МИНОБРНАУКИ РОССИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южный федеральный университет»
Институт высоких технологий и пьезотехники

Отчёт по проекту по дисциплине “Большие данные”

Выполнили студенты 3 курса 2_ВТ-09.03.03.01-о_ группы:

_____ Ивахненко И.С.
подпись

_____ Симаков Д.В.
подпись

Проверил старший преподаватель:

_____ Турлюн А. С.
подпись

Ростов-на-Дону – 2024

Цель проекта: разработка модели машинного обучения для предсказания наиболее подходящей культуры по почвенным условиям в среде Pyspark. Определить, какая культура лучше всего подходит для выращивания в условиях конкретного типа почвы, основываясь на различных параметрах почвы и погодных условиях.

Задачи проекта:

- Сбор данных и их предварительная обработка
- Разработка моделей машинного обучения
- Проведение кросс-валидации и оценки модели
- Выбор и интерпретация лучшей модели
- Визуализация результатов

Актуальность проекта: в современном сельском хозяйстве эффективное использование земельных ресурсов является ключевым фактором для обеспечения продовольственной безопасности и устойчивого развития агропромышленного комплекса. Основные аспекты актуальности проекта:
1. повышение урожайности продукции
2. снижение риска и затрат

Гипотеза проекта: для себя мы выставили следующую гипотезу: возможно ли с помощью машинного обучения определить наиболее подходящую культуру исходя из данных по почвенным и климатическим условиям?

Методология

В данном проекте используется набор данных, содержащих почвенные и климатические характеристики, а также типы сельскохозяйственных культур. Набор данных представлен в формате CSV и включает следующие столбцы:

Азот: содержание азота в почве

Фосфор: содержание фосфора в почве

Калий: содержание калия в почве

Температура: средняя температура воздуха в цельсиях

Влажность: относительная влажность в процентах

pH_Значение: значение pH почвы

Осадки: количество осадков в мм

Перед тем как приступить к моделированию, необходимо выполнить несколько этапов предобработки данных:

- Очистка данных – проверка и замена пропущенных значений. Преобразование типов данных всех столбцов в тип 'float', кроме столбца с типом культуры.
- Преобразование признаков – векторизация и нормализация признаков.
- Преобразование категориальных данных – преобразование столбца с типами культур в числовые метки.

Для решения задачи предсказания наиболее подходящей культуры по почвенным условиям используя следующие модели:

1. Decision Tree (Дерево решений) – простой и интерпретируемый алгоритм, который строит дерево решений на основе входных признаков.
2. Random Forest(Случайный лес) – Ансамблевый метод, который строит множество деревьев решений и объединяет их результаты для повышения точности и устойчивости модели
3. Naive Bayes(Наивный байесовский классификатор) – простая и быстрая модель, которая основывается на применении теоремы Байеса и предполагает независимость признаков.
4. Logistic Regression (Логистичекая регрессия) – Линейный классификатор, который используется для оценки вероятности принадлежности к тому или иному классу.
5. K-Means (Метод K-средних) – метод кластеризации, который группирует данные в кластеры на основе схожести признаков

Для каждой модели проводится настйрока гиперпараметров с использованием

техники grid search и кросс-валидации.

Оценка моделей

Для оценки производительности моделей используется метрика accuracy, кроме того, для более подробной оценки можно использовать матрицу ошибок, F-1scoreи другие метрики.

Далее на основании результатов кросс-валидации и тестирования выбирается модель с наилучшей производительностью. Анализируется важность признаков и интерпретируется результаты модели.

Для визуализации результатов используются графики и диаграммы, включающие:

- График матрицы ошибок
- Визуализация распределения данных и важных признаков.

Заключение.

Данный проект направлен на предсказание наиболее подходящей сельскохозяйственной культуры по почвенным условиям с использованием различных алгоритмов машинного обучения. В процессе работы были выполнены сбор и предобработка данных, построение и настройка моделей, оценка их производительности и визуализация результатов. Полученные модели помогут фермерам и агрономам в приятии решений о выборе культур для выращивания в конкретных почвенных условиях.