

# گزارش کار بخش پیاده سازی تمرین سوم - KNN

محمد لشکری

۶ دی ۱۴۰۰

مجموعه متغیرهای مستقل و وابسته به صورت زیر انتخاب شده‌اند:

$X = \{\text{Age, Year\_of\_operation, Number\_of\_positive\_cases}\}$

$Y = \{\text{Survival\_status}\}$

در مرحله پیش پردازش، دادگان آموزش و تست به حالت استاندارد درآمدند تا تأثیر تمام ویژگی‌ها در متر اقلیدسی برای پیش‌بینی یکسان باشد.

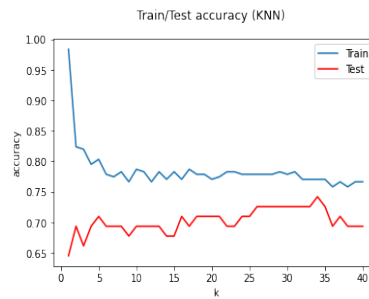
$$X_{train} := \frac{X_{train} - \mu}{\sigma}, X_{test} := \frac{X_{test} - \mu}{\sigma}$$

که در آن  $\mu$  و  $\sigma$  میانگین و واریانس  $X_{train}$  هستند.

## ۱ -k- نزدیک ترین همسایه

### ۱.۱ اجرای الگوریتم به ازای مقادیر مختلف k

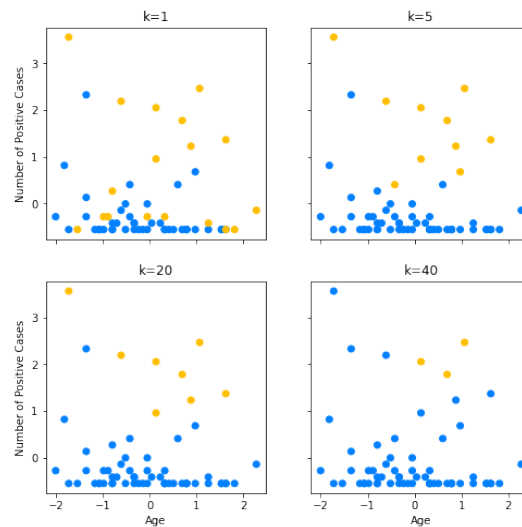
الگوریتم نزدیک ترین همسایگی به ازای اعداد بین ۱ تا ۴۰ به ازای پارامتر k اجرا شد که نتایج آن در شکل ۱.۱ قابل مشاهده است. اختلاف بین دقت دادگان آموزش و تست در عدد ۳۴ کمترین است.



شکل ۱.۱: نمودار دقت به ازای مقادیر مختلف k

## ۲.۱ دیداری سازی برای تحلیل و مقایسه

رنگ آبی نشان دهنده کلاس ۱ و رنگ نارنجی نشان دهنده کلاس ۲ است. همانطور که در شکل ۲.۱ مشاهده می شود با افزایش پارامتر k تمایل داده ها برای پیوستن به کلاس ۱ بیشتر می شود که علت این موضوع می تواند تعداد زیاد برچسب های ۱ در داده های آموزشی باشد.



شکل ۲.۱: نمودار دقت به ازای مقادیر مختلف k

### ۳.۱ تنظیم پارامتر k

در بخش ۱.۱ بهترین تعداد همسایه با توجه به معیار دقت روی دادگان آموزش و تست ۳۴ به دست آمد که علت آن می‌تواند زیاد بودن داده‌های با برچسپ کلاس یک در دادگان تست باشد. همانطور که در بخش ۲.۱ گفته شد با افزایش k همگرایی مدل به کلاس یک بالاتر می‌رود. اما با استفاده از Cross validation برای معیار دقت، این مقدار برابر ۱۱ شد.

پارامتر k	دقت روی داده‌های تست	میانگین دقت روی مجموعه اعتبارسنجی
۱۱	۰/۶۹	۰/۷۷
۳۴	۰/۷۵	۰/۷۴

جدول ۱.۱: دقت مدل برای مقادیر ۱۱ و ۳۴