|  |  |
| --- | --- |
|  | **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 ***\_\_\_\_\_\_\_\_\_***

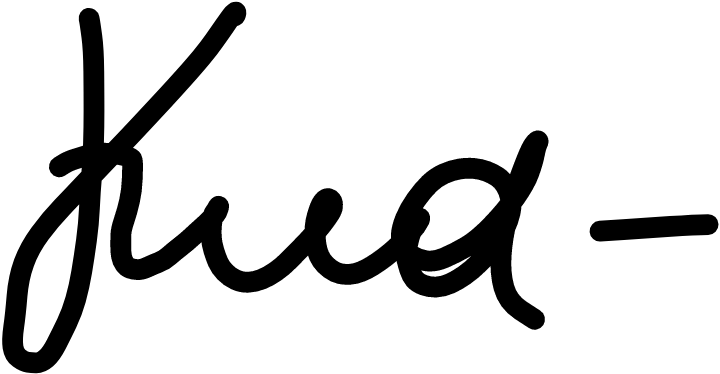
***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***



Студент \_\_\_\_\_ИУ5ц-83Б\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_Костников И.А.\_\_\_\_\_

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель курсовой работы **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_Гапанюк Ю.Е.\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_Гапанюк Ю.Е.\_\_\_\_

(Подпись, дата) (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.

**ЗАДАНИЕ**

**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.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

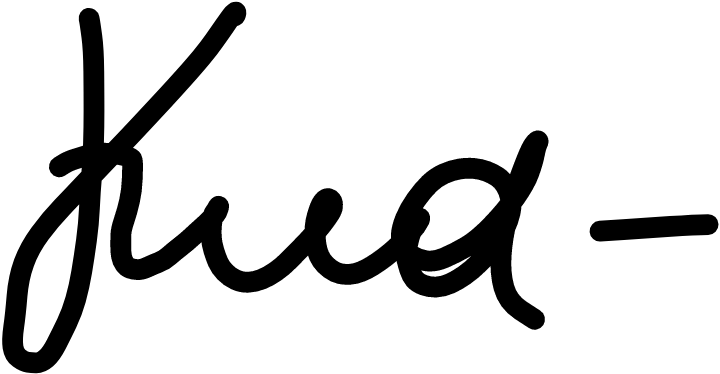
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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 6](#_Toc72608362)

[2 Цель работы 6](#_Toc72608363)

[3 Задание 6](#_Toc72608364)

[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. 6](#_Toc72608365)

[3.2 Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных. 6](#_Toc72608366)

[3.3 Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей. 6](#_Toc72608367)

[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. 6](#_Toc72608368)

[3.5 Выбор метрик для последующей оценки качества моделей. Необходимо выбрать не менее трех метрик и обосновать выбор. 6](#_Toc72608369)

[3.6 Выбор наиболее подходящих моделей для решения задачи классификации или регрессии. Необходимо использовать не менее пяти моделей, две из которых должны быть ансамблевыми. 6](#_Toc72608370)

[3.7 Формирование обучающей и тестовой выборок на основе исходного набора данных. 6](#_Toc72608371)

[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. 7](#_Toc72608372)

[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. 7](#_Toc72608373)

[3.10 Повторение пункта 8 для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей. 7](#_Toc72608374)

[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. 7](#_Toc72608375)

[3.12 Формирование web-приложения 7](#_Toc72608376)

[4 Основная часть 7](#_Toc72608377)

[4.1 Описание данных 7](#_Toc72608378)

[4.2 Решаемая задача 7](#_Toc72608379)

[4.3 Выбранные модели для обучения 7](#_Toc72608380)

[4.4 Выбранные метрики для оценки качества 7](#_Toc72608381)

[4.5 Построение графиков для выбранных моделей без подбора гипер-парамтров 8](#_Toc72608382)

[4.5.1 AdaBoostClassifier 8](#_Toc72608383)

[4.5.2 BaggingClassifier 8](#_Toc72608384)

[4.5.3 ExtraTreesClassifier 8](#_Toc72608385)

[4.5.4 GradientBoostingClassifier 9](#_Toc72608386)

[4.5.5 RandomForestClassifier 9](#_Toc72608387)

[4.5.6 LinearSVC 9](#_Toc72608388)

[4.5.7 SVC 10](#_Toc72608389)

[5 Основная часть 10](#_Toc72608390)

[5.1 Описание данных 10](#_Toc72608391)

[5.2 Решаемая задача 10](#_Toc72608392)

[5.3 Выбранные модели для обучения 10](#_Toc72608393)

[5.4 Выбранные метрики для оценки качества 10](#_Toc72608394)

[5.5 Построение графиков для выбранных моделей с подбором гипер-парамтров 11](#_Toc72608395)

[5.5.1 AdaBoostClassifier 11](#_Toc72608396)

[5.5.2 BaggingClassifier 11](#_Toc72608397)

[5.5.3 ExtraTreesClassifier 11](#_Toc72608398)

[5.5.4 GradientBoostingClassifier 11](#_Toc72608399)

[5.5.5 RandomForestClassifier 11](#_Toc72608400)

[5.5.6 LinearSVC 11](#_Toc72608401)

[5.5.7 SVC 11](#_Toc72608402)

[6 AutoML 12](#_Toc72608403)

[6.1 Результат обучения 12](#_Toc72608404)

[7 Вывод 12](#_Toc72608405)

[8 Использованные источники 12](#_Toc72608406)

[9 Приложение 12](#_Toc72608407)

[9.1 Исходный код 12](#_Toc72608408)

[9.2 Web-приложение 12](#_Toc72608409)

# Введение

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

# 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.

# Задание

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

## 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.

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

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

## 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.

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

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

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

## 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.

## 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.

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

## 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.

## Формирование web-приложения

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

# Основная часть

## Описание данных

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

## Решаемая задача

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

## Выбранные модели для обучения

Мной были выбраны следующие модели:

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

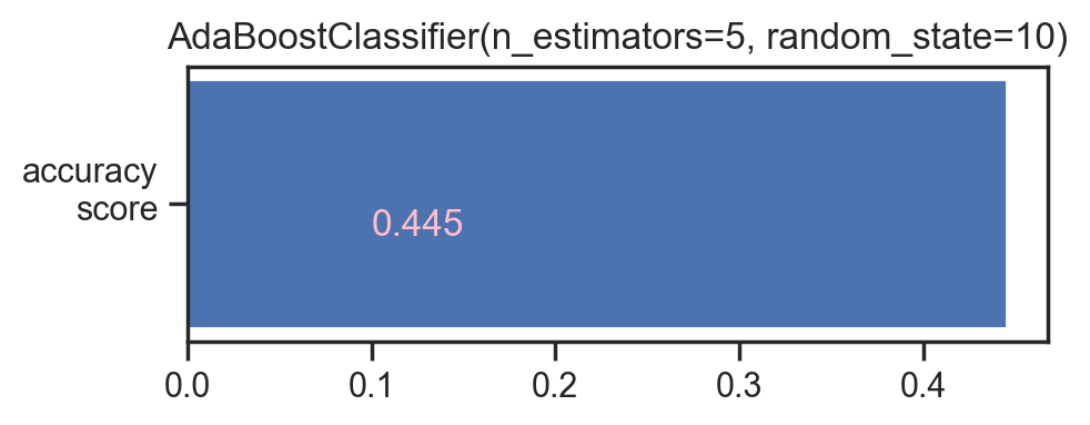
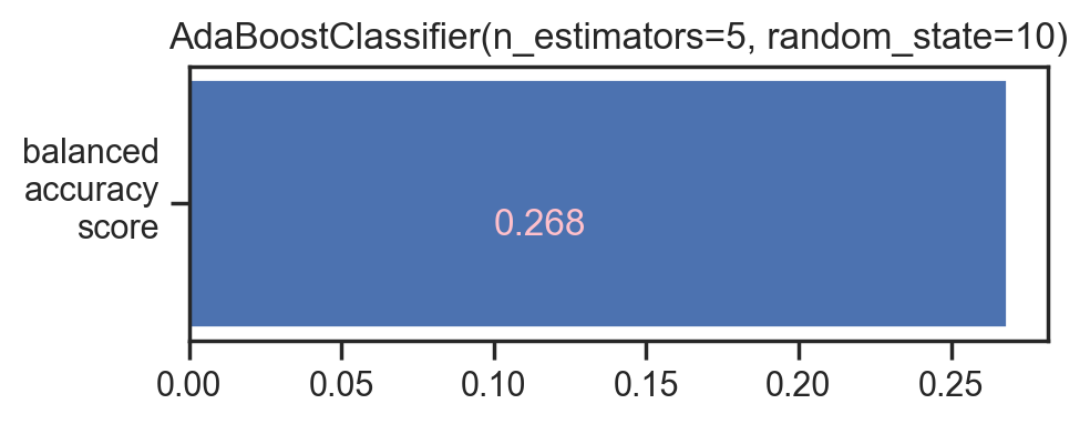
## Выбранные метрики для оценки качества

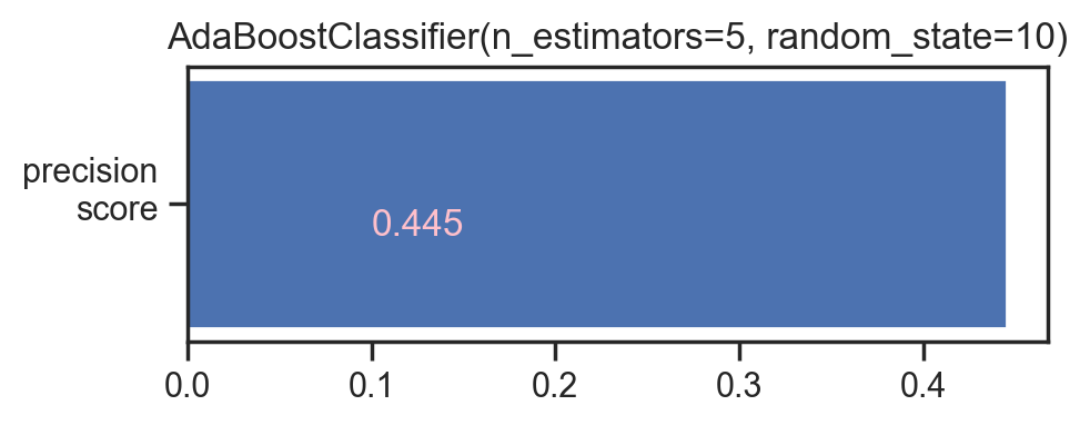
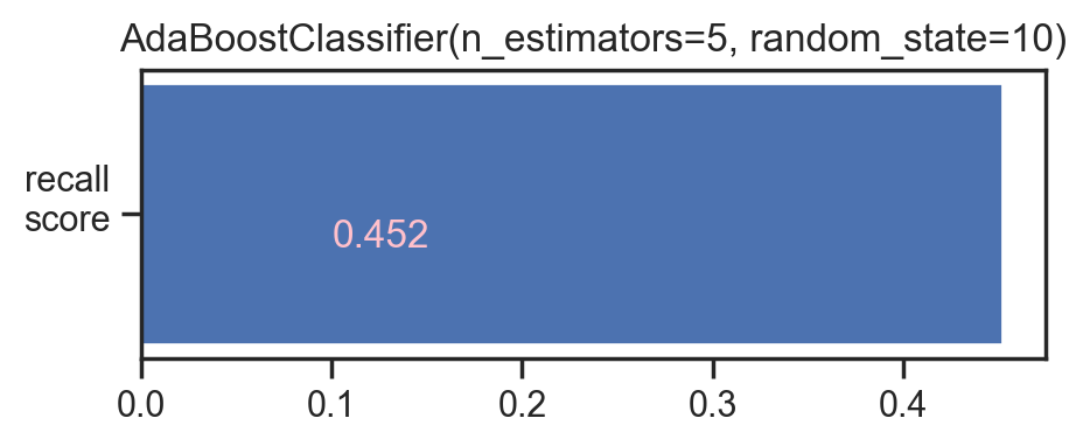
Мной были выбраны следующие модели:

1. accuracy\_score,
2. precision\_score,
3. recall\_score,
4. balanced\_accuracy\_score

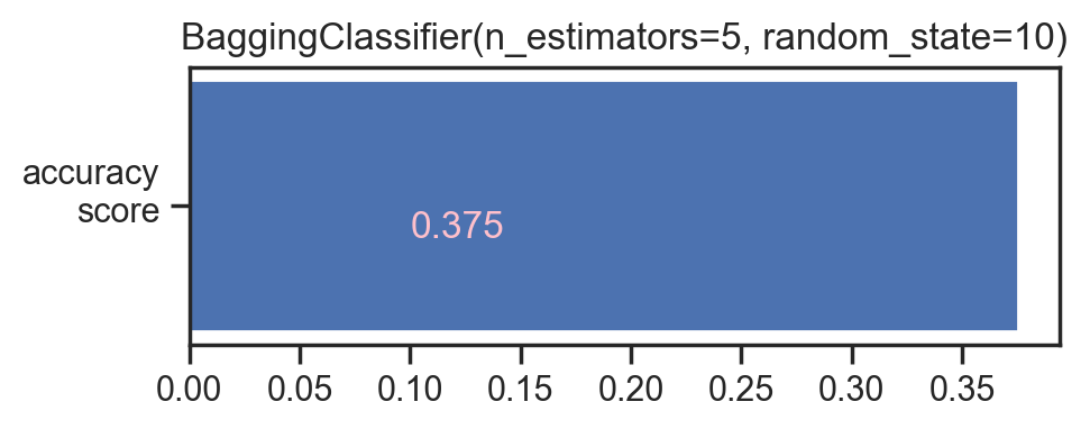
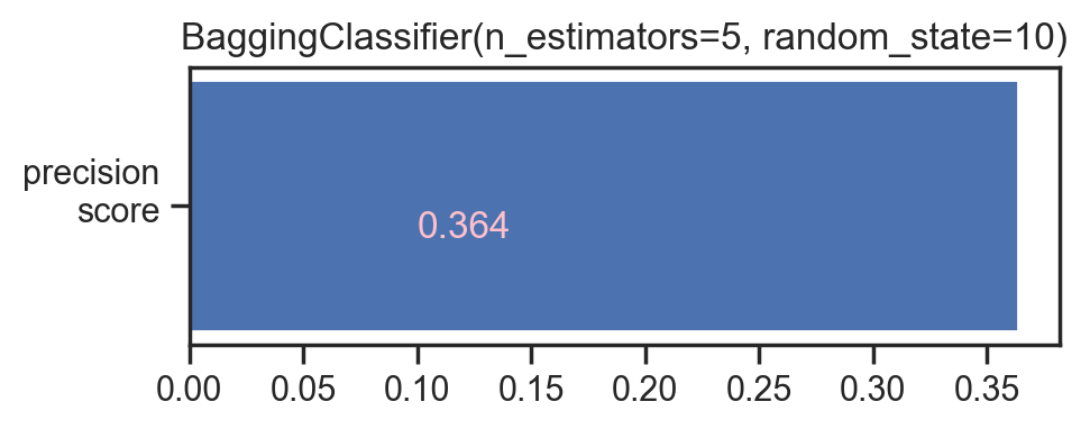
## Построение графиков для выбранных моделей без подбора гипер-парамтров

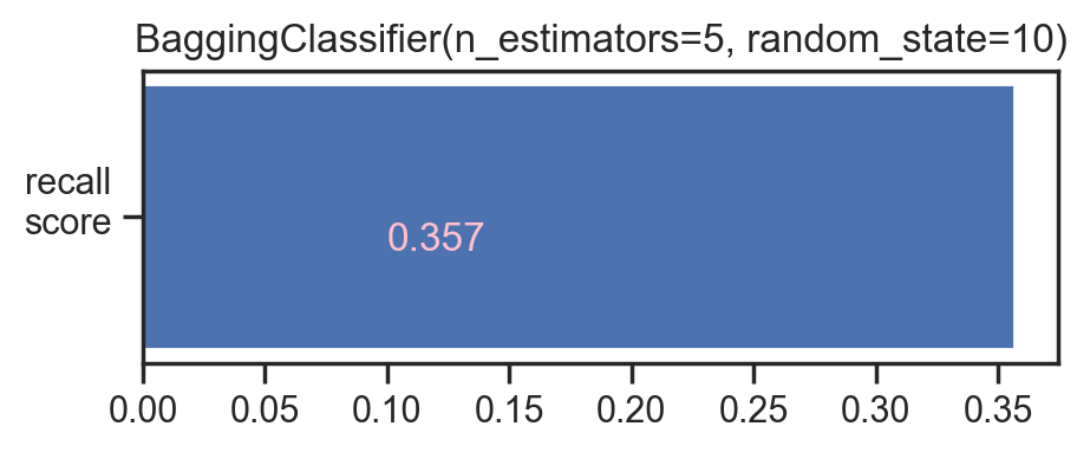
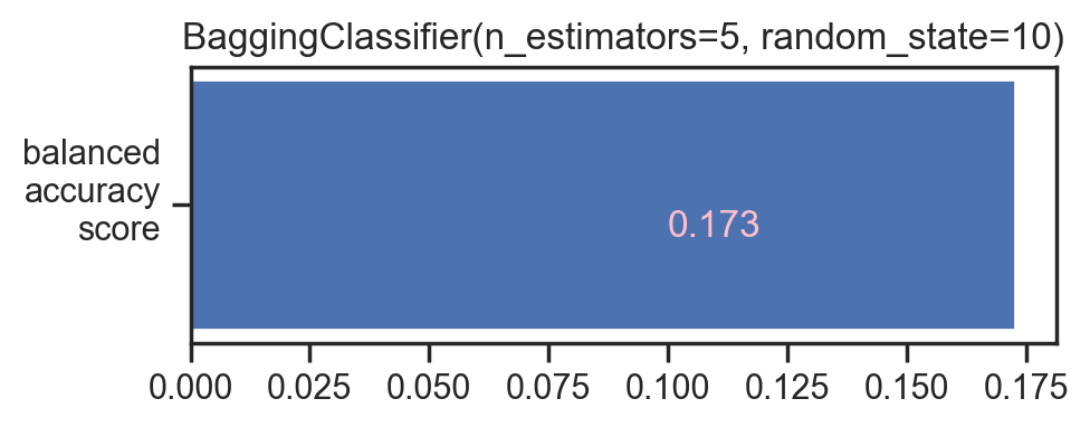
### AdaBoostClassifier

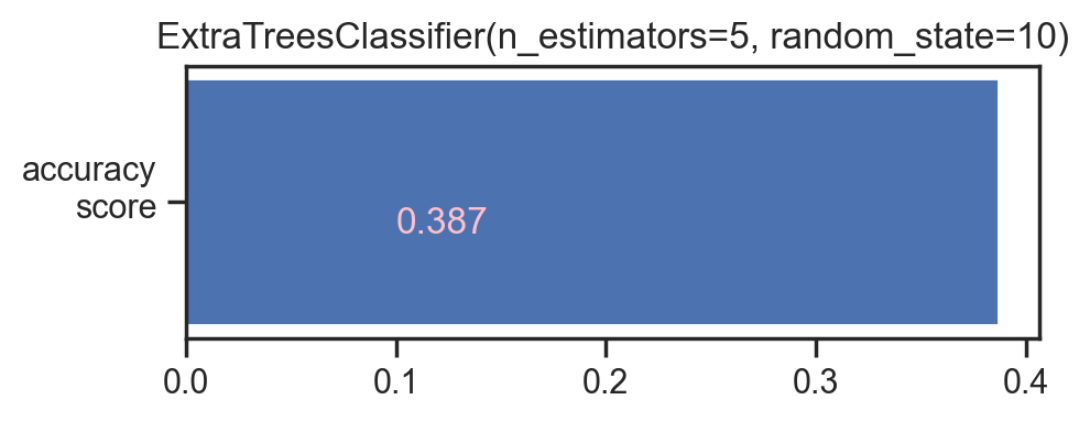
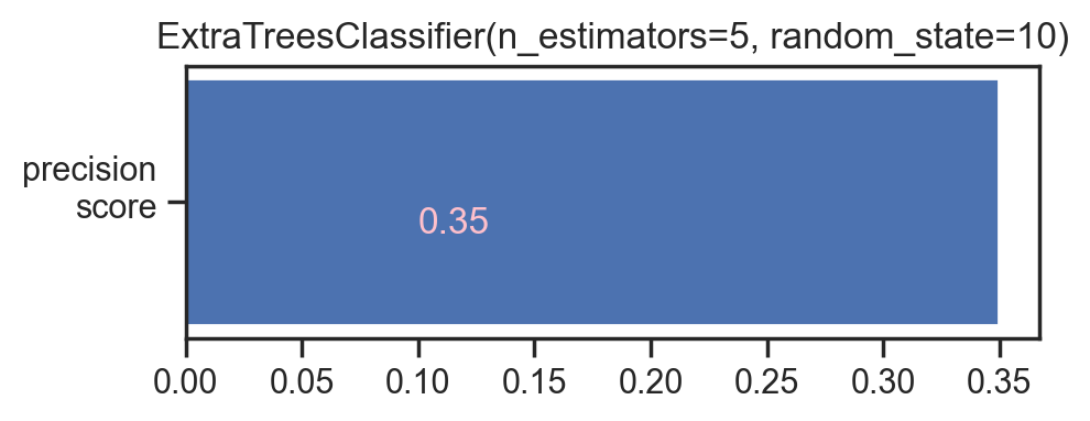
 

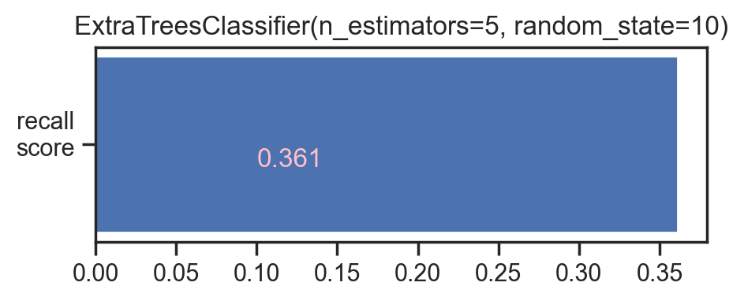
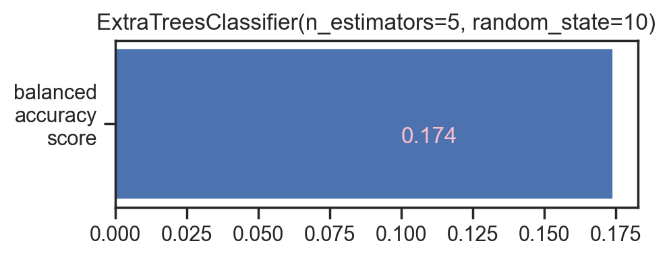
### BaggingClassifier

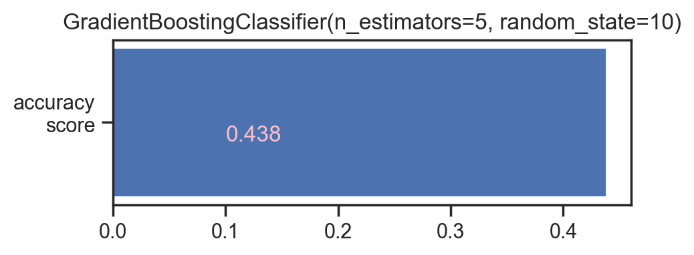
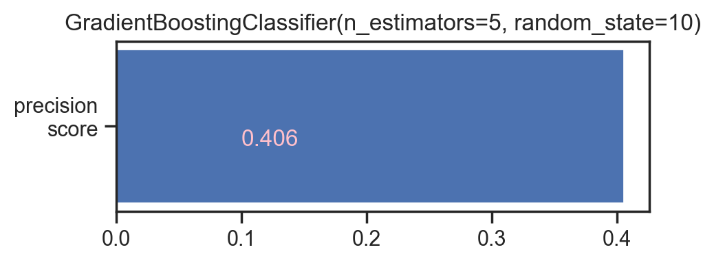
 

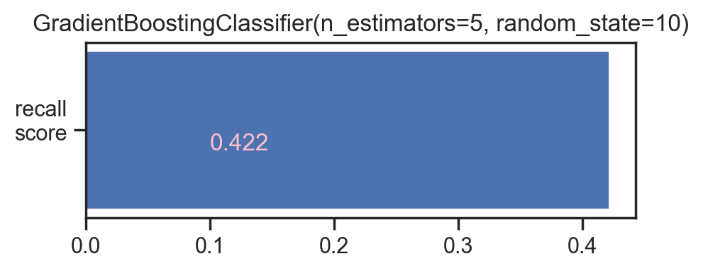
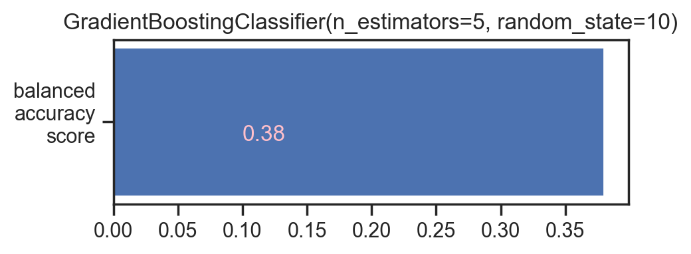
### ExtraTreesClassifier

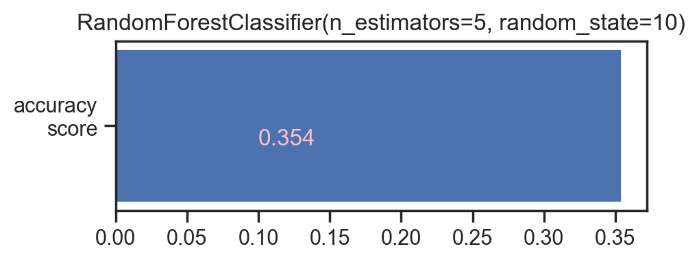
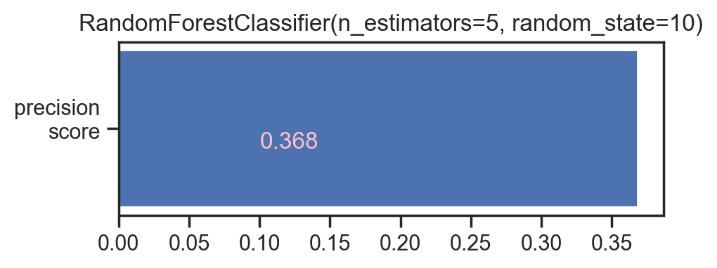
 

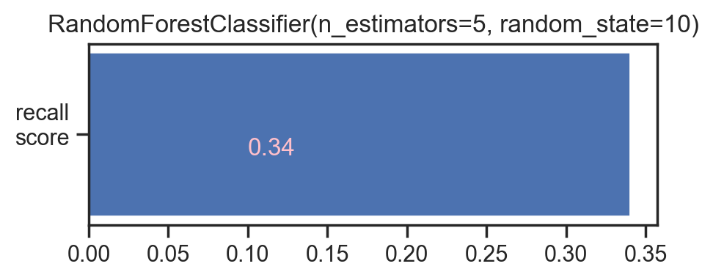
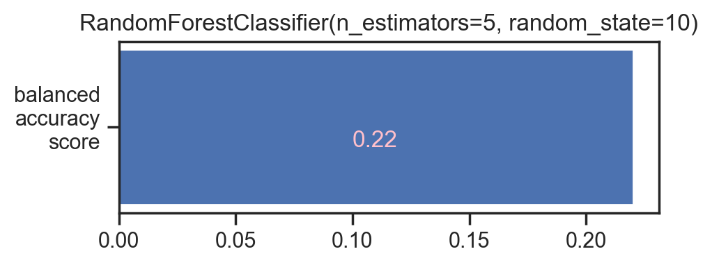
### GradientBoostingClassifier

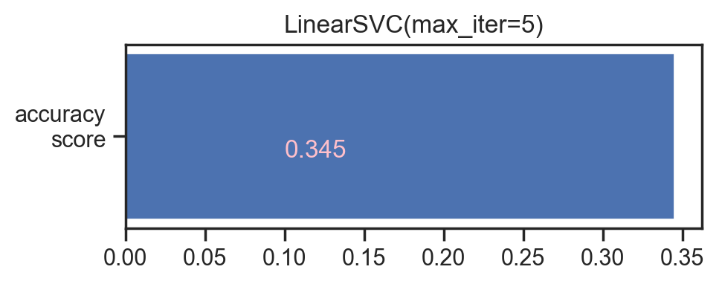
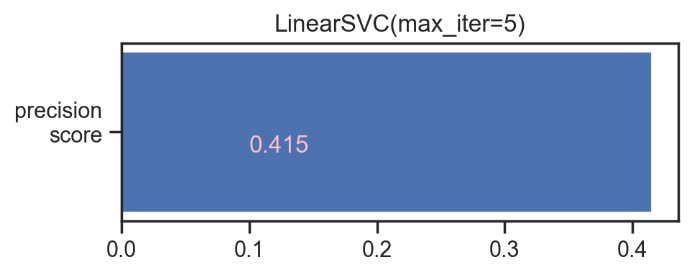
 

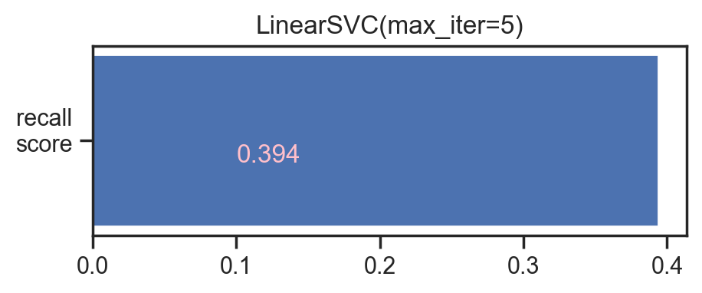
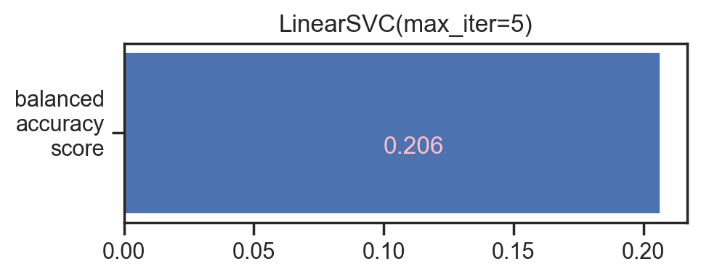
### RandomForestClassifier

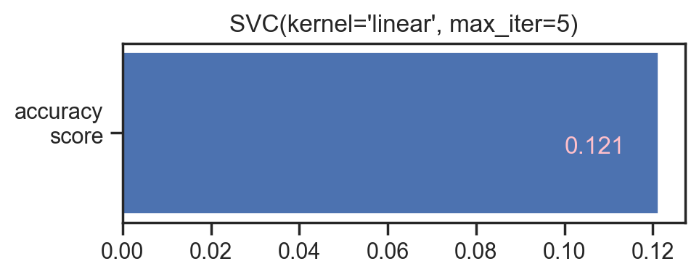
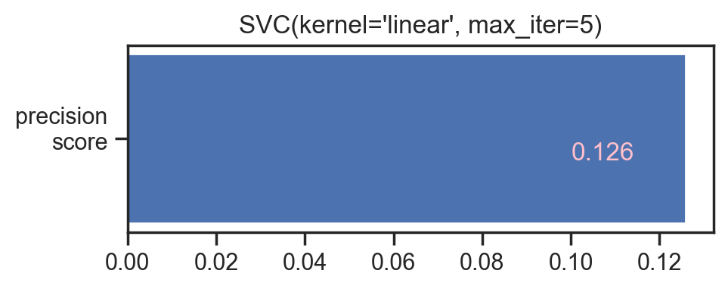
 

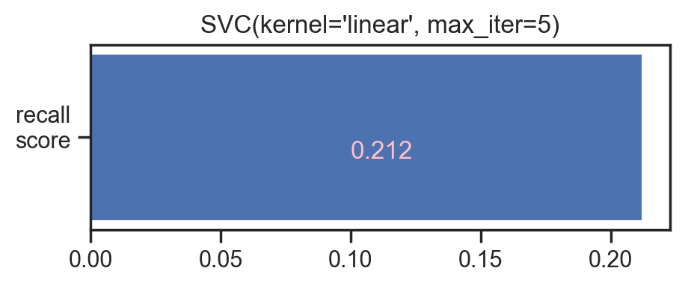
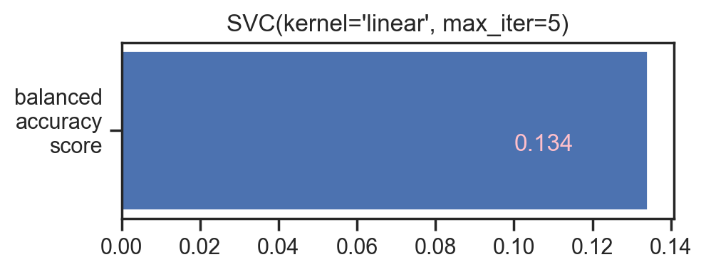
### LinearSVC

### SVC

# Основная часть

## Описание данных

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

## Решаемая задача

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

## Выбранные модели для обучения

Мной были выбраны следующие модели:

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

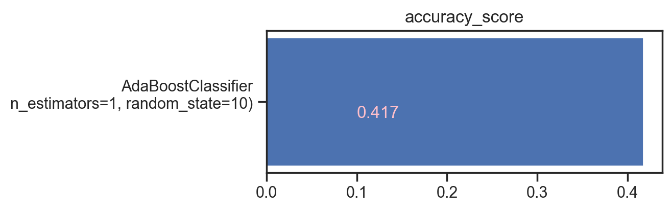
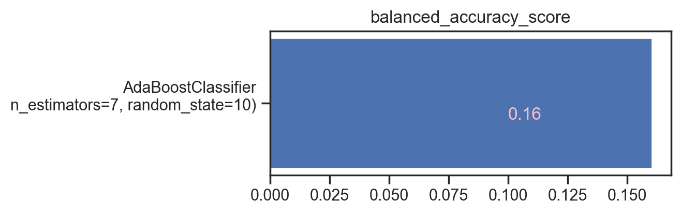
## Выбранные метрики для оценки качества

Мной были выбраны следующие модели:

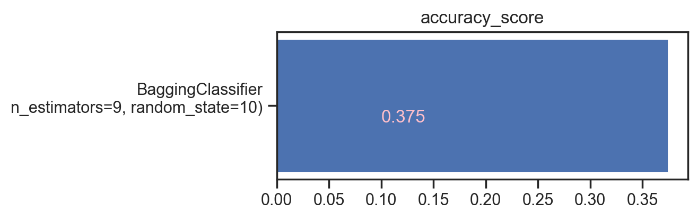
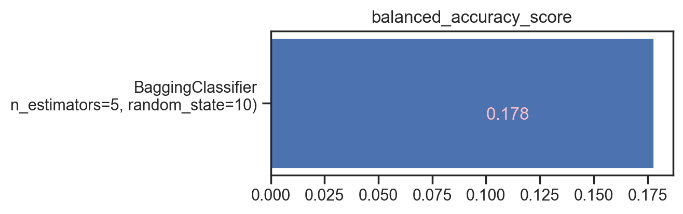
1. accuracy\_score,
2. precision\_score,
3. recall\_score,
4. balanced\_accuracy\_score

## Построение графиков для выбранных моделей с подбором гипер-парамтров

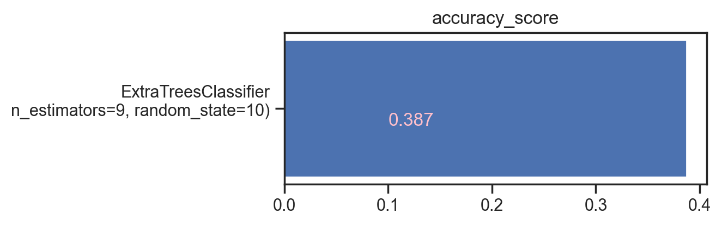
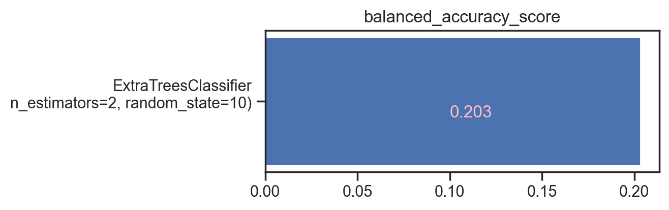
### AdaBoostClassifier

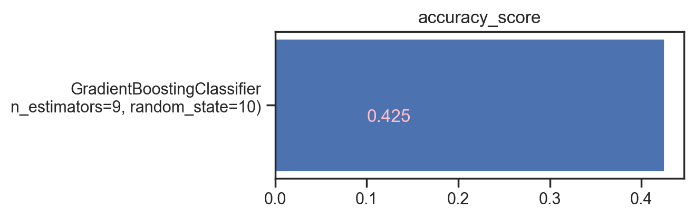
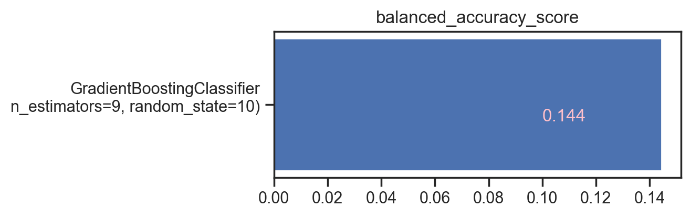
### BaggingClassifier

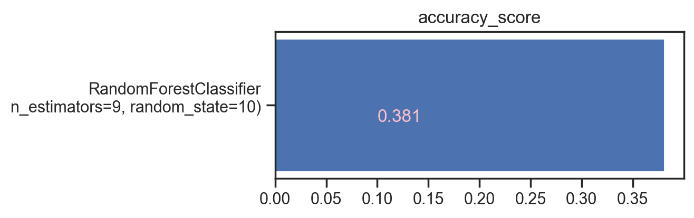
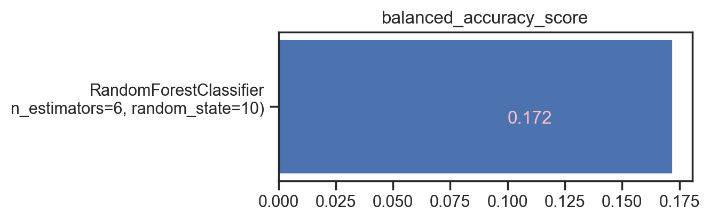
### ExtraTreesClassifier

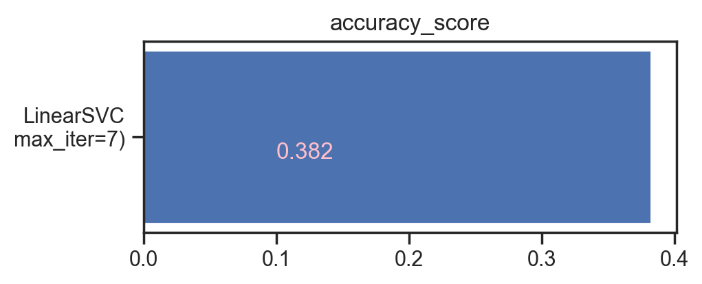
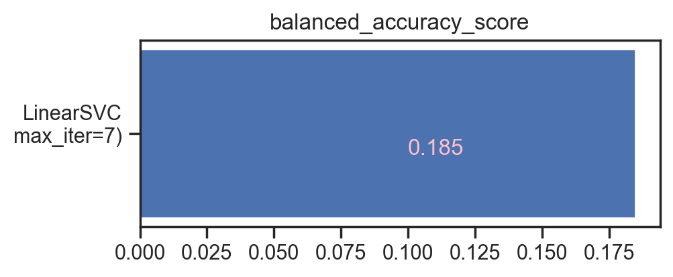
### GradientBoostingClassifier

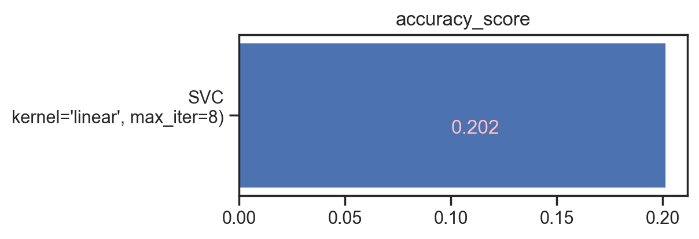
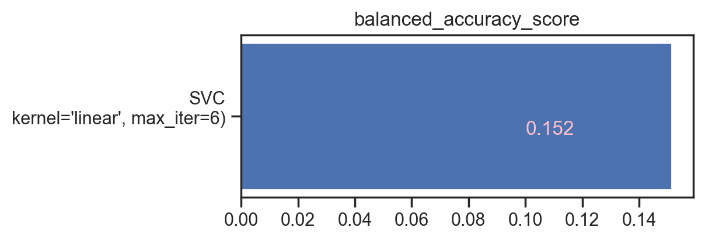
### RandomForestClassifier

### LinearSVC

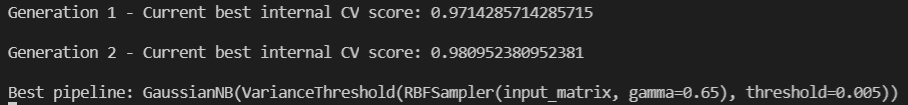
### SVC

# 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

## Результат обучения



# Вывод

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

# Использованные источники

1. [sklearn](https://www.sklearn.org)
2. [pandas](https://pandas.pydata.org)
3. [numpy](https://numpy.org)
4. [kaggle](https://www.kaggle.com/hugodarwood/epirecipes)

# Приложение

## 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

## Web-приложение

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