گزارش کار بخش پیاده سازی تمرین دوم درس یادگیری ماشین

محمد لشکری ۴۰۰۱۱۲۰۸۷

۱۳ آذر 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 `KNNImpuer` با در نظر گرفتن ۲۰ همسایه مقادیر گمشده در ویژگی های سن و کرایه بلیت جایگزین شده‌اند. ویژگی های 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 مقیاس کننده استاندارد آنها را مقیاس کرده‌ایم تا عددی در بازه $[-۳, ۳]$ 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 اموزش دیده‌اند:

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

- Gaussian Naïve bayes (GNB)

که درآن کرنل SVM خطی است. نتایج زیر با استفاده از cross validation با تعداد ۵ فلدر به دست آمده است:

	macro ave. of precision
SVM	0.78
LR	0.79
GNB	0.78

	macro ave. of recall
SVM	0.76
LR	0.77
GNB	0.78

	macro ave. of f1-score
SVM	0.77
LR	0.78
GNB	0.78

	accuracy
SVM	0.79
LR	0.80
GNB	0.79

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 می‌شود مقدار این کمیت در مدل LR بالاتر است. همچنین دقت عمومی (accuracy) برای این مدل بالاتر از مدل‌های 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.