Отчёт по заданию 3 «Классификация методом kNN»

Выполнил студент 4 курса 7.1 группы Путин Павел

Вариант 11 «2004 New Car and Truck Data»

1 Описание данных:

Приведены технические характеристики 428 новых автомобилей, выпущенных в 2004 году. Зафиксированные параметры включают в себя цену, размеры автомобиля и его топливную экономичность.

Среди них 60 автомобилей – внедорожники (14%).

Колонка	Имя параметра
1-45	Название транспортного средства
47	Спортивный автомобиль? (1=да, 0=нет)
49	Спортивный внедорожник? (1=да, 0=нет)
51	Универсал? (1=да, 0=нет)
53	Минивэн? (1=да, 0=нет)
55	Пикап? (1=да, 0=нет)
57	Полный привод? (1=да, 0=нет)
59	Задний привод? (1=да, 0=нет)
61-66	Рекомендованная розничная цена (в долларах США)
68-73	Дилерская стоимость (в долларах США)
75-77	Объем двигателя (л)
79-80	Количество цилиндров (=-1 для роторного двигателя)
82-84	Лошадиных силы
86-87	Миль на галлон по городу
89-90	Миль на галлон по шоссе
92-95	Вес (фунтов)
97-99	Колесная база (дюймов)
101-103	Длина (дюймов)
105-106	Ширина (дюймы)

2 Решение

2.1 Выделение обучающей и текстовой выборки

Обучающая выборка состоит из 84 объектов (17,9% от общего числа), среди них 12 внедорожников (14%) и 72 других автомобиля (86%) (см рисунок 1). Обучение проходило по 6 критериям: наличие полного привода, объём двигателя, количество цилиндров, количество лошадиных сил, вес, колёсная база.

```
def get_train_and_real_data(df):
  df_train_sport_utility = df[df[column_names.SPORT_UTILITY_VEHICLE] == True].sample(12)
  df_train_other = df[df[column_names.SPORT_UTILITY_VEHICLE] == False].sample(72)
  df_train = pd.concat([df_train_sport_utility, df_train_other])
  total_train, su_train, other_train = describe(df_train)
  train_classes = df_train[column_names.SPORT_UTILITY_VEHICLE].apply(get_prediction_name)
  temp = df_train.copy()
  df_train = prepare_for_prediction(df_train)
  data_train = Data(df_train, train_classes, total_train, su_train, other_train)

  df_real = pd.concat([df, temp]).drop_duplicates(keep=False)
  total_real, su_real, other_real = describe(df_real)
  real_classes = df_real[column_names.SPORT_UTILITY_VEHICLE].apply(get_prediction_name)
  df_real = prepare_for_prediction(df_real)
  data_real = Data(df_real, real_classes, total_real, su_real, other_real)

  return data_train, data_real
```

Рисунок 1 - Метод выделения обучающей и тестовой выборки и контрольных значений классов

2.2 Определение наилучшего значения k

Для определения наилучшего значения k было выполнено по 50 запусков обучения модели на 50 разных обучающих данных (см рисунок 2):

```python
def find_target_k(df):
    df_accuracies = pd.DataFrame(columns=('k', 'accuracy'))
    accuracies = []
    for k in range(5, 11):
      temp = []
      for _ in range(50):
        train, real = get_train_and_real_data(df)
        df_train, train_classes = train.values, train.classes
        knn = KNeighborsClassifier(n_neighbors=k)
        knn.fit(df_train, train_classes)
        knn_accuracy_test_predictions = knn.predict(df_train)
        temp.append(accuracy_score(train_classes, knn_accuracy_test_predictions))
      accuracies.append(statistics.mean(temp))

    train, real = get_train_and_real_data(df)

    df_accuracies['k'] = list(range(5, 11))
    df_accuracies['accuracy'] = accuracies

    idx = df_accuracies['accuracy'].idxmax()
    target_k = df_accuracies['k'][idx]
    return df_accuracies,target_k
```

Рисунок 2 - Определение оптимального значения k

На основе максимального среднего значения выбирается лучшее k. В силу случайности выборок, k может принимать значения 5 (чаще) или 6 (см рисунок 3).

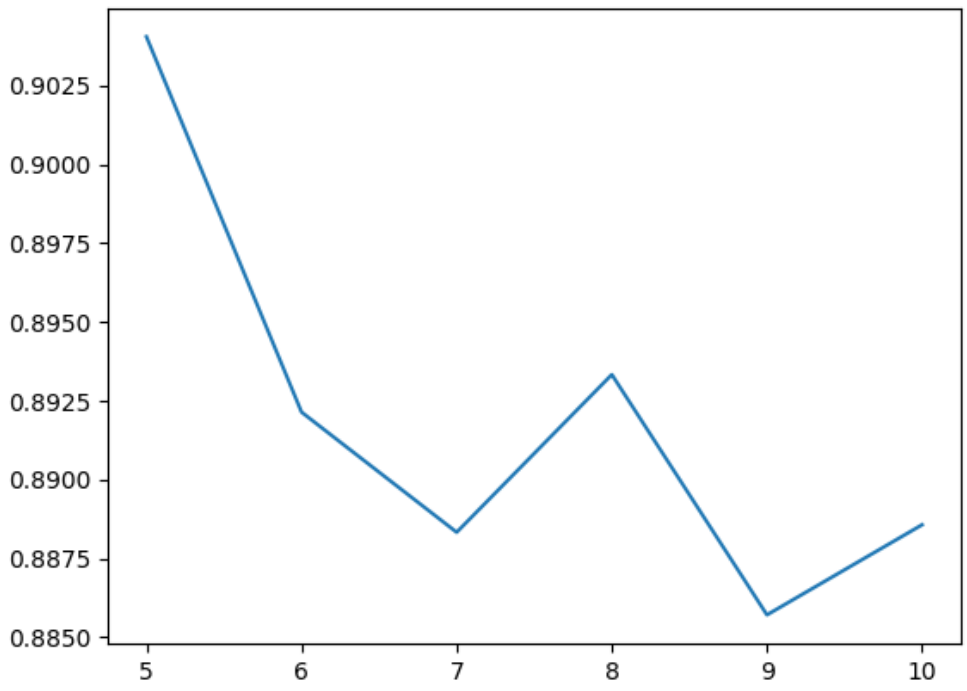

Рисунок 3 - Зависимость точности классификатора от значения k

2.3 Оценка качества прогноза на тестовой выборке с помощью таблицы сопряженности

На основе зависимости количества автомобилей от типа и одного из параметров, участвовавших в обучении классификатора, была выявлена явная зависимость типа автомобиля от веса, наличия полного привода и колёсной базы (см рисунки 4 - 9).

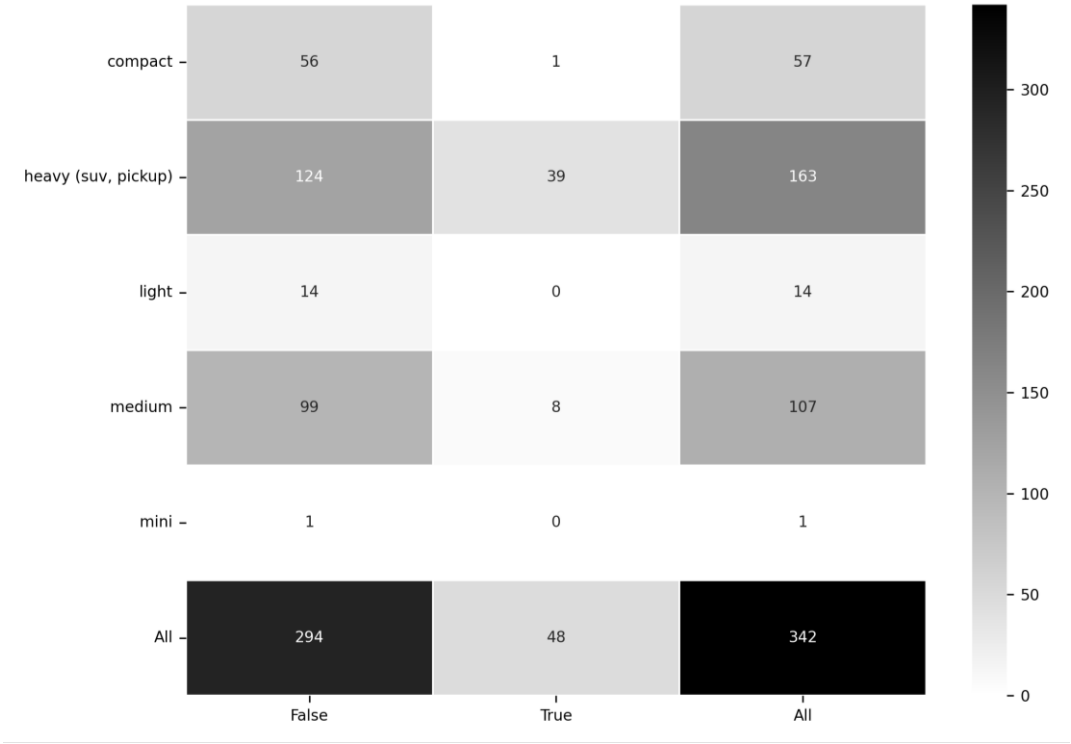

Рисунок 4 - Связь веса и типа автомобиля

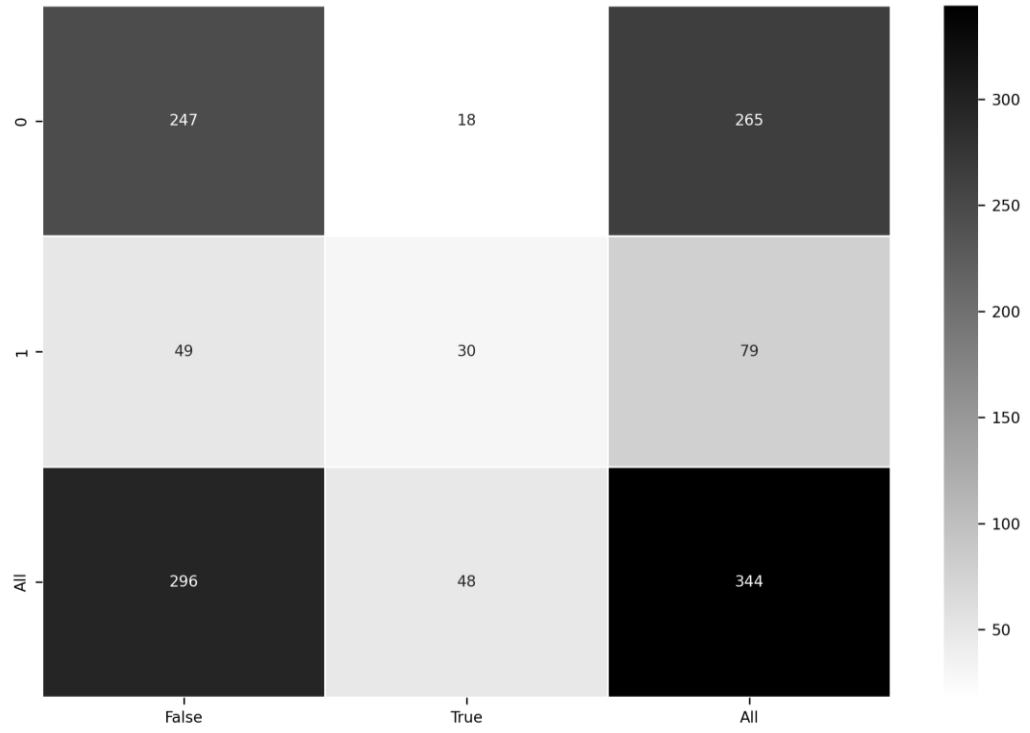

Рисунок 5 - Связь наличия полного привода и типа автомобиля

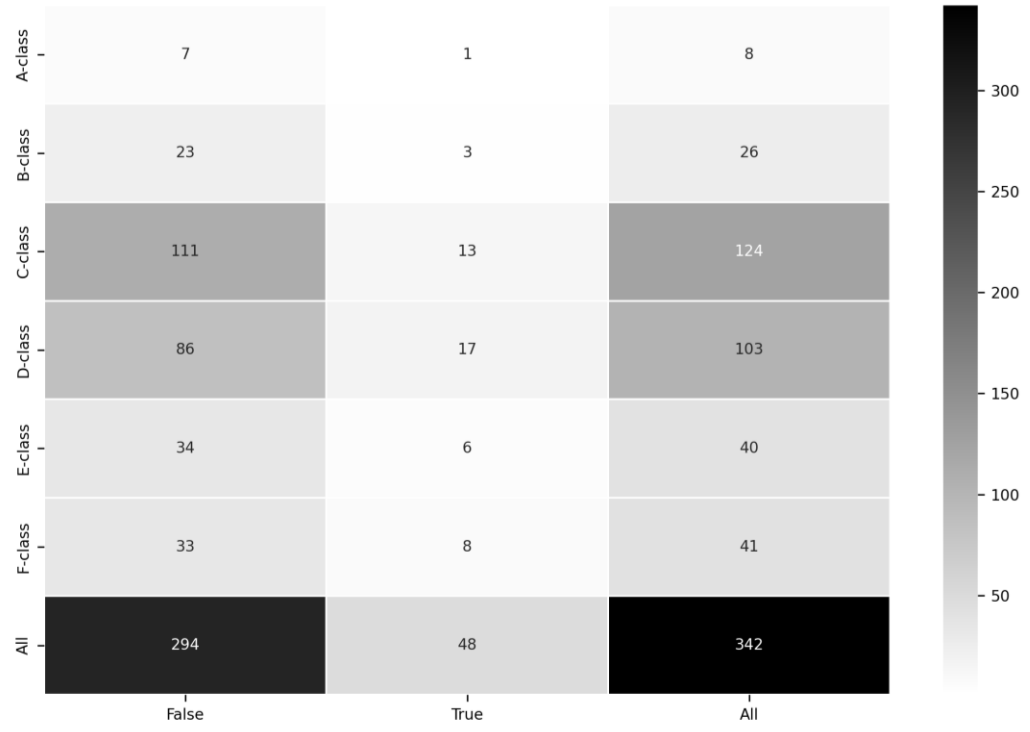

Рисунок 6 - Связь колёсной базы и типа автомобиля

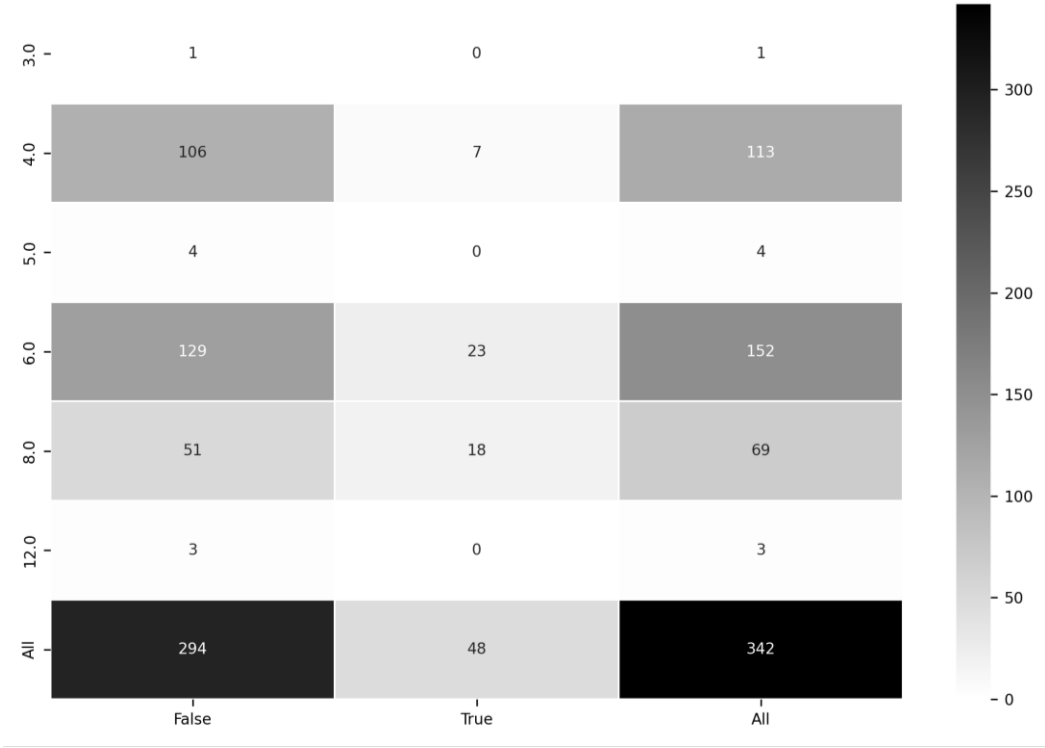

Рисунок 7 - Связь количества цилиндров и типа автомобиля

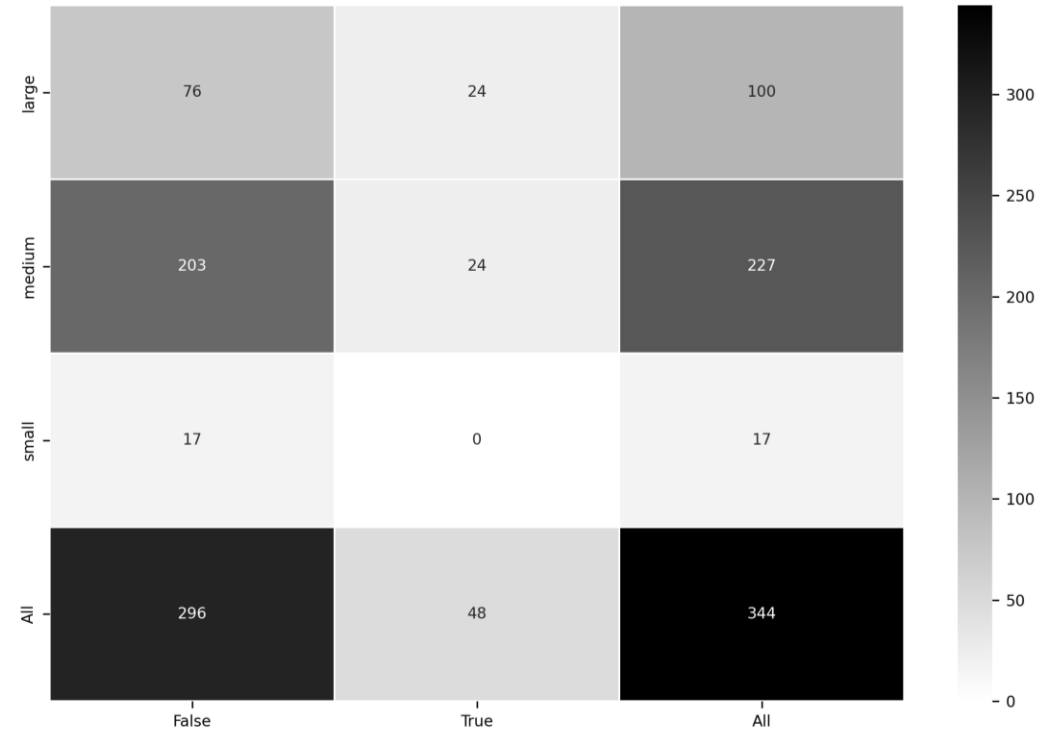

Рисунок 8 - Связь объёма двигателя и типа автомобиля

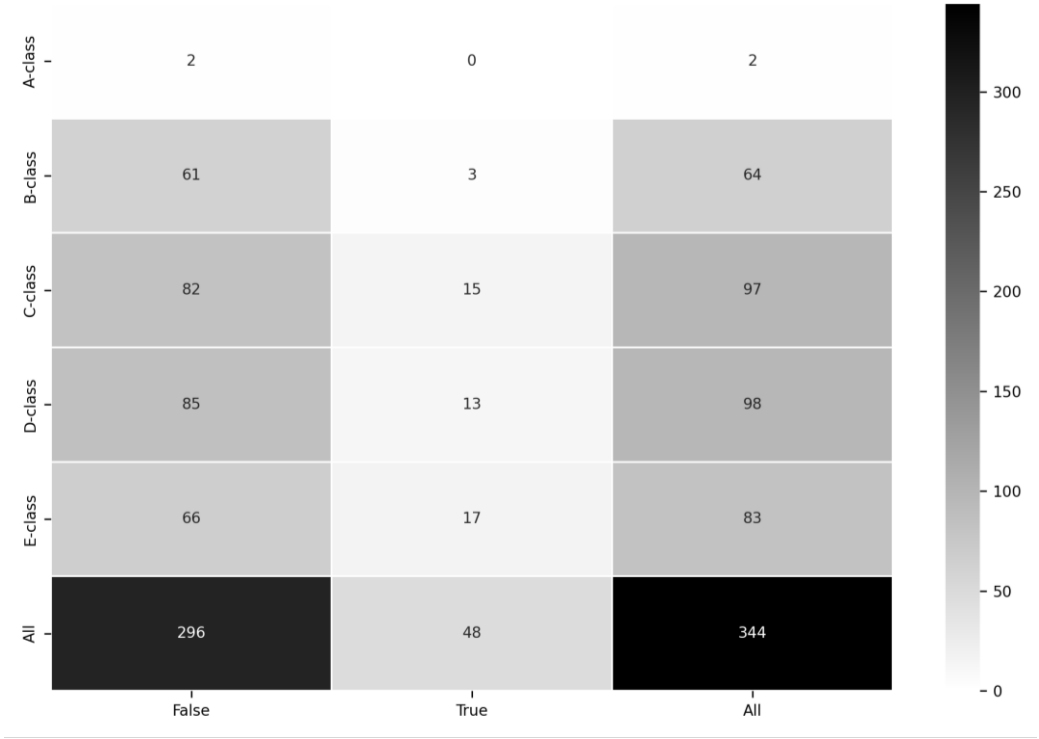

Рисунок 9 - Связь количества лошадиных сил и типа автомобиля

3 Выводы

В итоге, был получен следующий результат точности классификации (k равен 5):

Набор данных	Точность классификатора
Обучающие	89,29%
Тестовые	86,92%