Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и Системы управления»

КАФЕДРА «Автоматизированные системы обработки информации и управления»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К КУРСОВОЙ РАБОТЕ

НА ТЕМУ:

Решение комплексной задачи машинного обучения

Студент ИУ5ц-83Б (Группа) _____ (Подпись, дата) Костников И.А. (И.О.Фамилия)

Руководитель курсовой работы _____ (Подпись, дата) Гапанюк Ю.Е. (И.О.Фамилия)

Консультант _____ (Подпись, дата) Гапанюк Ю.Е. (И.О.Фамилия)

2021 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой _____
(Индекс)
_____ _____
(И.О.Фамилия)
« _____ » _____ 2021 г.

З А Д А Н И Е
на выполнение курсовой работы

по дисциплине ____ Технологии машинного обучения _____

Студент группы _____ИУ5ц-83Б_____

____Костников Иван Алексеевич_____
(Фамилия, имя, отчество)

Тема курсовой работы ____ решение комплексной задачи машинного обучения _____
_____
_____

Направленность КР (учебная, исследовательская, практическая, производственная, др.)
_____
Источник тематики (кафедра, предприятие, НИР) _____

График выполнения работы: 25% к ___ нед., 50% к ___ нед., 75% к __ нед., 100% к ___ нед.

Задание ____ решение задачи машинного обучения на основе материалов дисциплины. Выполняется студентом единолично. _____
_____
_____
_____

Оформление курсовой работы:

Расчетно-пояснительная записка на 12 листах формата А4.

_____
_____
_____
_____

Дата выдачи задания « 10» мая 2021 г.

Руководитель курсовой работы _____ Гапанюк Ю.Е.
(Подпись, дата) (И.О.Фамилия)

Студент _____ Костников И.А.
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

1 Введение

Данная работа является курсовым проектом по дисциплине технологии машинного обучения и включает в себя весь изученный материал, который был представлен в течении семестра

2 Цель работы

Целью работы является исследование предметной области, изучение базы данных, её преобразование и на основе полученных чистых данных обучить наиболее подходящие для решения задачи моделей.

3 Задание

Схема типового исследования, проводимого студентом в рамках курсовой работы, содержит выполнение следующих шагов:

3.1 Поиск и выбор набора данных для построения моделей машинного обучения. На основе выбранного набора данных студент должен построить модели машинного обучения для решения или задачи классификации, или задачи регрессии.

3.2 Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.

3.3 Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.

3.4 Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения. В зависимости от набора данных, порядок выполнения пунктов 2, 3, 4 может быть изменен.

3.5 Выбор метрик для последующей оценки качества моделей. Необходимо выбрать не менее трех метрик и обосновать выбор.

3.6 Выбор наиболее подходящих моделей для решения задачи классификации или регрессии. Необходимо использовать не менее пяти моделей, две из которых должны быть ансамблевыми.

3.7 Формирование обучающей и тестовой выборок на основе исходного набора данных.

3.8 Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров. Производится обучение моделей на основе обучающей выборки и оценка качества моделей на основе тестовой выборки.

3.9 Подбор гиперпараметров для выбранных моделей. Рекомендуется использовать методы кросс-валидации. В зависимости от используемой библиотеки можно применять функцию GridSearchCV, использовать перебор параметров в цикле, или использовать другие методы.

3.10 Повторение пункта 8 для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей.

3.11 Формирование выводов о качестве построенных моделей на основе выбранных метрик. Результаты сравнения качества рекомендуется отобразить в виде графиков и сделать выводы в форме текстового описания. Рекомендуется построение графиков обучения и валидации, влияния значений гиперпарметров на качество моделей и т.д.

3.12 Формирование web-приложения

Приведенная схема исследования является рекомендуемой. В зависимости от решаемой задачи возможны модификации.

4 Основная часть

4.1 Описание данных

В качестве основного датафрейма выбрана база данных, содержащая информацию о рейтингах блюд, оцененных посетителями, основываясь на ингредиентах, содержащихся в них.

4.2 Решаемая задача

Основной задачей выбранных данных является классификация блюд по рейтингу, основываясь на ингредиентах, содержащихся в них.

4.3 Выбранные модели для обучения

Мной были выбраны следующие модели:

1) AdaBoostClassifier,
2) BaggingClassifier,
3) ExtraTreesClassifier,
4) GradientBoostingClassifier,
5) RandomForestClassifier,
6) LinearSVC,
7) SVC

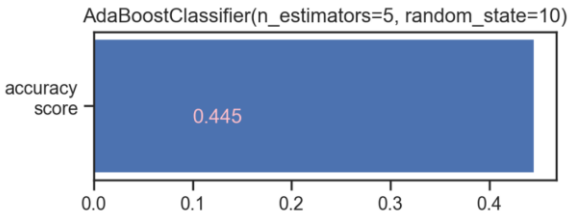
4.4 Выбранные метрики для оценки качества

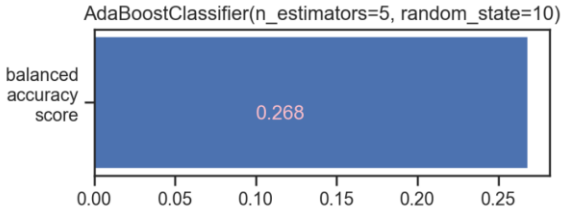
Мной были выбраны следующие модели:

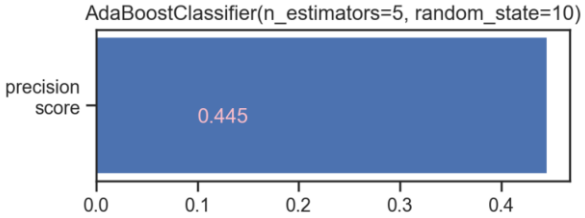
1) accuracy_score,
2) precision_score,
3) recall_score,
4) balanced_accuracy_score

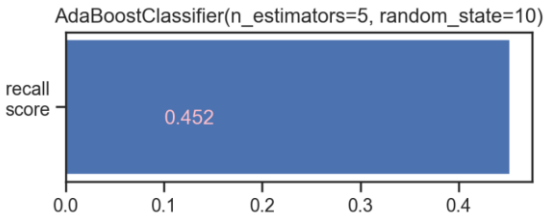
4.5 Построение графиков для выбранных моделей без подбора гипер-парамтров

4.5.1 AdaBoostClassifier

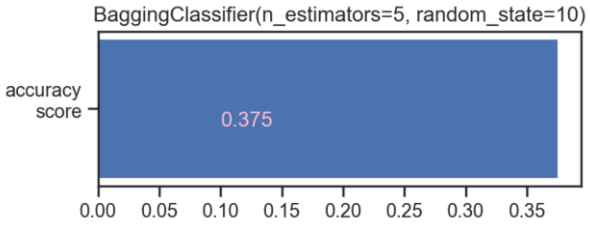


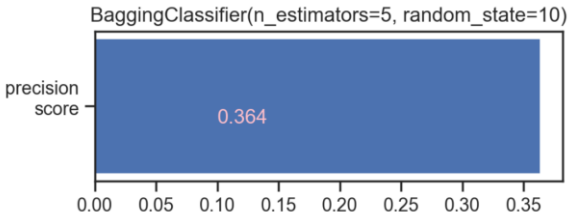


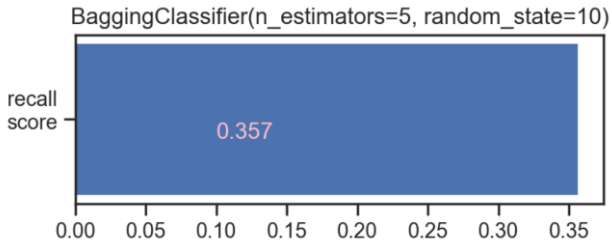


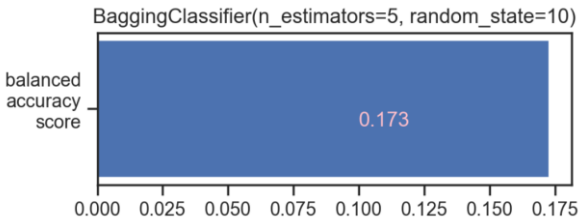


4.5.2 BaggingClassifier

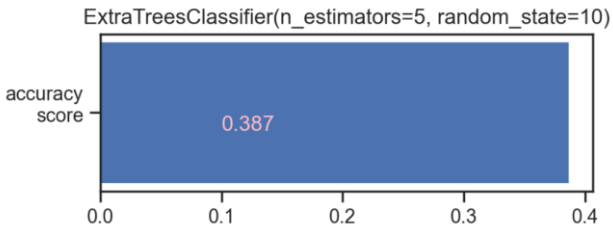


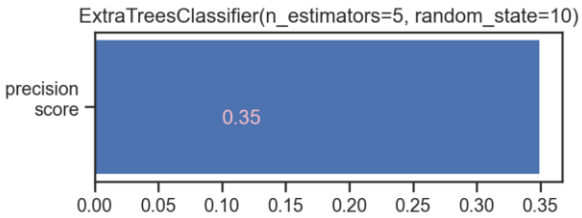


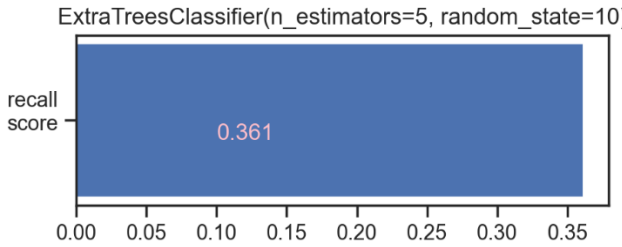


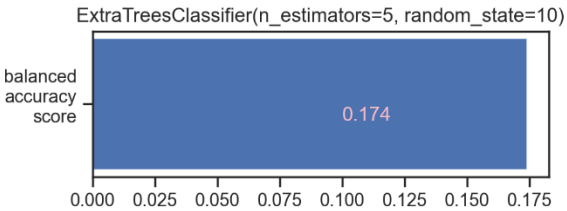


4.5.3 ExtraTreesClassifier

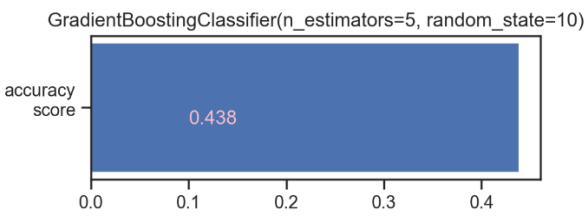


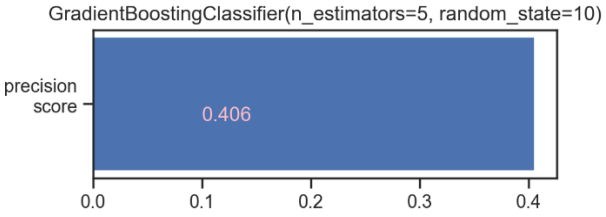


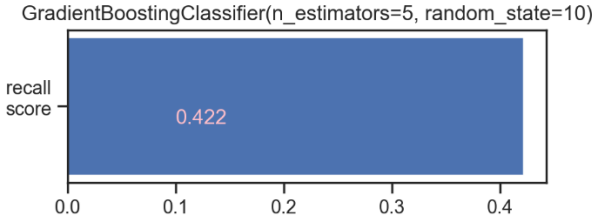


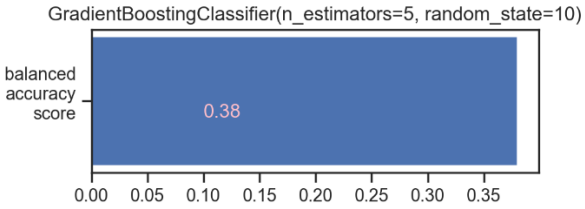


4.5.4 GradientBoostingClassifier

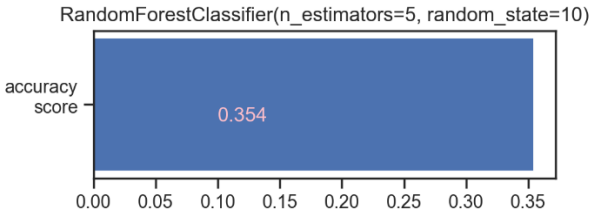


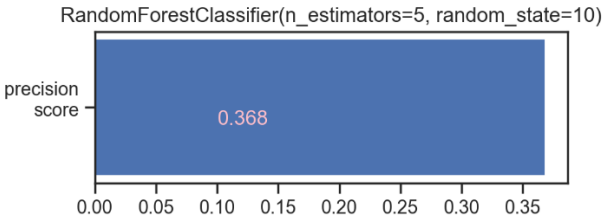


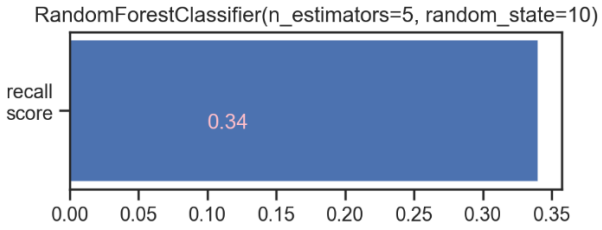


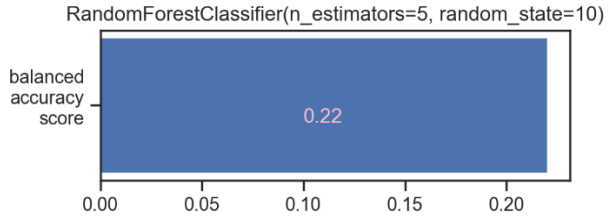


4.5.5 RandomForestClassifier

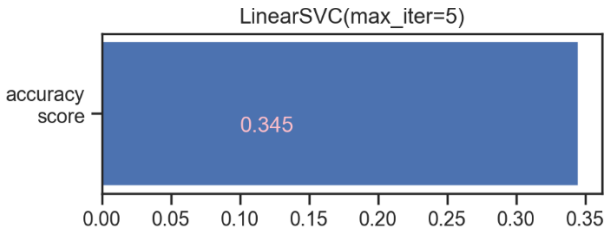


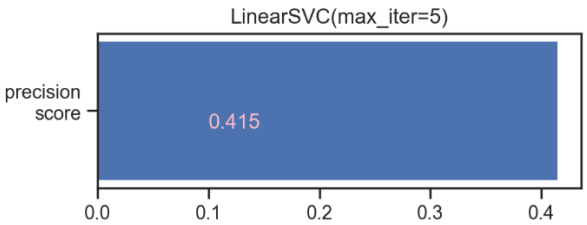


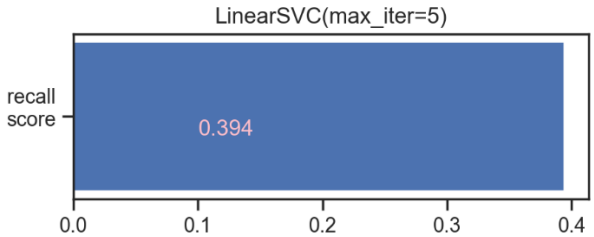


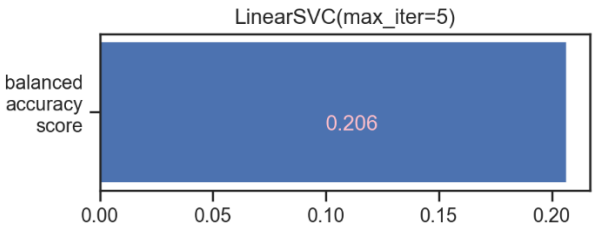


4.5.6 LinearSVC

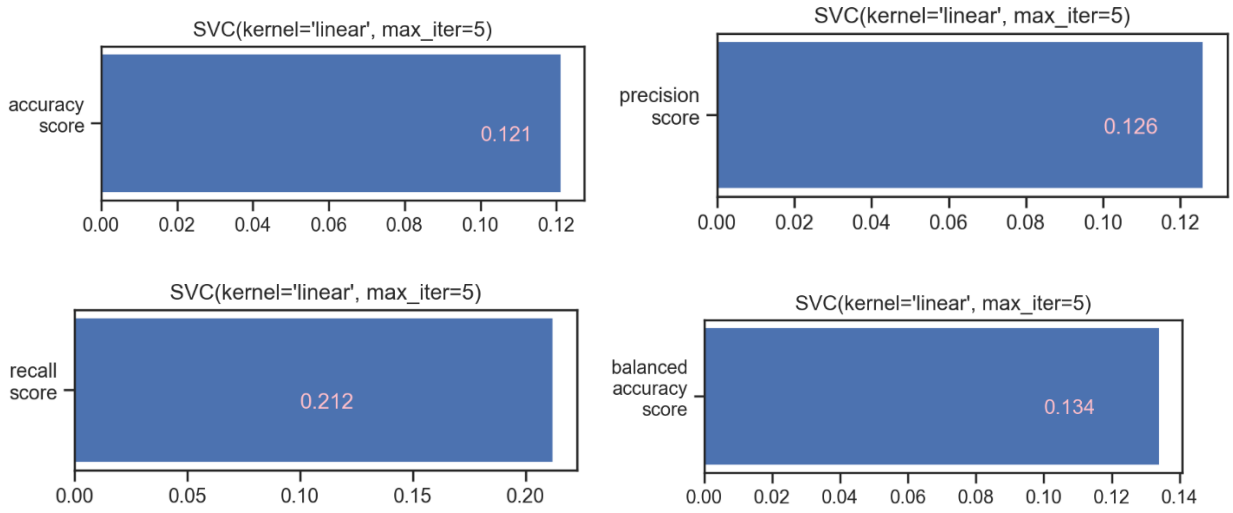








4.5.7 SVC



5 Основная часть

5.1 Описание данных

В качестве основного датафрейма выбрана база данных, содержащая информацию о рейтингах блюд, оцененных посетителями, основываясь на ингредиентах, содержащихся в них.

5.2 Решаемая задача

Основной задачей выбранных данных является классификация блюд по рейтингу, основываясь на ингредиентах, содержащихся в них.

5.3 Выбранные модели для обучения

Мной были выбраны следующие модели:

8) AdaBoostClassifier,
9) BaggingClassifier,
10) ExtraTreesClassifier,
11) GradientBoostingClassifier,
12) RandomForestClassifier,
13) LinearSVC,
14) SVC

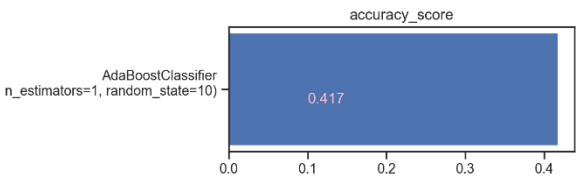
5.4 Выбранные метрики для оценки качества

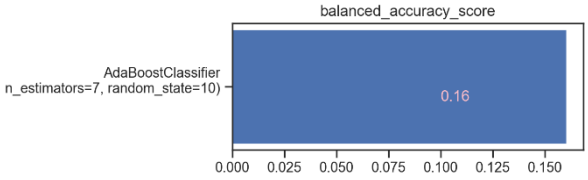
Мной были выбраны следующие модели:

5) accuracy_score,
6) precision_score,
7) recall_score,
8) balanced_accuracy_score

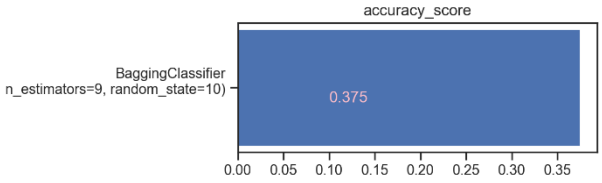
5.5 Построение графиков для выбранных моделей с подбором гипер-паратров

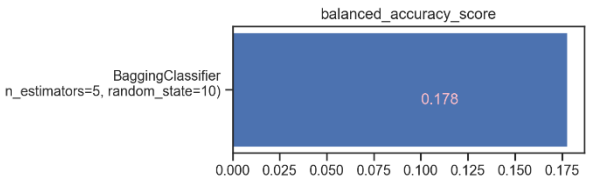
5.5.1 AdaBoostClassifier



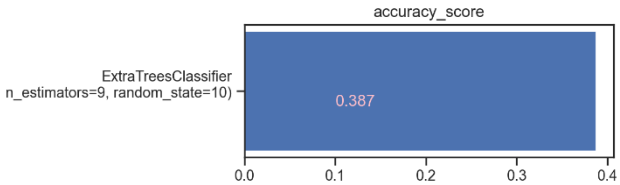


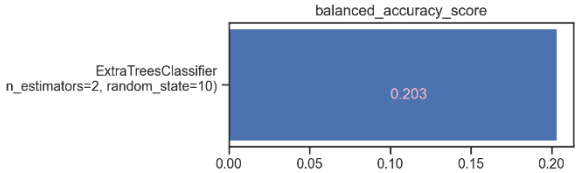
5.5.2 BaggingClassifier



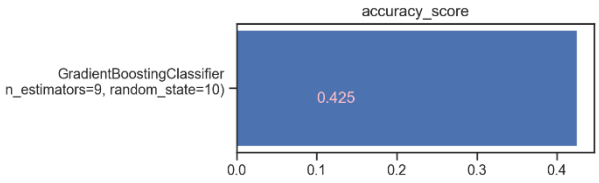


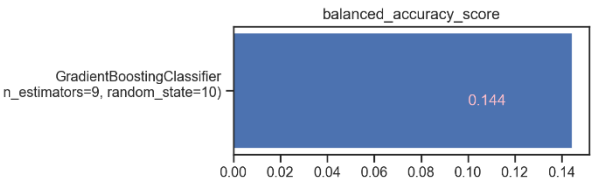
5.5.3 ExtraTreesClassifier



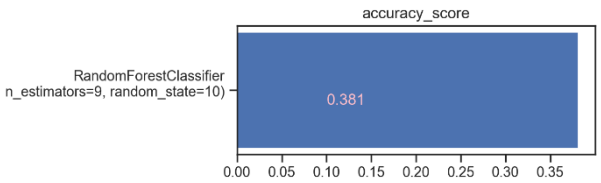


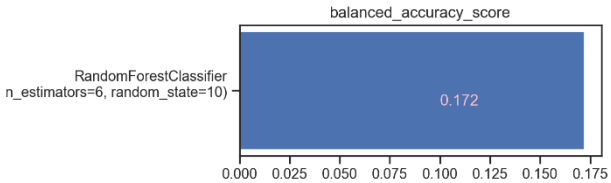
5.5.4 GradientBoostingClassifier



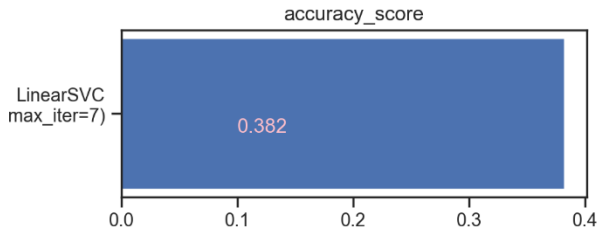


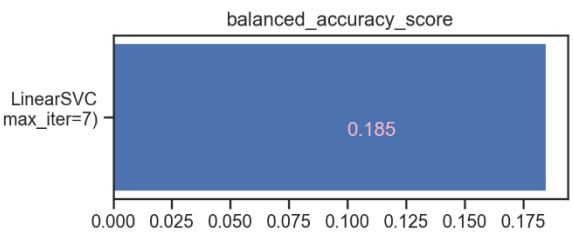
5.5.5 RandomForestClassifier



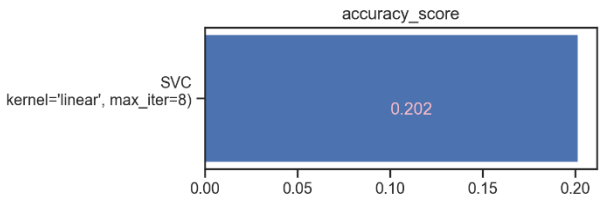


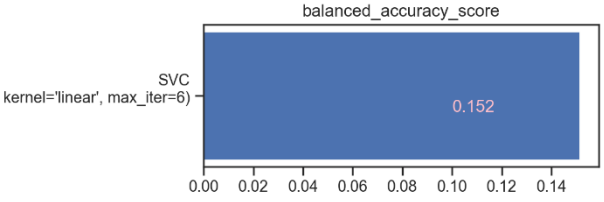
5.5.6 LinearSVC





5.5.7 SVC





6 AutoML

Так как основная база данных слишком большая, поэтому мной было принято решение взять еще одну базу данных, но поменьше. И для нее применить AutoML библиотеку TPOT

6.1 Результат обучения

```
Generation 1 - Current best internal CV score: 0.9714285714285715

Generation 2 - Current best internal CV score: 0.980952380952381

Best pipeline: GaussianNB(VarianceThreshold(RBFSampler(input_matrix, gamma=0.65), threshold=0.005))
```

7 Вывод

В этом курсовом проекте было много исследований, среди которых модель ExtraTreesClassifier оказалась лучшей в отличие от остальных.

8 Использованные источники

1) sklearn
2) pandas
3) numpy
4) kaggle

9 Приложение

9.1 Исходный код

Исходный код курсовой работы представлен в прилагаемых файлах:

1) CourseWork.pdf
2) CourseWork.ipynb
3) CourseWork.html
4) base.py
5) funcs.py

9.2 Web-приложение

1) web.py
2) web · Streamlit - AutoML.pdf
3) web · Streamlit - Description.pdf
4) web · Streamlit - Main.pdf