УДК: 311 + 338.432

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ УРОЖАЙНОСТИ В РЕГИОНАХ РФ

Сергеев Степан Иванович, *студент 3 курса Института экономики и управления АПК, ФГБОУ ВО РГАУ-МСХА имени К. А. Тимирязева, sstepan5725@gmail.com*

Научный руководитель – Титов Артем Денисович*, ассистент кафедры Статистики и кибернетики, ФГБОУ ВО РГАУ-МСХА имени К. А. Тимирязева, a.titov@rgau-msha.ru*

Аннотация: *В данной статье рассматривается использование статистических методов и методов машинного обучения для прогнозирования урожайности. В статье показаны преимущества и недостатки каждого метода. Сравнение статистических методов и методов машинного обучения показывает, что последние обеспечивают более высокую точность предсказаний, что подчеркивает их значимость для оптимизации агропрактики и повышения продовольственной безопасности.*

Ключевые слова: *Урожайность, машинное обучение, агропрактика, прогнозирование, статистические методы, информационные системы***.**

THE USE OF STATISTICAL METHODS AND MACHINE LEARNING METHODS TO PREDICT YIELDS IN THE REGIONS OF THE RUSSIAN FEDERATION

Sergeev Stepan Ivanovich, *3th year undergraduate student of the Institute of Economics and Management of the Agro–Industrial Complex, Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, sstepan5725@gmail.com*

Scientific supervisor – Titov Artyom Denisovich*, qssistant at the Department of Statistics and Cybernetics, Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, a.titov@rgau-msha.ru*

Annotation: *This article discusses the use of statistical methods and machine learning methods to predict yields. The article shows the advantages and disadvantages of each method. A comparison of statistical methods and machine learning methods shows that the latter provide higher accuracy of predictions, which underlines their importance for optimizing agricultural practices and improving food security.*

Key words: *Productivity, machine learning, agropractic, forecasting, statistical methods, information systems*

Современное сельское хозяйство России сталкивается с многочисленными вызовами, вызванными климатическими и экономическими факторами. Прогнозирование урожайности сельскохозяйственных культур становится важной задачей для эффективного использования природных ресурсов. Актуальность исследования обусловлена несколькими аспектами. Климатические изменения влияют на сельскохозяйственные процессы, особенно в регионах с разнообразными климатическими зонами. Экстремальные условия снижают урожайность, что делает прогнозирование критически важным для минимизации рисков. Точные прогнозы способствуют улучшению планирования сельскохозяйственного производства и стабилизации продовольственного рынка. Традиционные статистические подходы, такие как линейная регрессия, имеют ограничения и чувствительны к отсутствию данных, что снижает точность прогнозов. Методы машинного обучения могут обрабатывать большие объемы данных и учитывать нелинейные зависимости, что делает их перспективными для точного прогнозирования урожайности.

В статье рассматриваются ключевые различи между статистическими методами и методами ML, а также их применение в регионах РФ. Статистические модели предполагают линейные зависимости между переменными, что может упрощать сложные взаимосвязи и игнорировать нелинейные эффекты, влияющие на урожайность. Модели ML, такие как Random Forest, учитывают нелинейные зависимости и взаимодействия, улучшая точность предсказаний. Оба метода имеют разные требования к данным и устойчивость к выбросам. Статистические модели чувствительны к качеству данных и требуют предварительной нормализации, что увеличивает трудоемкость подготовки данных. Напротив, методы ML более устойчивы к выбросам и могут обрабатывать большие объемы данных различных форматов, снижая потребность в детальной предобработке. Интерпретируемость моделей также имеет значение: статистические модели, такие как линейная регрессия, легко интерпретируемы и позволяют определить вклад каждого фактора в прогноз, тогда как ML-модели могут быть сложнее для интерпретации из-за нелинейных взаимодействий.

Для данного исследования использовались агрономические, климатические и почвенные данные из ключевых сельскохозяйственных регионов России: Орловской области, Республики Татарстан и Ставропольского края. Эти регионы выбраны из-за их значительной роли в производстве основных культур и разнообразия климатических условий, влияющих на урожайность. Данные были получены из Единой межведомственной информационно-статистической системы (ЕМИСС), что обеспечило их достоверность и единообразие. Агрономические данные: собраны сведения по урожайности, посевным площадям и внесению удобрений для пшеницы, картофеля и сахарной свеклы за 2000–2023 годы. Климатические данные включают средние значения

температуры, давления, влажности, скорости ветра, осадков и снежного покрова, сгруппированные по декадам с 2005 по 2023 год. Данные, полученные с архива погоды rp5.ru, агрегированы по 10-дневным периодам для упрощения анализа и повышения точности прогнозирования. Почвенные данные содержали информацию о типах почв и их распространении в регионах, включая процентное содержание различных типов от общего объема почвенных ресурсов. Эти данные были получены из Единого государственного реестра почвенных ресурсов России.

Данные из ЕМИСС были объединены и очищены от пропусков, проведена нормализация и стандартизация для корректной работы моделей. Этот массив использовался для построения статистической модели и модели машинного обучения (Random Forest), что обеспечило корректное сравнение методов. Данные были разделены на обучающую и тестовую выборки для оценки точности моделей на незадействованных данных.

После обработки составлена матрица корреляций для всех субъектов и культур, что позволило выявить взаимосвязи между агрономическими, климатическими и вегетационными показателями (Рисунок 1). Для визуализации зависимостей использовалась тепловая карта, где значения корреляции представлены цветовой шкалой. Анализ матрицы позволяет выделить ключевые факторы для дальнейшего построения моделей машинного обучения.

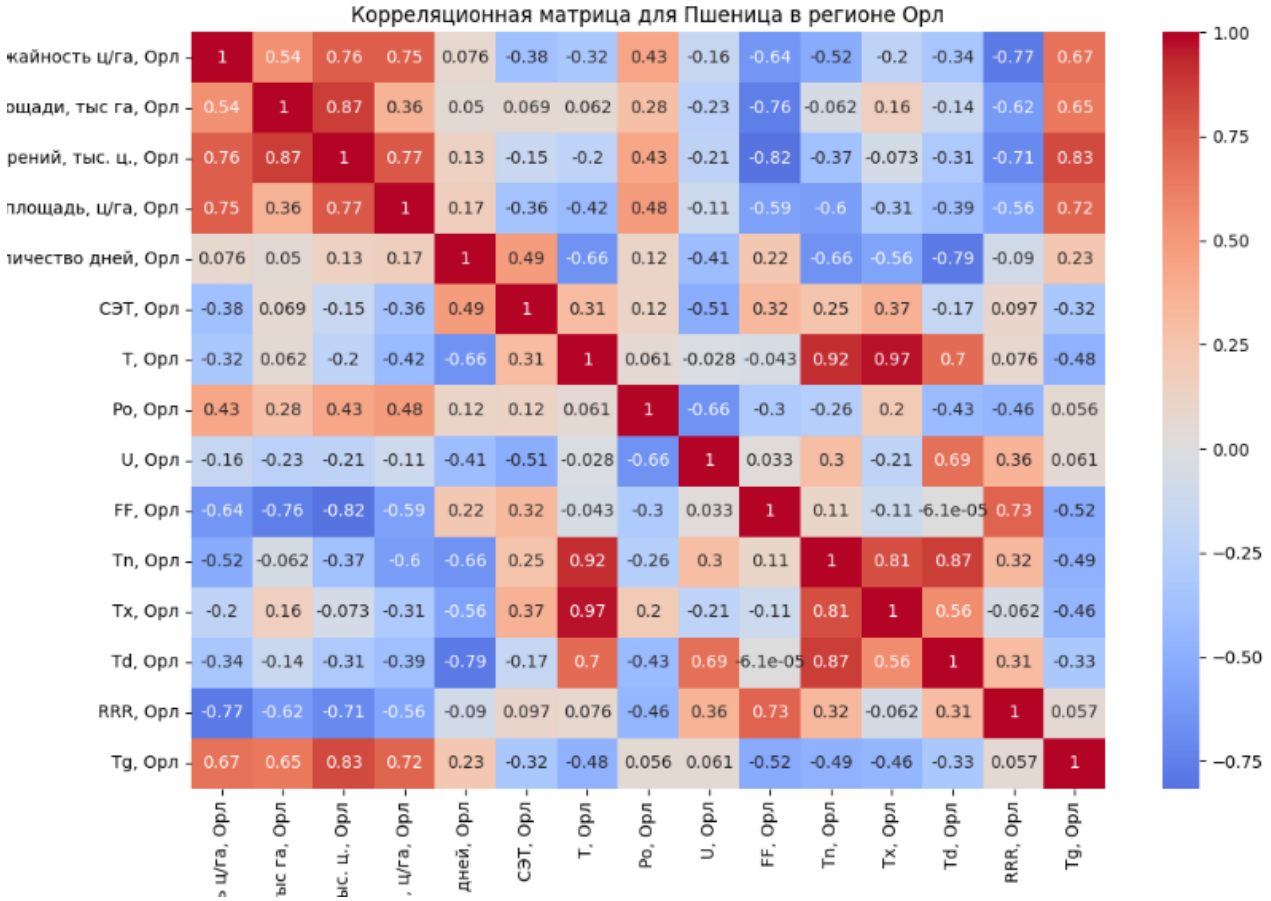

Рисунок 1 – **Пример матрицы корреляции**

В ходе исследования было обнаружено, что определенные климатические условия, такие как температура и количество осадков, оказывают значительное влияние на урожайность, что подтверждается высокой степенью корреляции. Эти данные будут служить основой для разработки предсказательных моделей.

Для построения статистической модели использовался метод множественной линейной регрессии, который позволяет оценить влияние различных факторов на урожайность сельскохозяйственных культур. Модель построена на основании следующих переменных. Зависимая переменная: урожайность (ц/га). Независимые переменные: климатические данные (температура, осадки и др.), агрономические показатели (посевные площади, объем внесенных удобрений), а также почвенные характеристики. Программный код для реализации модели на языке Python был организован в несколько этапов. Сначала данные были отфильтрованы и подготовлены для анализа. Затем переменные были разделены на обучающую и тестовую выборки. С помощью библиотеки statsmodels была создана линейная регрессионная модель, которая позволила оценить влияние независимых переменных на урожайность. Модель была оценена по среднеквадратичной ошибке (MSE) и коэффициенту детерминации (R^2). Полученные результаты продемонстрировали удовлетворительную точность прогнозов, что подтверждается значениями MSE и R^2, находящимися на приемлемом уровне.

В рамках исследования для прогнозирования урожайности с помощью методов машинного обучения была использована модель Random Forest. Основное преимущество этой модели заключается в ее устойчивости к выбросам и высокой способности к обобщению данных. Для построения прогноза модель учитывает средний результат предсказаний множества деревьев, что позволяет снизить переобучение и повысить точность предсказаний. Кросс-валидация модели проводилась с 5-кратным разбиением данных (cv=5). Этот метод позволяет объективно оценить обобщающую способность модели на новых данных и выбрать оптимальные параметры. Кросс-валидация позволила выбрать оптимальные гиперпараметры модели и обеспечить баланс между точностью и обобщающей способностью. На каждом этапе модель обучалась и оценивалась, что позволило выбрать такие параметры, как количество деревьев (n_estimators=50) и максимальная глубина (max_depth=3), при которых достигается минимальная ошибка.

После применения статистического метода и методов машинного обучения для всех культур и субъектов была составлена сравнительная таблица, которая демонстрирует значительное преимущество машинного обучения (ML) по сравнению с традиционным статистическим методом.

Для всех рассматриваемых культур и регионов среднеквадратичная ошибка (MSE) ML-метода оказалась значительно ниже, чем у статистического. Например, для картофеля в Ставропольском крае MSE составляет 13,27 для ML, против 93,39 для статистического метода. Аналогичные тенденции наблюдаются и для других культур: для пшеницы в Орле MSE снизилась с 24,68 до 2,92, а для свеклы — с 119,77 до 46,41. Это говорит о том, что ML метод обеспечивает более

точные предсказания урожайности, что, вероятно, связано с его способностью учитывать нелинейные зависимости и взаимодействия между признаками.

Таблица 1

Сравнительная таблица использования статистических методов и методов машинного обучения

Метод	Показатель	Карт, Орл	Карт, Став	Карт, Тат	Пшен, Орл	Пшен, Став	Пшен, Тат	Свекл, Орл	Свекл, Став	Свекл, Тат
Статистический	Среднеквадратичная ошибка (MSE)	119,77	93,39	266,93	24,68	9,91	20,15	119,77	93,39	266,93
	Коэффициент детерминации (R^2)	0,80	0,94	0,82	0,73	0,66	0,61	0,80	0,94	0,82
ML	Среднеквадратичная ошибка (MSE)	46,41	13,27	62,84	2,92	1,53	4,85	46,41	13,27	62,84
	Коэффициент детерминации (R^2)	0,92	0,99	0,94	0,96	0,95	0,90	0,92	0,99	0,94

Коэффициент детерминации (R^2) также показывает более высокие значения для методов машинного обучения. Например, для картофеля в Ставропольской области R^2 увеличился с 0,94 до 0,99, что свидетельствует о том, что ML метод объясняет большую долю вариации в данных.

Выводы из этих результатов показывают, что применение методов машинного обучения позволяет значительно улучшить качество предсказаний урожайности, что в свою очередь может оказать положительное влияние на принятие решений в агрономии. Более высокая точность прогнозов дает фермерам возможность оптимально планировать свои ресурсы и адаптироваться к изменениям в климате и других факторах, что в конечном итоге приводит к повышению урожайности и эффективности сельского хозяйства. В заключении, проведенное исследование демонстрирует значительный потенциал применения методов машинного обучения в области агрономии для прогнозирования урожайности различных культур. Сравнение результатов, полученных с использованием статистических методов и методов машинного обучения, подтверждает, что последние обеспечивают более точные и надежные предсказания.

Библиографический список

1. Информационно-аналитическое обеспечение устойчивого развития сельского хозяйства / М. В. Кагирова, В. В. Демичев, Ю. Н. Романцева [и др.]. – Москва : Российский государственный аграрный университет - МСХА им. К.А. Тимирязева, 2023. – 307 с. – ISBN 978-5-9675-2013-6. – EDN RTMIDX.

2. Титов, А. Д. Методы и алгоритмы интеллектуального анализа больших данных в сельском хозяйстве / А. Д. Титов // Материалы международной научно-практической конференции «Тренды развития сельского хозяйства и агрообразования в парадигме Зеленой экономики»: сборник статей, Москва, 14–15 июня 2023 года. – Москва: Российский государственный аграрный университет- Московская сельскохозяйственная академия им. К.А. Тимирязева, 2023. – С. 29-33. – EDN QZGBTG.

3. Сайфетдинов А.Р., Максименко А.А. Применение машинного обучения и искусственного интеллекта для анализа данных сельского хозяйства и повышения урожайности / А.Р. Сайфетдинов, А.А. Максименко // Контентус. – 2023. – № 7S. – Т.8. – С. 28 – 34.

4. Paudel, Dilli & Boogaard, Hendrik & Wit, Allard & Janssen, Sander & Osinga, Sjoukje & Pylianidis, Christos & Athanasiadis, Ioannis. (2020). Machine learning for large-scale crop yield forecasting. Agricultural Systems. 187. 103016. 10.1016/j.agsy.2020.103016.

5. Корреляционно-регрессионный анализ влияния экономических факторов на урожайность пшеницы / В. И. Хоружий, Д. В. Быков, А. В. Уколова, А. Г. Ибрагимов // Бухучет в сельском хозяйстве. – 2024. – № 8. – С. 557-571. – DOI 10.33920/sel-11-2408-04. – EDN MMQTOR.

6. Зинченко, А. П. Практикум по статистике / А. П. Зинченко, О. Б. Тарасова, А. В. Уколова. – Москва : Российский государственный аграрный университет - МСХА им. К.А. Тимирязева, 2013. – 314 с. – EDN WEDVEZ.

7. Уколова, А. В. Эконометрика : практикум / А. В. Уколова, Е. В. Шайкина. – Москва : Российский государственный аграрный университет - МСХА им. К.А. Тимирязева, 2011. – 105 с. – EDN WEDTMJ.