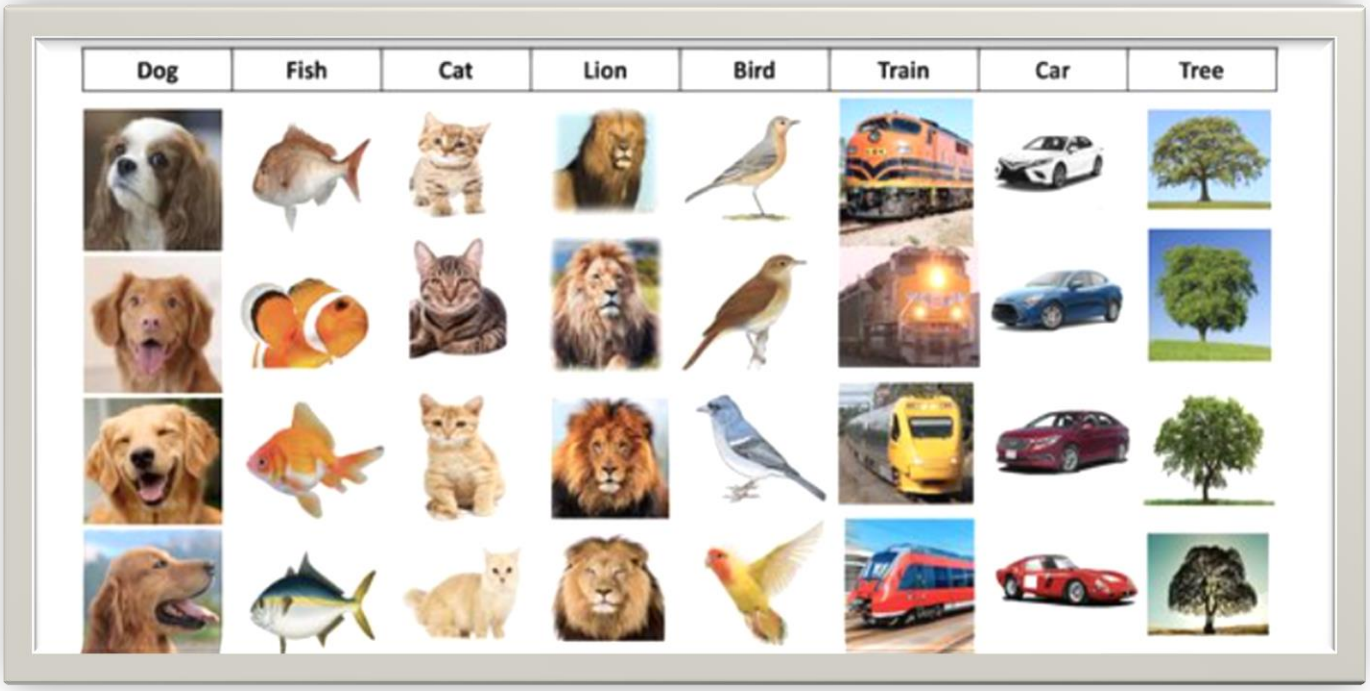
به نام خدا

پروژه هوش مصنوعی



یادگیری ماشین

تخمین بر اساس داده های اموزشی

دانشجو: مهدی براتی

شماره دانشجوی: 9912358008

مقدمه:

در این پروژه پیش پردازشی بر روی داده ها انجام شده تا داده های غیر نرمال حذف شوند و در مرحله بعد فرایند یادگیری صورت گرفته و در نهایت تعدادی داده جهت تست داده شده و نتیجه را به صورت اماری نمایش داده ایم

توضیحات بخش های مختلف

1. ابتدا داده ها را نمایش میدهیم تا نوع انها را و فراوانی و همچنین بازه مقادیر را جهت پردازش بدست اوریم
2. نمایش اماری داده ها را انجام میدهیم تا متوجه شویم کدام یک از ویژگی ها دارای داده دور افتاده است.
3. برای ویژگی هایی که داده پرت دارند نمودار scatter plotرا برای انها نمایش میدهیم تا داده های پرت واضح تر نمایان شوند
4. fcوpx_height دارای داده دور افتاده هستند. پس با حذف داده های دور به روش محدود کردن بازه مقادیر میتوان داده های پرت را حذف نمود (باید در نظر بگیریم که رکورد هایی ک ان مقدار برایشان null در نظر گرفته شده را حذف نمیکنیم)
5. در مرحله حذف داده های تکراری چک میکنیم که داده تکراری نباشد.و باید توجه داشته باشیم که index یک ویژگی منحصر به فرد است و باید انرا حذف کنیم سپس چک کنیم که داده نتکراری نباشد.

6. باید داده هایی ک مقادیر انها توصیفی است را به مقادیر عددی تبدیل کنیم برا ی این کار اطلاعات ویژگی ها را نمایش میدهیم تا مقادیری که توصیفی هستند را پیدا کنیم. و انها را به یک عدد مناسب متناظر کنیم.

7. برای برخی از رکورد ها مقدار یک یا چند ویژگی nullقرار گرفته است و باتوجه به این که بعضی از این ویژگی هااز حالت توصیفی تبدیل به عدد شده اند لذا مجاز به استفاده از تابع میانگین نیستیم زیرات ممکن است مقداری بجای nullقرار دهد که از نظر مفهومی صحیح نیست مانند بلوتوث داشتن یا نداشتن که با 0و1 متناظر شده را مقدار میانگین قرار دهیم،عدد (0.7) مقدار معتبری نخواهد بود . به همین دلیل از تابع میانه برای پر کردن بخش های nullاستفاده کرده ایم.

8. ستون price_range را جدا میکنیم و داده ها را به دو بخش trainو تست تقسیم میکنیم train اطلاعات برای اموزش و تست جهت بررسی اینکه مدل تا چه حدی درست کار می کند استفاده میکنیم.

9. پس از جدا کردن price_range مقادیر را به بین 0 و1 scaleمیکنیم و سپس مدل های مختلف را برای پیش بینی استفاده میکنیم

تخمین محدوده قیمت با توجه به ویژگی ها

برای تخمین از 5 الگوریتم Decision classifierو

kneighborsclassifierوLogisticRegressionوRandomforestclassifierو

GassianNB

مدلLogisticRegression از سایر مدل ها نتیجه بهتری برای پیشبینی داشت

```
[ ]  model = LogisticRegression()
     model.fit(X_train, y_train)
     y_pred = model.predict(X_test)
     print(classification_report(y_test, y_pred))
```

```
                  precision    recall  f1-score   support

               0       0.90      0.90      0.90        51
               1       0.87      0.85      0.86        46
               2       0.95      0.92      0.93        38
               3       0.97      1.00      0.98        63

        accuracy                           0.92       198
       macro avg       0.92      0.92      0.92       198
    weighted avg       0.92      0.92      0.92       198
```

```
print('Accuracy: ', accuracy_score(y_test, y_pred))
print('Precision: ', precision_score(y_test, y_pred,average='micro'))
print('Recall: ', recall_score(y_test, y_pred,average='micro'))
print('F1-Score: ', f1_score(y_test, y_pred,average='micro'))
```

```
Accuracy:  0.9242424242424242
Precision:  0.9242424242424242
Recall:  0.9242424242424242
F1-Score:  0.9242424242424242
```