outlier data:

Model accuracy for first cleaning method = 66.15194564546016

Model accuracy for second cleaning method = 60.5200191418089



با توجه به نمودارهای بدست آمده، به طور کلی می‌توان به موارد زیر اشاره کرد:

برای نرخ‌های خیلی پایین، ۰/۰۱، سرعت همگرایی به نقطه کمینه کم است و در نتیجه برای تکرارهای کم، ۱۰۰۰۰، خطای مدل زیاد خواهد بود؛ اما با افزایش تکرار، ۱۰۰۰۰۰، نقطه کمینه به درستی شناسایی می‌شود.

برای نرخ‌های بالا، ۱ و ۱۰، رفتار مدل بسته به تعداد تکرار، تقریبا سینوسی است. علت این امر آن است که دیتاست کاملا درهم است و به هیچ وجه قابل جداسازی با یک خط نیست؛ در نتیجه اندازه گرادیان تابع loss بزرگ است و ممکن است برای نرخ‌های بالا حول نقطه کمینه به شدت نوسان کند.

به طور کلی با حذف داده‌های پرت عملکرد مدل در روش اول بهبود می‌یابد و در روش دوم تضعیف می‌شود.

با توجه به مشاهدات نرخ ۰/۰۱ و تکرار ۱۰۰۰۰۰ برای بدست آوردن یک خط خوب مناسب است.

اگر برای بدست آوردن خط مناسب از تمام ۱۲ ویژگی استفاده کنیم، به دقت‌های زیر می‌رسیم:

Model accuracy for first cleaning method = 64.33811802232854

Model accuracy for second cleaning method = 65.74227831285287

به طور مشخص این کار نیز وضعیت را آنچنان بهتر نمی‌کند.

۳. تمرین Linear regression

آ) موارد خواسته شده در فایل کد آورده شده است.

ب) موارد خواسته شده در فایل کد آورده شده است.

ج) با بررسی نرخ‌های مختلف، بهترین حالت ممکن با مقادیر زیر بدست می‌آید:

W = [2.08237101 0.45511227 2.90067044 -0.07512324]

MSE = 0.095 Number of iterations = 3094 Eta = 0.018

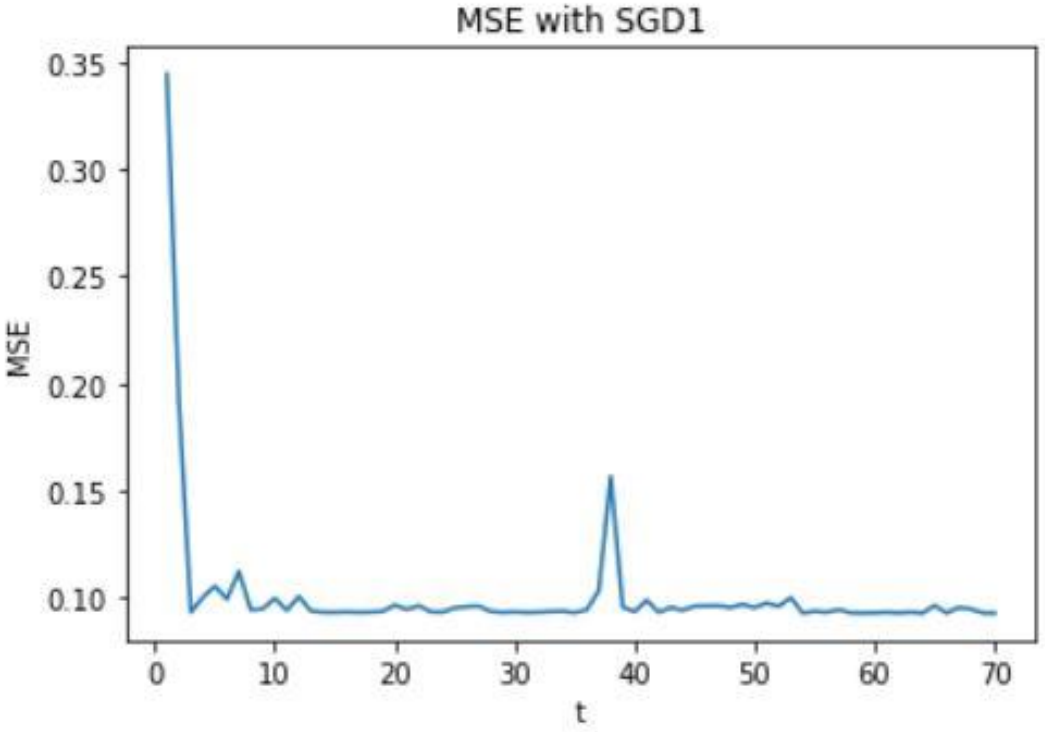
د) از آنجایی که r مقدارهای ۱ تا m را با احتمال برابر قبول می‌کند، احتمال اینکه r برابر i بین ۱ تا m باشد، برابر $\frac{1}{m}$ است. بنابراین داریم:

$$E\left[V_t|X^{(t)}\right] = \frac{1}{m}\sum_{i=1}^{m}\nabla l_i\left((W,b)^{(t)}\right) = \nabla L\left(X^{(t)}\right)$$

ه) با بررسی نرخ‌های مختلف، بهترین حالت ممکن با مقادیر زیر بدست می‌آید:

W = [2.07625352 0.47774205 2.96044576 -0.05663001]

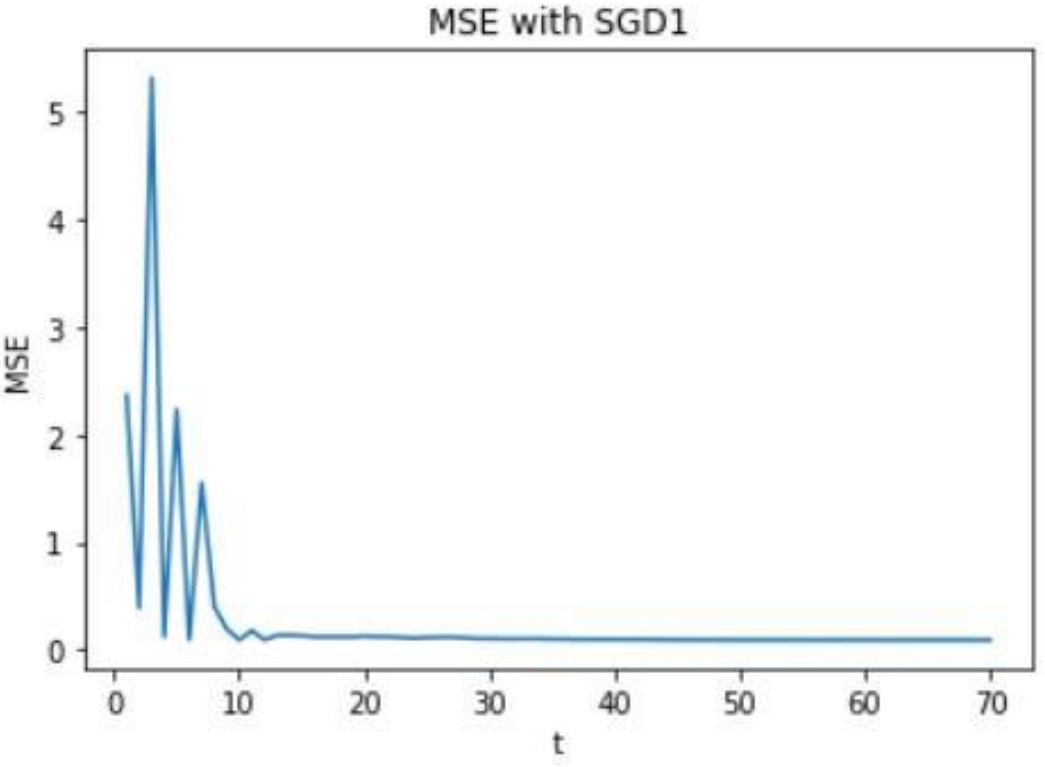
MSE = 0.09299078635466838 Learning rate (η_a) = 0.018



و) با بررسی نرخ‌های مختلف، بهترین حالت ممکن با مقادیر زیر بدست می‌آید:

W = [2.06839727e+00 6.08551849e-01 2.83627162e+00 2.88536508e-04]

MSE = 0.09145476204301962 Learning rate (η_b) = 0.055

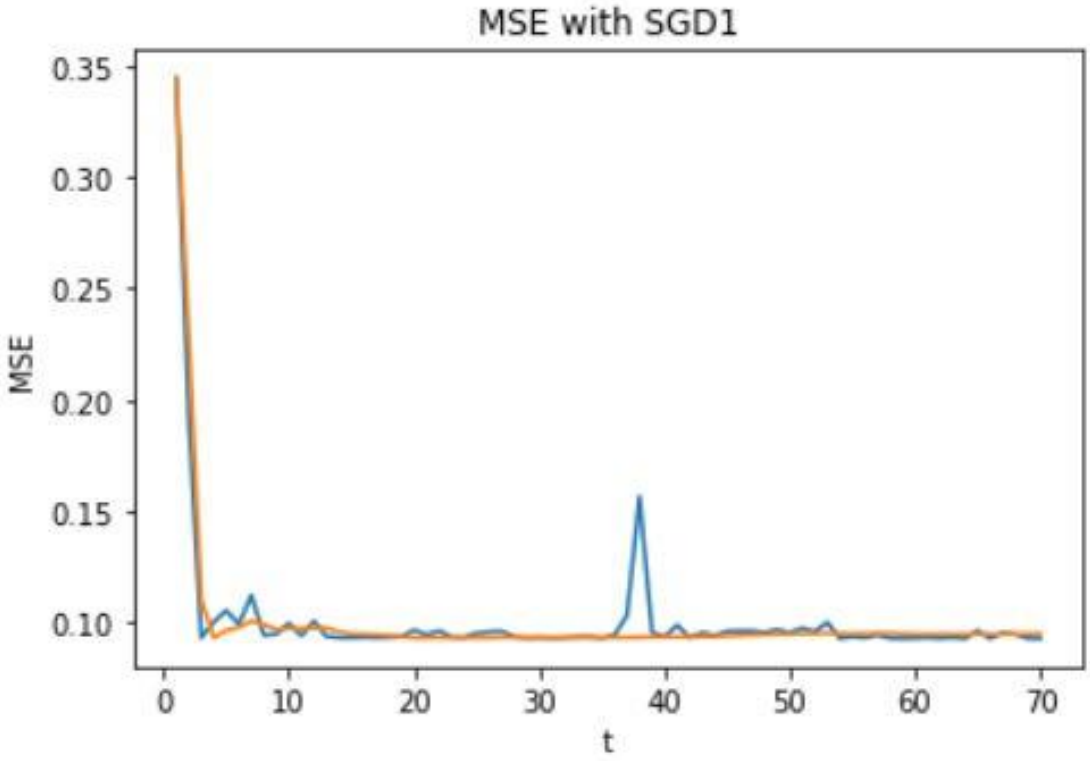


از آنجایی که نرخ یادگیری با افزایش t کاهش می‌یابد، داده‌های جدید تاثیر کمتری روی وزن نهایی دارند.

ز) با بررسی نرخ‌های مختلف، بزرگترین نرخ یادگیری که به ازای آن هر دو روش همگرا می‌شوند، برابر است با:

Maximum learning rate (η_*) = 0.018

برای این نرخ داریم:



تغییرات و نوسان روش دوم نسبت به روش اول کمتر است. این از آن جهت است در روش دوم نرخ یادگیری با افزایش t کاهش می‌یابد.