

# گزارش کار بخش پیاده سازی تمرین دوم درس یادگیری ماشین

محمد لشکری

۶ دی ۱۴۰۰

مجموعه متغیرهای مستقل و وابسته به صورت زیر انتخاب شده‌اند:

$X = \{Pclass, Sex, Age, SibSp, Parch, Fare, Embarked\}$

$Y = \{Survived\}$

همانطور که مشاهده می‌شود بعضی از ویژگی‌ها مانند شناسه مسافر و بلیت و کابین به دلیل یکتا بودن مقادیر حذف شده‌اند.

## ۱ پیش پردازش داده‌ها

با استفاده از KNNImputer با در نظر گرفتن ۲۰ همسایه مقادیر گم‌شده در ویژگی‌های سن و کرایه بلیت جایگزین شده‌اند. ویژگی‌های جنسیت و Embarked به نوع عددی تغییر پیدا کردند. نسبت دادن بازه به داده در ویژگی‌های سن و کرایه بلیت باعث حذف اطلاعات از دادگان می‌شود اما چون دامنه این متغیرها گسترده است با استفاده از مقیاس‌کننده استاندارد آنها را مقیاس کرده‌ایم تا عددی در بازه  $[-3, 3]$  اختیار کنند. لازم به ذکر است منفی شدن اعداد در نتیجه تأثیر ندارد.

## ۲ مدل‌سازی

قبل از فرایند مدل‌سازی، دادگان آموزش به دو بخش آموزش و تست با نسبت ۰/۲ تقسیم شدند. سه الگوریتم زیر روی دادگان آموزش دیده‌اند:

- Support Vector Machine (SVM)
- Logistic Regression (LR)

- Gaussian Naïve bayes (GNB)

که در آن کرنل SVM خطی است. نتایج زیر با استفاده از cross validation با تعداد ۵ فلد به دست آمده است:

	precision(macro ave.)	precision of class 0(Test set)	precision of class 1(Test set)
SVM	0.78	0.84	0.70
LR	0.79	0.87	0.74
GNB	0.78	0.78	0.77

	recall(macro ave.)	recall f class 0(Test set)	recall of class 1(Test set)
SVM	0.76	0.80	0.75
LR	0.77	0.83	0.80
GNB	0.78	0.83	0.71

	f1-score(macro ave.)	f1-score of class 0(Test set)	f1-score of class 1(Test set)
SVM	0.77	0.82	0.73
LR	0.78	0.85	0.77
GNB	0.78	0.80	0.74

	accuracy(CV)	accuracy(Test set)
SVM	0.79	0.78
LR	0.80	0.82
GNB	0.79	0.78

پیش بینی درست زنده ماندن یا نماندن افراد در این مسأله اهمیت بیشتری دارد. پس مدلی که بالاترین مقدار precision را دارد بهتر است. همانطور که مشاهده می شود مقدار این کمیت برای CV در مدل LR بالاتر است و برای داده های تست نیز نزدیک به سایرین است. همچنین دقت عمومی (accuracy) برای این مدل بالاتر از مدل های دیگر است. پس این مدل، از دو مدل دیگر عملکرد بهتری داشته است. در فایل کد مجموعه دادگان موجود در test.csv با X\_submission نمایش داده شده و نتایج مربوط به آن، در فایل gender\_submission.csv قرار گرفته است.