گزارش کار بخش پیاده سازی تمرین دوم درس یادگیری ماشین

محمد لشکری ۴۰۰۱۱۲۰۸۷

۱۲ 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

سه الگوریتم زیر روی دادگان اموزش دیده‌اند:

- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Gaussian Naïve bayes (GNB)

که درآن کرنل SVM خطی است.

کلاس صفر نمایانگر زنده ماندن فرد و کلاس یک نشان‌دهنده زنده نمانده آن از حادثه کشتی تایتانیک است. نتایج ارزیابی دادگان مدل‌ها روی داده‌های تست به شرح زیر است:

	class 0	class 1
SVM	1.00	1.00
LR	0.96	0.90
GNB	0.97	0.80

Table 1: precision

	class 0	class 1
SVM	1.00	1.00
LR	0.94	0.93
GNB	0.87	0.95

Table 2: recall

	class 0	class 1
SVM	1.00	1.00
LR	0.95	0.92
GNB	0.91	0.87

Table 3: f1-score

	Accuracy
SVM	1.00
LR	0.94
GNB	0.90

Table 4: accuracy

پیش بینی درست زنده ماندن افراد بیش از درست پیش بینی کردن زنده نماندن آنها اهمیت دارد. پس مدلی که بالاترین مقدار precision برای کلاس یک(زنده ماندن)را دارد بهتر است. همانطور که مشاهده می‌شود دقت مدل SVM برای کلاس یک(زنده ماندن) بالاتر است. همچنین دقت عمومی مدل (accuracy) برای مدل SVM بالاتر از مدل‌های دیگر است. پس این مدل، از دو مدل دیگر برای دادگان تست عملکرد بهتری داشته است. لازم به ذکر است در دادگان این فایل فرض شده که فقط بانوان زنده مانده‌اند. پس این دادگان اریب است.