Отчет

**Выбранный набор данных:**

Набор данных диабета индейцев Пима

Набор данных о диабете индейцев пима включает прогнозирование возникновения диабета в течение 5 лет у индейцев пима с учетом медицинских данных.

Это бинарная (2-классная) задача классификации. Количество наблюдений для каждого класса не сбалансировано. Есть 768 наблюдений с 8 входными переменными и 1 выходной переменной. Считается, что пропущенные значения кодируются нулевыми значениями. Имена переменных следующие:

1. Количество раз беременных.
2. Концентрация глюкозы в плазме через 2 часа при оральном тесте на толерантность к глюкозе.
3. Диастолическое артериальное давление (мм рт. Ст.).
4. Толщина трехглавой кожной складки (мм).
5. 2-часовой сывороточный инсулин (мю Ед / мл).
6. Индекс массы тела (вес в кг / (рост в м) ^ 2).
7. Родословная функция диабета.
8. Возраст (годы).
9. Переменная класса (0 или 1). (target value)

Базовая эффективность прогнозирования наиболее распространенного класса - точность классификации приблизительно 65%. Лучшие результаты достигают точности классификации примерно 77%.

<https://machinelearningmastery.ru/standard-machine-learning-datasets/>

**Методы классификации:**

1. KNN-model classifier in sk-learn library
2. SVM-model classifier in sk-learn library
3. Linear-regression classifier in sk-learn library

|  |
| --- |
| import sklearn.linear\_model as lnm  logistic\_model=lnm.LogisticRegression()  logistic\_model.fit(data,target) |

Метод линейной регрессии из библиотеки sklearn.

Линейная регрессия соответствует линейной модели с коэффициентами, минимизирующими остаточную сумму квадратов между наблюдаемыми целями в наборе данных и целями, предсказанными линейной аппроксимацией.

**Сравнение качество работы методов**:

Accuracy of test cases(20% from data set):

|  |  |
| --- | --- |
| **Model name** | **prediction accuracy** |
| Support Vector Machine | 0.8116883116883117 |
| K-Nearest-Neighbor | 0.8636363636363636 |
| Linear-regression | 0.8181818181818182 |

3 measures from test cases:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F-measure** |
| SVM | 0.38 | 0.35 | 0.36 |
| KNN | 0.42 | 0.44 | 0.43 |
| Linear-reg | 0.73 | 0.64 | 0.68 |

**Вывод:**

Большую часть из наблюдении наблюдалось высшим prediction accuracy в моделе KNN. Но 3-меры наблюдалось в линейном регрессии большим из 3 моделей.