**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**

**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра вычислительной техники**

**ОТЧЕТ**

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

**Тема: Классификация**

| Студент |  |  |
| --- | --- | --- |
| Преподаватель |  | Гатауллин Р.И. |

Санкт-Петербург

2023

# Оглавление

[Оглавление 2](#_gjdgxs)

Цель, задачи и краткая теоретическая информация3

Ход работы [5](#_2et92p0)

[Вывод 23](#_tyjcwt)

# Цель работы: целью данной работы является изучение и практическое применение методов классификации в контексте машинного обучения. Основная цель заключается в обучении моделей и подборе параметров, оценки моделей.

# Задачи работы: в рамках данной работы необходимо выполнить следующие задачи - изучить основные принципы и методы классификации, подобрать параметры и обучить модели, оценить модели.

**Краткая теоретическая информация:** классификация - это задача в области машинного обучения, которая заключается в присвоении объектам определенных классов на основе их характеристик или признаков. Она основывается на предварительно подготовленном наборе данных, где каждому объекту уже известно, к какому классу он относится. Используя эти данные, модель классификации обучается так, чтобы предсказывать классы для новых, еще неизвестных объектов.

Ниже представлена краткая информация о некоторых алгоритмах классификации:

1. Метод k-ближайших соседей (K-NN): K-NN относит объекты к определенным классам, основываясь на их близости к K ближайшим обучающим примерам. Он измеряет расстояние между точками в пространстве признаков и использует голосование ближайших соседей для определения класса объекта.

2. Метод опорных векторов (SVM): SVM строит гиперплоскость или набор гиперплоскостей в многомерном пространстве для разделения объектов разных классов. Он стремится найти гиперплоскость, которая наилучшим образом разделяет точки двух классов и имеет наибольший отступ от ближайших обучающих примеров.

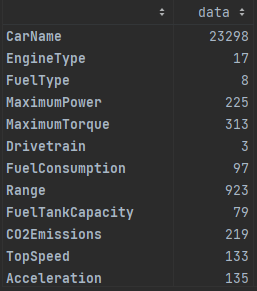
3. Деревья решений: Деревья решений используются для классификации и прогнозирования путем построения дерева, где каждый узел представляет собой разделение данных по одному из признаков. Дерево последовательно принимает решения на основе значений признаков, чтобы достичь окончательного решения. Оно представляет собой структуру в виде ветвей и листьев, где каждый лист соответствует определенному классу или прогнозу. Деревья решений позволяют легко интерпретировать результаты и принимать решения на основе логических правил.

**Ход работы**

Датасет: <https://www.kaggle.com/datasets/shrirangmhalgi/world-cars-data>

# Подготовка датасета

# Распределение данных (Рис.2), которые имеют большое количество уникальных значений (Рис.1) и имеют сильную корреляцию с целевой функцией – TopSpeed, по категориям (TopSpeed, MaximumPower, Acceleration):



# Рис.1

# 

# Рис. 2

# Результат выполнения данных функций:

# 

# Рисунок 2

# Все не числовые параметры кодируются номерами:

# 

# 

# Рисунок 3

# Исходный вид датасета после подготовки:

# 

# Рисунок 4

# Целевая переменная – TopSpeed.

# 

# Рисунок 5

# K-ближайших соседей (KNN):

# Для подбора оптимальных параметров используем GridSearchCV:

# 

# Рисунок 6

# Матрица ошибок:

# 

# Рисунок 7

# Оценка качества прогноза:

# 

# Рисунок 8

# 1. Точность (precision): Для класса 1 точность составляет 0.87, для класса 2 - 0.79 и для класса 3 - 0.90. Общее средневзвешенное значение точности (weighted avg) равно 0.85. Это означает, что модель достаточно точно классифицирует данных для каждого из классов.

# 2. Отзыв (recall): Отзыв для класса 1 составляет 0.89, для класса 2 - 0.79 и для класса 3 - 0.88. Это означает, что модель справляется лучше с классом 1 и 3 (имеет высокий отзыв), чем с классом 2.

# 3. F1-мера (f1-score): F1-мера для класса 1 составляет 0.88, для класса 2 - 0.79 и для класса 3 - 0.89. F1-мера является средним гармоническим между точностью и отзывом, поэтому она предоставляет общую оценку производительности модели. Значения F1-меры указывают на хорошую сбалансированность между точностью и отзывом для всех классов.

# 4. Поддержка (support): Поддержка для каждого класса указывает на количество образцов, принадлежащих каждому классу. В данном случае, класс 1 имеет поддержку 1747, класс 2 - 1642, и класс 3 - 1668.

# 5. Общая точность (accuracy): Общая точность модели составляет 0.85, что означает, что модель правильно классифицировала 85% образцов из общего количества образцов. Более высокое значение accuracy указывает на лучшую производительность модели.

# 6. ROC/AUC: Значение ROC/AUC равно 0.95, что указывает на хорошую производительность модели в предсказании мультиклассовых вероятностей. Значение ROC/AUC ближе к 1 указывает на лучшую способность модели различать классы.

# В целом, модель KNN показывает хорошие результаты в классификации мультиклассовых данных, с достаточно высокой точностью, отзывом и F1-мерой.

# Машина опорных векторов (SVM)

# Для подбора оптимальных параметров используем GridSearchCV:

# 

# Рисунок 9

# Результат подбора параметров:

# 

# Рисунок 10

# Оцениваем точность модели:

# 

# Рисунок 11

# Матрица ошибок:

# 

# Рисунок 12

# Оценка качества прогноза:

# 

# Рисунок 13

# - Precision (точность): Показывает, насколько модель точно классифицирует каждый класс. Значения precision для каждого класса составляют 92% для класса 1, 85% для класса 2 и 93% для класса 3. Это означает, что модель довольно точна в предсказании всех трех классов.

# - Recall (полнота): Оценивает способность модели правильно обнаруживать примеры каждого класса. Значения recall для каждого класса составляют 93% для класса 1, 86% для класса 2 и 91% для класса 3. Это означает, что модель хорошо находит всех 3 классов.

# - F1-score: Среднее гармоническое между precision и recall. Он учитывает как точность, так и полноту. Значения f1-score для каждого класса составляют 93% для класса 1, 86% для класса 2 и 92% для класса 3.

# - Support (поддержка): Показывает количество примеров в каждом классе.

# - Accuracy (точность): Показывает общую точность модели в предсказании всех классов. В данном случае, модель достигла точности 90%.

# - Macro avg: Среднее значение precision, recall и f1-score для всех классов. В данном случае, среднее значение составляет 90%.

# - Weighted avg: Средневзвешенное значение precision, recall и f1-score, учитывая количество примеров в каждом классе. В данном случае, средневзвешенное значение составляет 90%.

# Относительно ROC/AUC (Receiver Operating Characteristic / Area Under Curve), это метрика, которая оценивает способность модели разделять классы на основе значений вероятности. Значение 0.963 указывает на высокую производительность модели, где ближе значение к 1, тем лучше.

# Общий вывод: Модель опорных векторов показывает хорошие результаты с точки зрения точности, полноты, F1-меры и показателя ROC/AUC.

# Дерево решений:

# Для подбора оптимальных параметров используем GridSearchCV:

# 

# Рисунок 14

# Оцениваем точность модели:

# 

# Рисунок 15

# Матрица ошибок:

# 

# Рисунок 16

# Оценка качества прогноза:

# 



# Рисунок 17

# 1. Точность (precision): Для класса 1 (Label 1) точность составляет 0.91, для класса 2 (Label 2) - 0.79 и для класса 3 (Label 3) - 0.92. Это означает, что модель дерева решений довольно точно классифицирует данные для каждого из классов.

# 2. Отзыв (recall): Отзыв для класса 1 составляет 0.88, для класса 2 - 0.84 и для класса 3 - 0.89. Это означает, что модель добротно предсказывает все 3 класса.

# 3. F1-мера (f1-score): F1-мера для класса 1 составляет 0.89, для класса 2 - 0.82 и для класса 3 - 0.90. F1-мера является средним гармоническим между точностью и отзывом, и она предоставляет общую оценку производительности модели. Значения F1-меры указывают на сбалансированность между точностью и отзывом для всех классов, но с небольшим преимуществом класса 1.

# 4. Поддержка (support): Поддержка для каждого класса представлена в последнем столбце отчета. Это количество экземпляров каждого класса в тестовой выборке.

# 5. ROC/AUC: Значение ROC/AUC равно 0.8939, что говорит о хорошей способности модели предсказывать классы в задаче классификации. Значение AUC (площадь под кривой ROC) близко к 1, что указывает на хорошую производительность модели.

# Производительность модели дерева решений, оцененная с использованием классификационного отчета и ROC/AUC, получилась довольно хорошей.

# Точность, отзыв и F1-мера для всех классов составляют 0.87, что указывает на способность модели достаточно точно классифицировать данные для каждого класса.

# Значение ROC/AUC равно 0.96 , что говорит о хорошей способности модели различать между классами.

# Модель дерева решений хорошо справляется с задачей классификации и может быть эффективно использована для предсказания классов на новых данных.

# Вывод дерева решений:

# 

# Рисунок 18

# Вывод по работе

# В ходе выполнения данной работы были изучены и применены методы классификации в контексте машинного обучения. Результаты работы показали хорошие результаты задачи классификации и могут в дальнейшем использоваться для предсказания классов на новых значениях.