УДК 332.055.2

БИБЛИОТЕКИ PYTHON ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ: ОБЗОР И ПРАКТИЧЕСКАЯ АПРОБАЦИЯ

Галимуллин Данил Рамильевич, *студент 1 курса института экономики им. А. В. Чаянова, ФГБОУ ВО РГАУ−МСХА имени К. А. Тимирязева, e-mail: danil.frompayday@gmail.com*
Казлаускас Анастасия Сергеевна, *студентка 1 курса института экономики им. А. В. Чаянова, ФГБОУ ВО РГАУ−МСХА имени К. А. Тимирязева, e-mail: anastasia12kazlauskas@gmail.com*
Научный руководитель – Демичев Вадим Владимирович, *к.э.н., доцент, доцент кафедры статистики и кибернетики ФГБОУ ВО РГАУ−МСХА имени К. А. Тимирязева, e-mail: demichev_v@rgau-msha.ru*

Аннотация. *В данной статье проведена апробация часто используемых для анализа больших данных загрузочных пакетов (библиотек), реализованных на языке программирования Python. По результатам практического применения библиотек, убедились в их эффективности и удобстве. В качестве содержательного вывода приведены выявленные причинно-следственные взаимосвязи результативных и факторных показателей эффективности сельского хозяйства.*
Ключевые слова: *анализ больших данных, Python, библиотеки, урожайность зерновых.*

Применение библиотек в современных языках программирования существенно упрощает и, вместе с тем, обогащает анализ данных [2]. В данном исследовании были рассмотрены и апробированы такие библиотеки как pandas, NumPy, Scikit-learn, Matplotlib, Seaborn.

Апробация библиотек была реализована на основе базы данных, содержащей информацию о факторах и результатах сельскохозяйственной деятельности регионов 77 регионов Российской Федерации в течение 15лет (панельные данные).

Библиотека Pandas позволила быстро и эффективно проанализировать и обработать данные. Загрузочный пакет NumPy позволил поменять форму массивов, не затрагивая при этом данные, которые в нем находятся. Для реализации модели машинного обучения применялась библиотека Scikit-learn. С ее помощью был реализован алгоритм регрессионного дерева решений, позволивший нам с высокой точностью предсказать значения определенных переменных в зависимости от других параметров [1]. Загрузочные пакеты Matplotlib и Seaborn позволили визуализировать полученные в ходе вышеупомянутых манипуляций данные [4].

Результатом исследования представленных данных стало построение корреляционной матрицы [3, 5], в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции между соответствующими показателями (рисунок 1).

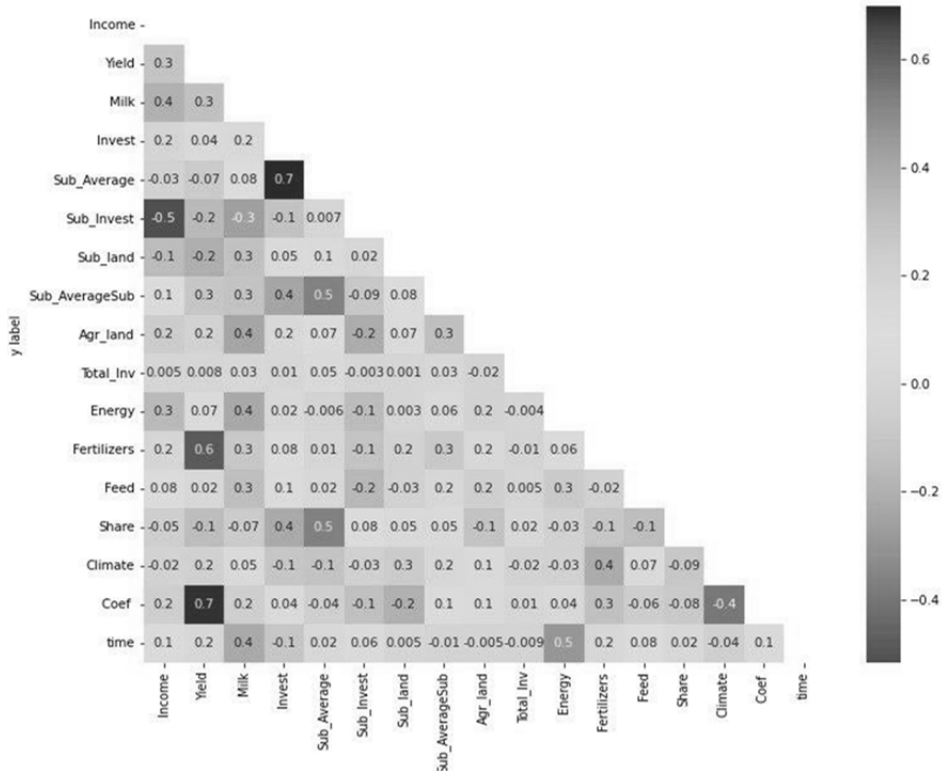

Рисунок 1 – Корреляционная матрица

В качестве индикаторов были использованы следующие показатели: Income – прибыль с 1 га посевов с.-х. культур, тыс. руб.; Yield – урожайность зерновых с 1 га посевов с.-х. культур, ц.д.в.; Milk– получено молока в расчете на одну корову, кг.; Invest – инвестиции в основной капитал СХО, тыс. руб.; Energy – энергетические мощности, л.с.; Fertilizers – внесение удобрений на 1 га посевов с.-х. культур, ц. д. в.; Feed – расход кормов в расчете на одну голову крупнорогатого скота, ц.д.в.; Climate – средний балл продуктивности климата и другие.

Далее была построена, как было упомянуто выше, модель регрессионного дерева решений, которая базируется на зависимости урожайности от внесения минеральных удобрений, выбор именно этой зависимости аргументирован высоким значением коэффициента корреляции между соответствующими параметрами в корреляционной матрице (рисунок 1). В качестве примера можем также отметить, что на урожайность оказывают высокое значение такие показатели как внесение удобрений, средний балл продуктивности климата, уровень государственной поддержки и другие факторы.

Для оценки качества модели рассчитана среднеквадратическая ошибка (RMSE). Ее значение составило приблизительно 10,365, что является допустимым показателем.

Таким образом, в ходе работы, выявили, что использование таких библиотек для анализа данных, как Pandas, NumPy, Sklearn, Seaborn, Matplotlib позволяет изучать причинно-следственные взаимосвязи с высоким уровнем эффективности.

Библиографический список

1. **Гуручаран, М. К.** Основы машинного обучения: регрессионное дерево решений / М. К. Гуручаран – Режим доступа: https://towards-datascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda (Дата обращения 08.11.2022).

2. **Дэви, С.** Основы Data Science и Big Data. Python и наука о данных / С. Дэви, М. Арно, А. Мохамед. − СПб. : Питер, 2018. − 336 с.

3. **Мохтар, Ибрагим** Туториал по созданию корреляционной матрицы – Режим доступа: https://likegeeks.com/python-correlation-matrix (Дата обращения 08.11.2022).

4. Пять ключевых библиотек и пакетов для анализа данных на Python [Электронный ресурс]. − Режим доступа: https://techrocks.ru/2018/07/22/5-key-libraries-and-packets-for-data-analysis-in-python/ (Дата обращения: 21.09.2022).

5. Тринадцать способов настроить визуализацию матрицы корреляции. [Электронный ресурс]. − Режим доступа: https://datastart.ru/blog/read/seaborn-heatmaps-13-sposobov-nastroit-vizualizaciyu-matricy-korrelyacii (Дата обращения: 30.09.2022).