Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский Томский политехнический Университет»



Центр цифровых

образовательных технологий

01.03.02 «Прикладная математика и информатика»

Табличное представление данных

Лабораторная работа № 4

по дисциплине:

**Python для анализа данных**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Исполнитель:** |  | | | | |
| студент группы | 0В02 |  | Редько Д.А. |  |  |
|  |  |  |  |  |  |
| **Руководитель:** |  | | | | |
| преподаватель |  |  | Гоморов А.Л. |  |  |
|  |  |  |  |  |  |

Томск – 2023

**Цель работы:**

Применить алгоритмы деревьев решений и случайного леса для решения задачи классификации на реальном наборе данных, проанализировать различные техники оптимизации и выбрать подходящую метрику качества.

**Часть 1. Предобработка данных**

1. Импортировать необходимые библиотеки и загрузить данные
2. Изучить и проанализировать данные (определить пропущенные значения, основные статистики, типы данных и т.д.)
3. Создать пайплайн с использованием Pipeline и ColumnTransformer для обработки пропущенных значений и преобразования категориальных признаков (one-hot encoding, label encoding)
4. Разделить данные на обучающую и тестовую выборки

**Часть 2. Обучение базовых моделей**

1. Создать и обучить модель дерева решений с использованием Pipeline
2. Создать и обучить модель случайного леса с использованием Pipeline
3. Сравнить результаты моделей на тестовой выборке

**Часть 3. Исследование различных техник оптимизации**

1. Применить кросс-валидацию для оценки качества моделей
2. Исследовать важность признаков для каждой из моделей
3. Настроить гиперпараметры моделей с использованием GridSearchCV или RandomizedSearchCV
4. Сравнить результаты оптимизированных моделей

Часть 4. Выбор метрики качества

1. Оценить качество моделей с использованием различных метрик (Accuracy, Precision, Recall, F1-score, ROC-AUC)
2. Выбрать наиболее подходящую метрику для данной задачи и обосновать свой выбор.

**Ход работы:**

Часть 1. Предобработка данных

Импортируем все необходимые в лабораторной работе библиотеки и модули (рис 1).

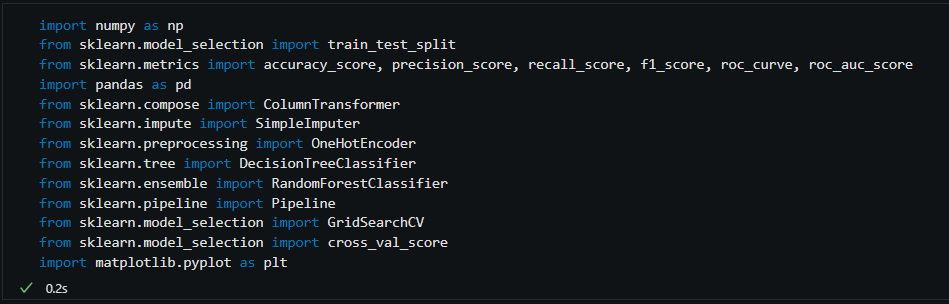


Рисунок 1. Импорт библиотек.

При помощи pandas подключим датасет, проверим, какие есть типы данных, получим их описание, проверим на наличие NaN значений (рис 2,3)

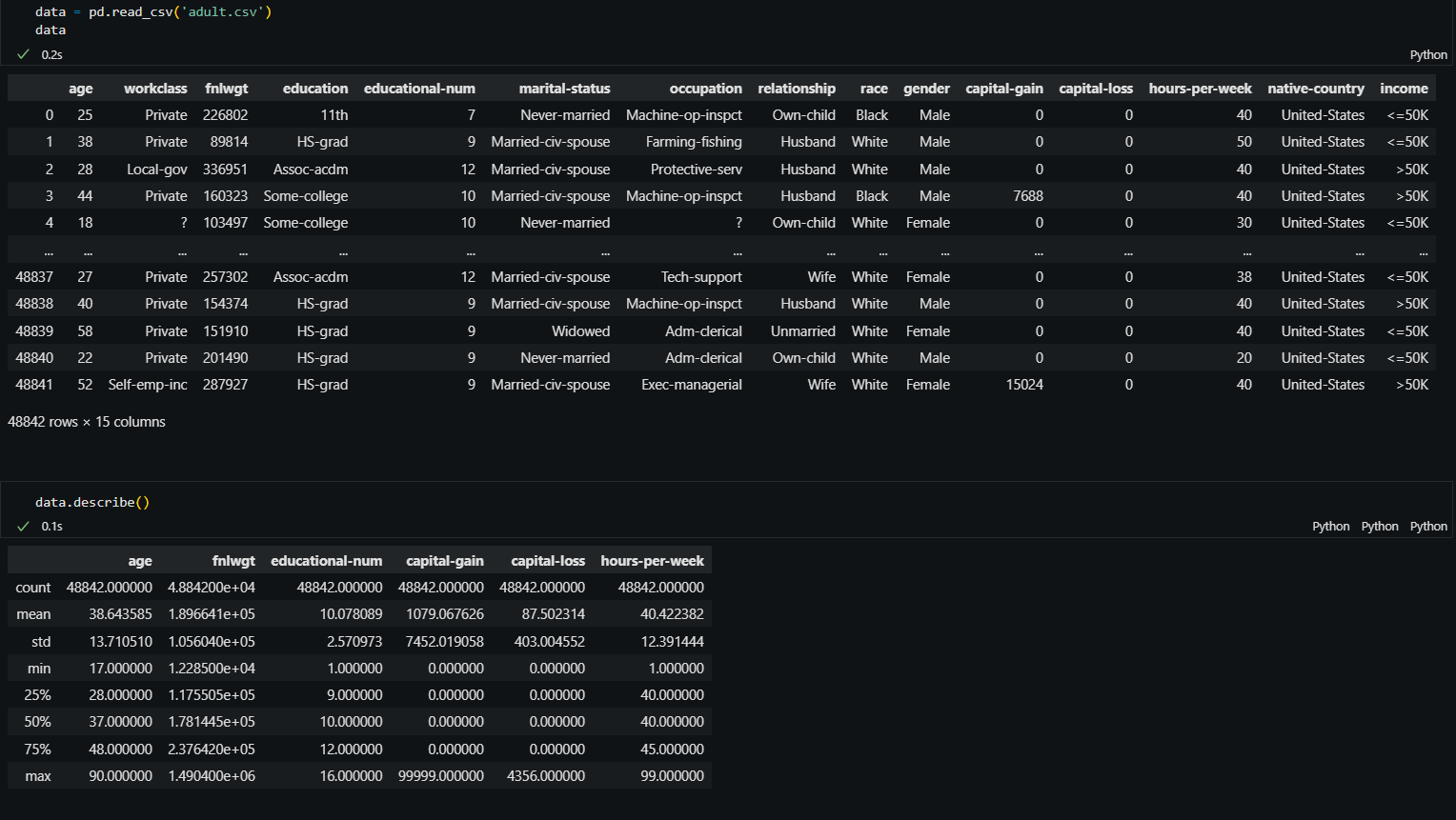


Рисунок . Подключение датасета и краткая информаця.

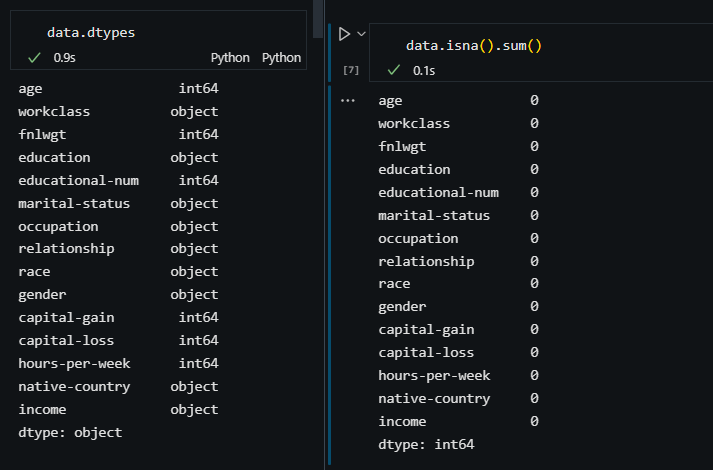


Рисунок . Получение типов признаков и проверка на пропуски.

Выделим категориальные и числовые признаки:

1. Возраст (age): числовой признак, представляющий возраст человека.
2. Работа (workclass): категориальный признак, описывающий тип занятости (например, государственная служба, самозанятость, частная компания и т.д.).
3. Финальный вес (fnlwgt): числовой признак, который является весом, присвоенным человеку на основе переписи населения. Этот признак используется для коррекции несбалансированности выборки.
4. Образование (education): категориальный признак, показывающий уровень образования человека (например, начальное, среднее, высшее и т.д.).
5. Образование-число (education-num): числовой признак, соответствующий количеству лет обучения.
6. Состояние брака (marital-status): категориальный признак, указывающий на семейное положение человека (например, женат/замужем, разведен(а), холост(а) и т.д.).
7. Род занятий (occupation): категориальный признак, описывающий тип выполняемой работы.
8. Семейное положение (relationship): категориальный признак, показывающий родственные связи человека в семье (например, муж, жена, ребенок и т.д.).
9. Раса (race): категориальный признак, указывающий на расовую принадлежность человека (например, белые, черные, азиаты и т.д.).
10. Пол (sex): категориальный признак, указывающий на пол человека.
11. Капитал (capital-gain): числовой признак, представляющий прирост капитала у человека.
12. Потери (capital-loss): числовой признак, представляющий потери капитала у человека.
13. Часы работы в неделю (hours-per-week): числовой признак, показывающий количество часов работы в неделю.
14. Происхождение (native-country): категориальный признак, указывающий на страну происхождения человека.
15. Доход (income): бинарный признак, представляющий уровень дохода человека (<=50K или >50K).

Разделим данные на обучающую и тестовую выборки (рис. 4).Начало формы

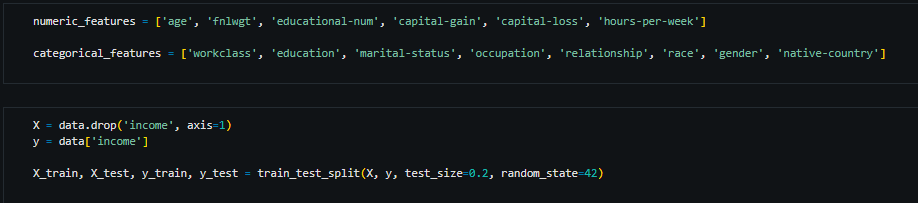
Конец формы

Рисунок 4. Категориальные и числовые признаки, разделение на обучающую и тестовую выборки.

**Часть 2. Обучение базовых моделей**

Создадим Pipline для DecisionTreeClassifier (рис. 5).

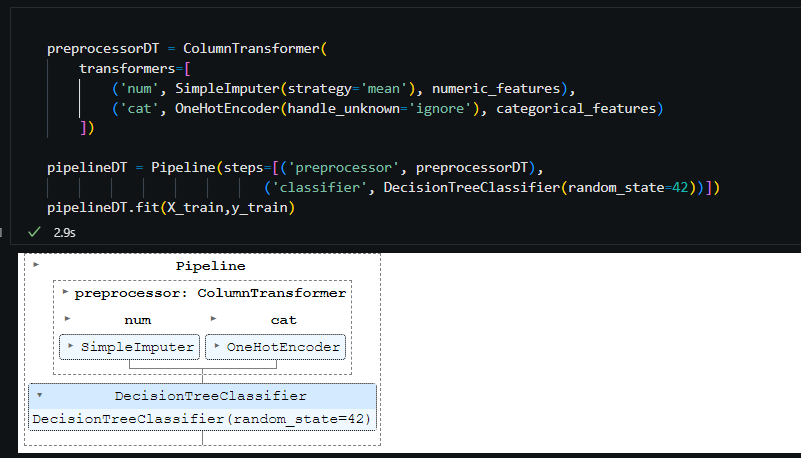


Рисунок 5. Pipeline для DecisionTreeClassifie

Создадим Pipline для RandomForestClassifier (рис. 6).

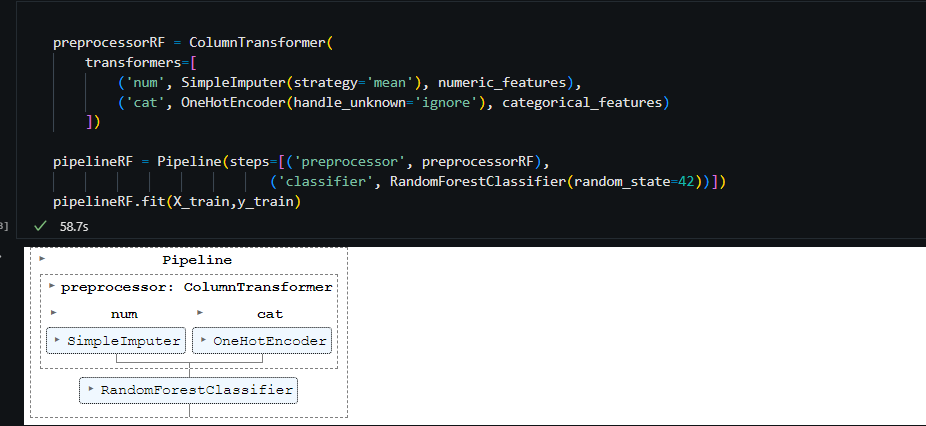


Рисунок 6. Pipeline для RandomForestClassifier

Сравним результаты на тестовой выборке (рис. 7).

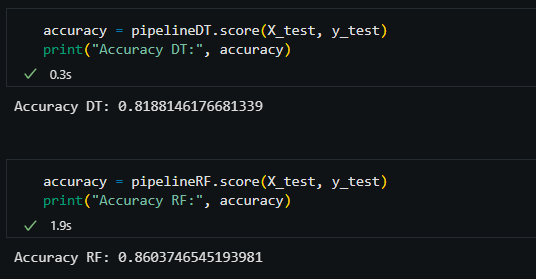


Рисунок 7. Сравнение результатов

**Часть 3. Исследование различных техник оптимизации**

Применим кросс-валидацию для оценки качества моделей (рис. 8).

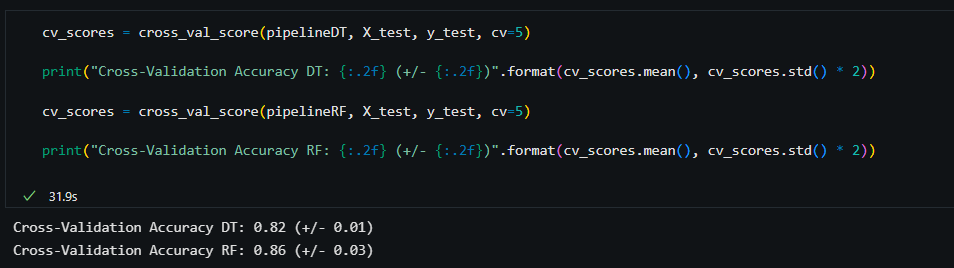


Рисунок 8. Cross-Validation accuracy.

Определим значимость признаков (рис 9, 10)

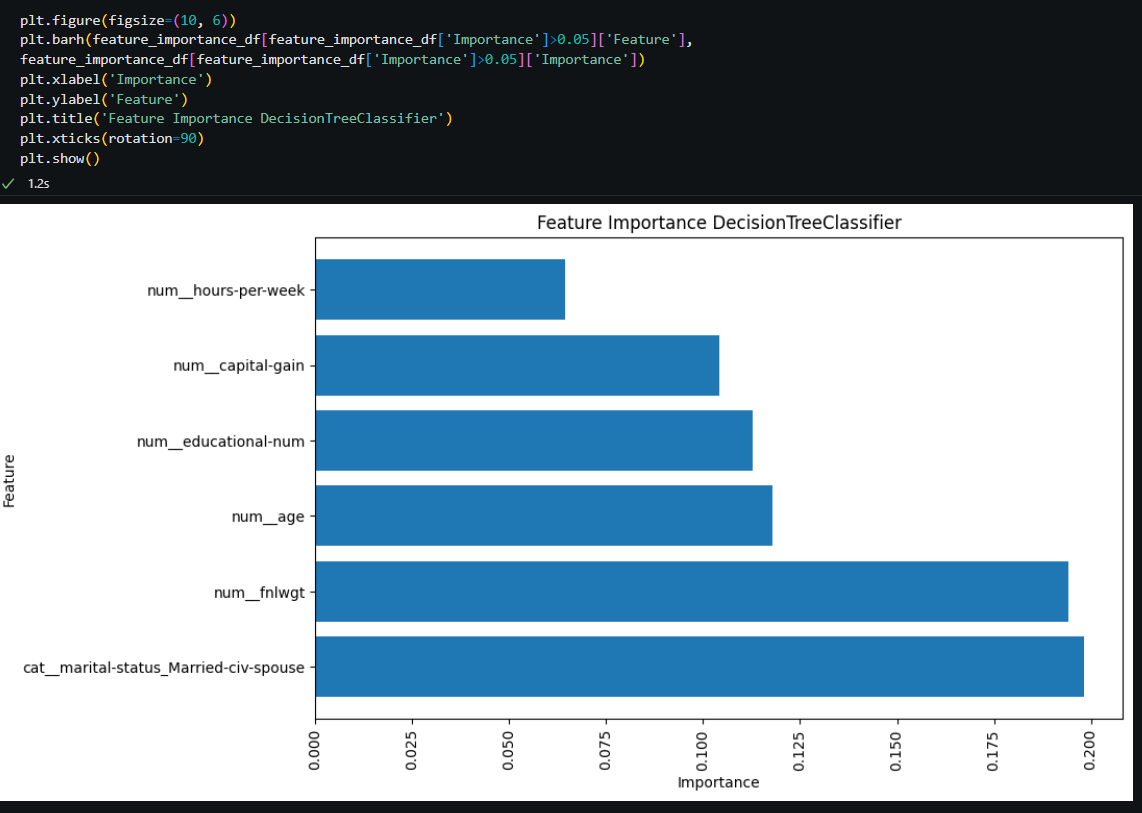


Рисунок 9. Значимость признаков DecisionTreeClassifier

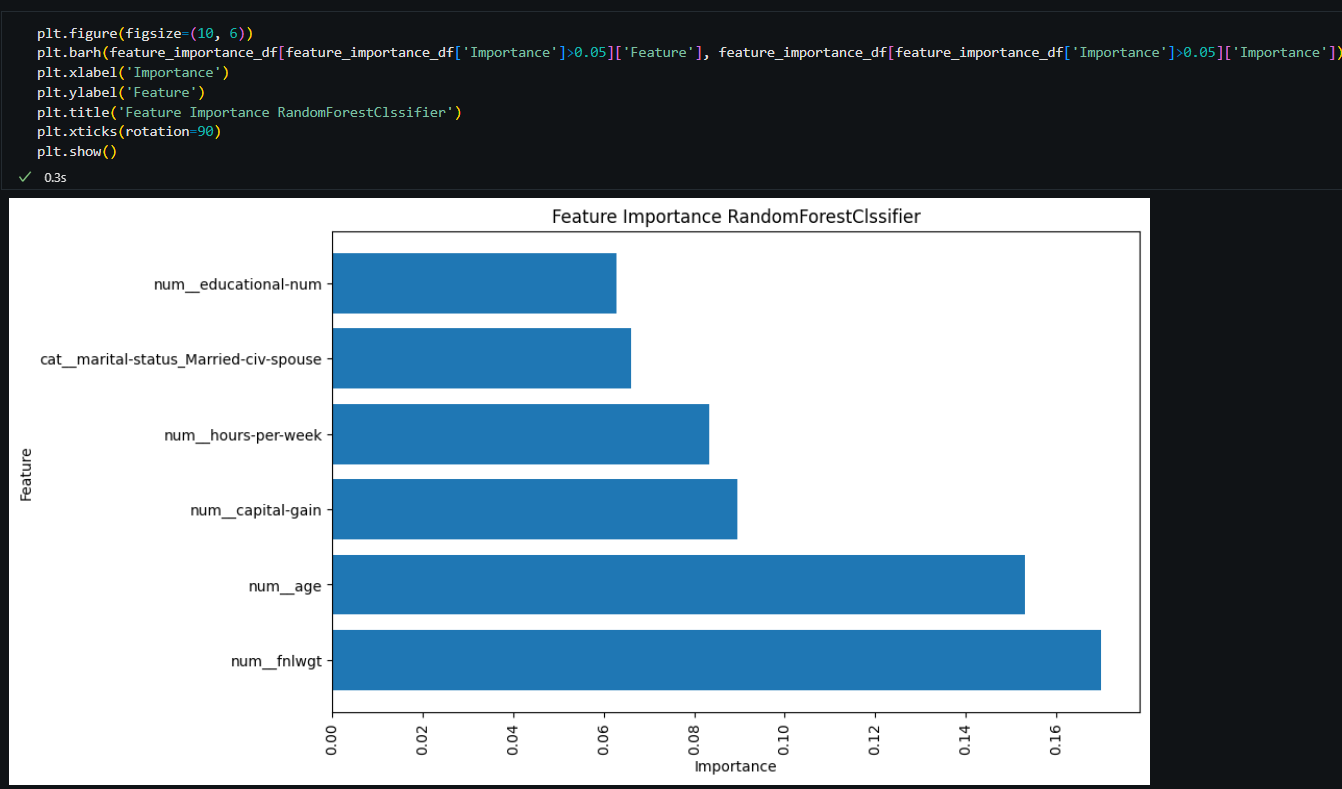


Рисунок . Значимость признаков RandomForestClassifier

Подберем наилучшие параметры при помощи GridSearchCV (рис 11,12)



Рисунок 11. Лучшие параметры для Pipeline DecisionTreeClassifier



Рисунок 12. Лучшие параметры для Pipeline RandomForestClassifier

Сравним результаты на тестовой выборке для лучших параметров (рис. 13):

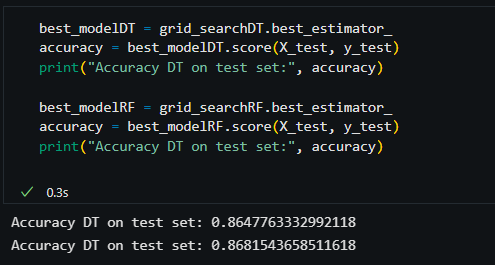


Рисунок 13. Наилучшие параметры

Видим, что значения DecisionTreeClassifier улучшились примерно на 0.05, при этом RandomForestClassifier результаты практически не изменились.

Часть 4. Выбор метрики качества

Определим основные метрики (рис 14, 15)

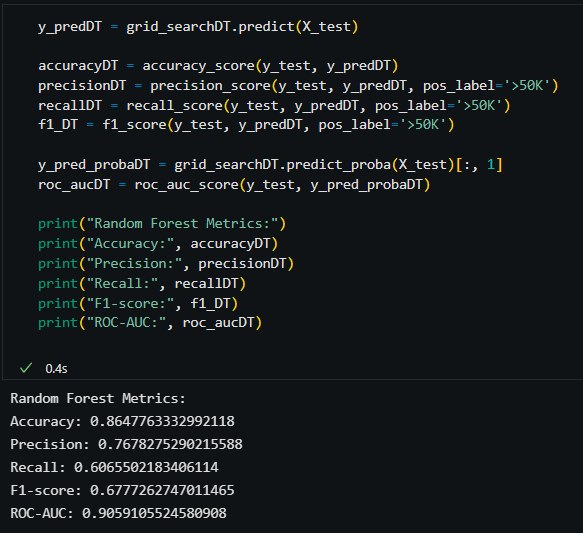


Рисунок 14. Метрики для DecisionTreeClassifier

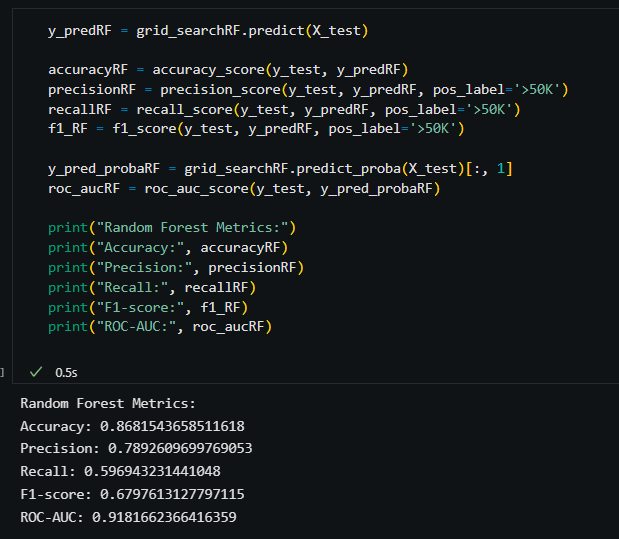


Рисунок 15. Метрики для RandomForestClassifier

Отобразим метрики графически в сравнении (рис. 16):

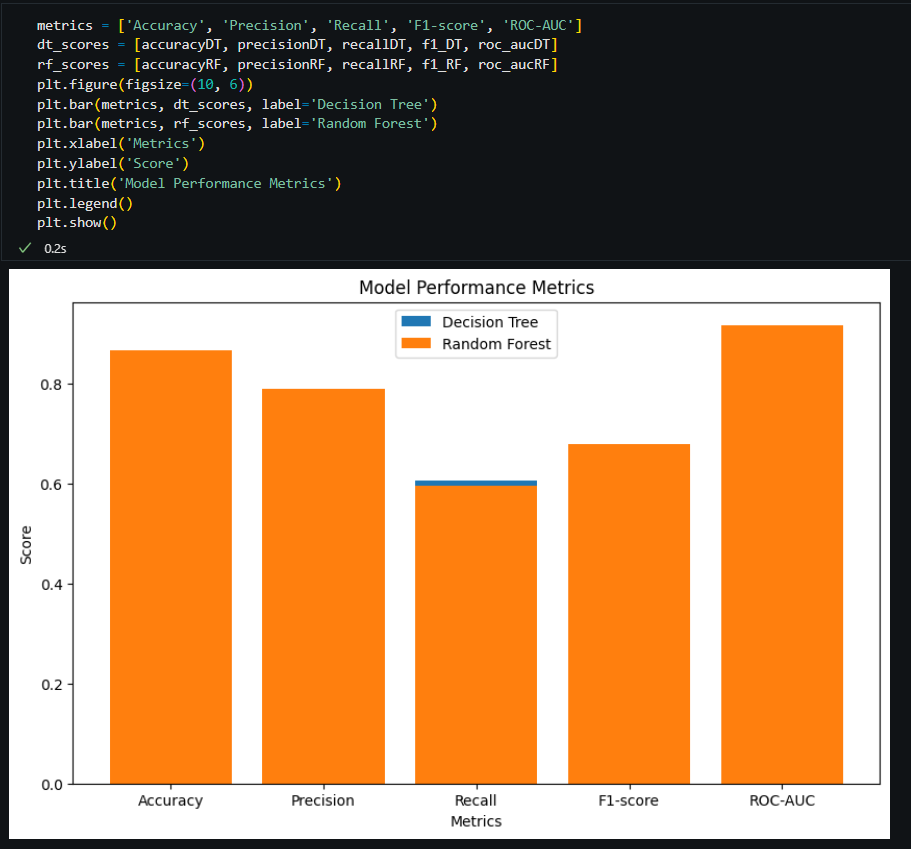
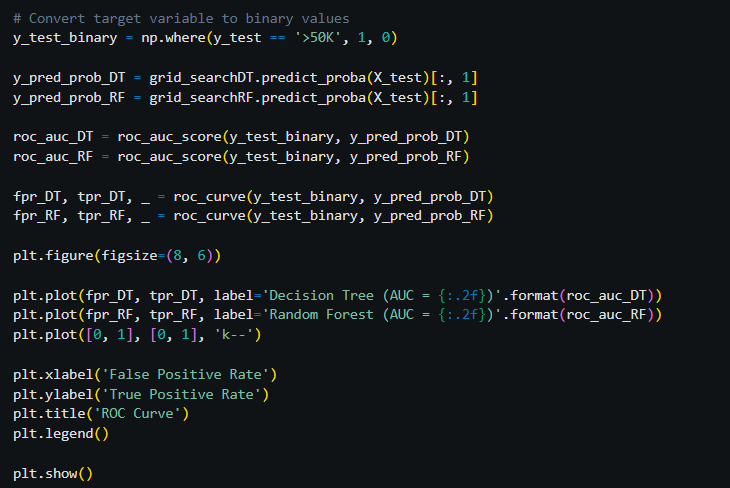


Рисунок 16. Сравнение метрик графически

Отобразим графически ROC-AUC кривую (рис. 17):



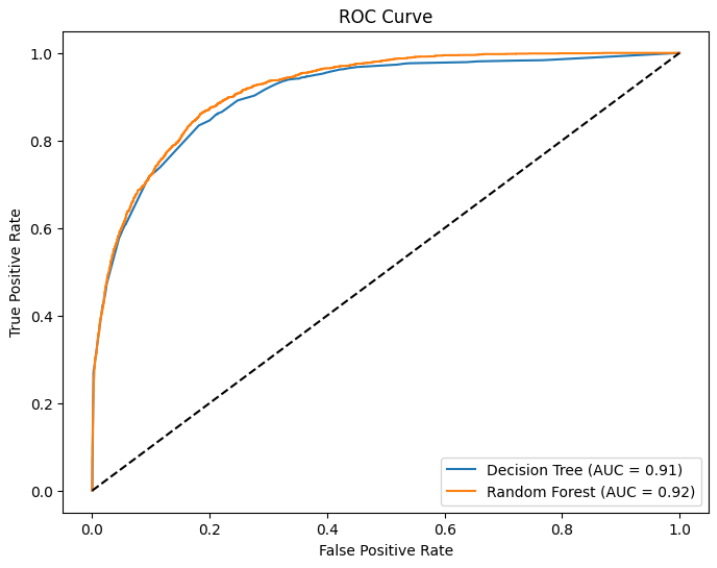


Рисунок 17. ROC-AUC кривая

Для выбора наиболее подходящей метрики для Adult Dataset можно рассмотреть характеристики и особенности задачи предсказания дохода людей. В данной задаче классификации, где необходимо предсказать, превышает ли доход человека порог в 50 тысяч долларов в год, все рассмотренные метрики могут быть полезны так как нет конкретной приоритезации( ошибка играет значимую роль):

1. Accuracy (точность): показывает, какая часть предсказаний была верной. Она может быть полезна для общей оценки качества модели, особенно если классы в данных сбалансированы.
2. Precision (точность): показывает, какая доля положительных предсказаний была верной. В данной задаче это может быть важно, если нам важно минимизировать ложноположительные результаты, то есть предсказывать, что доход превышает 50 тысяч долларов, когда это на самом деле не так.
3. Recall (полнота): показывает, какая доля истинных положительных результатов была найдена моделью. В данной задаче это может быть важно, если нам важно минимизировать ложноотрицательные результаты, то есть предсказывать, что доход не превышает 50 тысяч долларов, когда это на самом деле превышает.
4. F1-score: комбинирует точность и полноту в одну метрику. Он может быть полезен, когда нам нужно достичь баланса между точностью и полнотой.
5. ROC-AUC: площадь под кривой ROC (Receiver Operating Characteristic) показывает, насколько хорошо модель разделяет классы. Она может быть полезна, особенно если нам важно иметь хорошую способность модели разделять положительные и отрицательные примеры.

Выбор наиболее подходящей метрики зависит от конкретных целей и требований к модели. Например, если важно избегать ложноположительных результатов (предсказание превышения дохода, когда он на самом деле не превышает), то можно сосредоточиться на метрике Precision. Если же важно минимизировать ложноотрицательные результаты (предсказание непревышения дохода, когда он на самом деле превышает), то можно сосредоточиться на метрике Recall.Также можно оценивать модель по нескольким метрикам, чтобы получить более полное представление о ее качестве и сбалансированности предсказаний.

**Вывод:**

В данной работе применили методы предобработки данных для преобразования категориальных данных. Применили Pipeline на примере RandomForestClassifier и TreeDecisionClassifier. Скомбинировали Pipeline с GridSearchCV, определили значимость признаков и посчитали метрики для GridSearchCV моделей.