مراحل انجام عملیات:

بعد از آنکه دیتاست را در نرم افزار WEKA بارگذاری کردیم گزارش اولیه ای از وضعیت دیتاست بدست میدهد که این دیتاست شامل ۱۰۰۰ رکورد است و هر رکورد از ۲۱ ویژگی متشکل شده است. در وکا با انتخاب هر کدام از ویژگیها میتوان توزیع داده ای آن متغییرنسبت به متغییر هدف یعنی ویژگی class و همچنین توزیع آماری داده های آن متغییردر کل دیتاست را مشاهده کرد که نتایج این عملیات در قسمت نتایج آورده شده است.برای به نهایت رساندن پروژه در کنار نرم افزار وکا که برای راحتی مراحل تجزیه تحلیل اکتشافی داده ها بود از زبان برنامه نویسی R برای مراحل مدل سازی داده ها استفاده کردیم زیرا دارای محدوده ی گسترده ای از تکنیک های آماری از جمله مدل سازی خطی و غیر خطی،رده بندی ،خوشه بندی و ... میباشد.

اگر مجموعه های trainset و testset از قبل درست شده باشد فقط آنها را read میکنیم ولی اگر درست نشده باشد با استفاده از تابع split در برنامه ای که نوشته ایم به نسبت 7 به 3 جدا میکنیم و در فایل های جداگانه با فرمت arff ذخیره میکنیم.

ما تصمیم گرفتیم که دیتاست را با استفاده از درخت تصمیم، الگوریتم بیزین و قوانین انجمنی مدل سازی کنیم.و نتایج بر اساس جدول زیر آمده است.

Measures	Navive Bayesian	Decision tree
Accuracy	0.76	0.72
95%CI	(0.70,0.80)	(0.66,0.77)
Sensitivity	0.53	0.45
Specificity	0.85	0.84
P-Value	0.013	0.173

درخت تصمیم که در شکل زیر مشاهده میشود توسط R رسم شده که در هنگام اجرای برنامه در فایل پروژه ایجاد میشود.

Classification Tree