Московский государственный технический университет
им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Рубежный контроль № 1

По курсу «методы машинного обучения в АСОИУ»

Выполнил:

студент ИУ5-24М
Ширшов А.С.

Проверил:

Гапанюк Ю.Е.

Подпись:

29.02.2024

Москва, 2024

Задание

- Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием функции "обратная зависимость - 1 / X".
- Для набора данных проведите процедуру отбора признаков (feature selection). Используйте метод обертывания (wrapper method), алгоритм полного перебора (exhaustive feature selection).
- Для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Ход работы

Для выполнения работы возьмем набор хорошо известный тестовый набор данных про вино.

```
wine = load_wine()
df = pd.DataFrame(wine.data, columns=wine.feature_names)
df['target'] = wine.target
print("Первые 5 строк исходного набора данных:")
print(df.head())
```

✓ 0.0s

```
Первые 5 строк исходного набора данных:
   alcohol  malic_acid   ash  alcalinity_of_ash  magnesium  total_phenols  \
0    14.23        1.71  2.43               15.6      127.0           2.80
1    13.20        1.78  2.14               11.2      100.0           2.65
2    13.16        2.36  2.67               18.6      101.0           2.80
3    14.37        1.95  2.50               16.8      113.0           3.85
4    13.24        2.59  2.87               21.0      118.0           2.80

   flavanoids  nonflavanoid_phenols  proanthocyanins  color_intensity   hue  \
0        3.06                  0.28             2.29             5.64  1.04
1        2.76                  0.26             1.28             4.38  1.05
2        3.24                  0.30             2.81             5.68  1.03
3        3.49                  0.24             2.18             7.80  0.86
4        2.69                  0.39             1.82             4.32  1.04
```

Рисунок 1 - Вывод набора данных

Построение обратной зависимости для поля - “Алкоголь”

```
transformer = FunctionTransformer(func=lambda x: 1 / x, validate=True)
df['alcohol_transformed'] = transformer.transform(df[['alcohol']])
print("\nПервые 5 строк после нормализации признака 'alcohol':")
print(df[['alcohol', 'alcohol_transformed']].head())
```

[29] ✓ 0.0s

```
Первые 5 строк после нормализации признака 'alcohol':
   alcohol  alcohol_transformed
0    14.23             0.070274
1    13.20             0.075758
2    13.16             0.075988
3    14.37             0.069589
4    13.24             0.075529
```

Рисунок 2 - Нормализация числового признака при помощи обратной зависимости

Для набора данных проведём процедуру отбора признаков (feature selection). Алгоритм полного перебора (exhaustive feature selection)

```
model = LinearRegression()
# Выполняем исчерпывающий отбор признаков (exhaustive feature selection)
efs = EFS(model,
          min_features=1,
          max_features=5,
          scoring='r2',
          print_progress=True,
          cv=5)

efs = efs.fit(X, y)
```

✓ 19.3s

Рисунок 3 - Использование EFS

Ниже представлены итоговые полученные признаки.

```
# Проверка выбранных признаков
selected_features = X.columns[rfe.support_]
print("\nВыбранные признаки:", selected_features)
```

✓ 0.0s

```
Выбранные признаки: Index(['alcohol', 'flavanoids', 'hue', 'od280/od315_of_diluted_wines',
       'alcohol_transformed'],
      dtype='object')
```

Рисунок 4 - Итоговые полученные признаки

Также выполним небольшое доп. задание. Построим violin plot для поля alcohol.

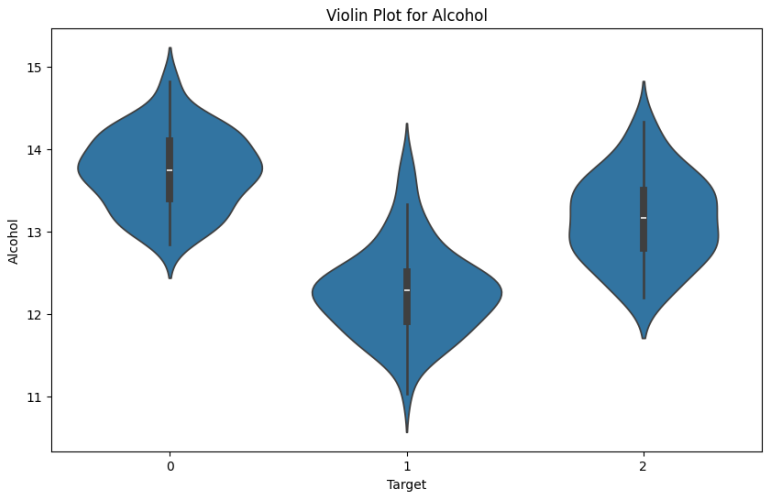

Рисунок 5 - График violin plot

Вывод

В ходе выполнения данного задания были проведены шаги по предварительной обработке и анализу набора данных, такие как нормализация числовых признаков, отбор наиболее значимых признаков и визуализация данных с помощью графиков. Они могут значительно улучшить качество и точность моделей машинного обучения, которые будут использоваться для анализа данного набора данных в будущем.