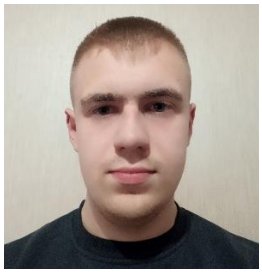
УДК 004.855.5

ПРОГНОЗИРОВАНИЕ ДНЕВНОГО КОЛИЧЕСТВА ОСАДКОВ МЕТОДАМИ МАШИННОГО БУЧЕНИЯ






В.Т. Кучеренко
студент БГУИР, инженер-программист

С.Н.Нестеренков
кандидат технических наук, доцент кафедры программного обеспечения информационных технологий, декан факультета компьютерных систем и сетей

И.В.Шилов
студент БГУИР, инженер-программист

А.Н. Марков
старший преподаватель, магистр технических наук, заместитель начальника Центра информатизации и инновационных разработок БГУИР

Центр информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники, Республика Беларусь
Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: vova.kucherenko.00@mail.ru, s.nesterenkov@bsuir.by, Ilyashilov@mail.by

В.Т. Кучеренко
Студент Белорусского государственного университета информатики и ридиоэлектроники.

С.Н. Нестеренков
Кандидат технических наук, декан факультета компьютерных систем и сетей Белорусского государственного университета информатики и радиоэлектроники, доцент кафедры Программного обеспечения информационных технологий. Автор публикаций на тему машинного обучения, алгоритмов принятия решений, искусственных нейронных сетей и автоматизации

И.В. Шилов
Студент Белорусского государственного университета информатики и ридиоэлектроники.

А.Н. Марков
Магистр технических наук, старший преподаватель кафедры ПИКС, заместитель начальника Центра информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники.

Аннотация. Прогноз ежедневных осадков играет важную роль в производительности сельского хозяйства и обеспечивает снабжение продовольствием и водой для поддержания здоровья граждан. Но неустойчивое распределение осадков в стране влияет на сельское хозяйство, от которого зависит экономика страны. Разумное использование дождевой воды должно планироваться и практиковаться в странах, чтобы свести к минимуму проблему засухи и наводнений. Основная цель этого исследования заключается в выявлении соответствующих атмосферных особенностей, которые вызывают осадки, и прогнозировании интенсивности ежедневных осадков с помощью методов машинного обучения. Для прогнозирования осадков было проведено несколько видов исследований с использованием методов интеллектуального анализа данных и машинного обучения наборами экологических данных. Метод корреляции Пирсона использовался для

выбора соответствующих переменных окружающей среды, которые использовались в качестве входных данных для модели машинного обучения. Набор данных был собран в местном метеорологическом отделении в Минске, Беларусь, для измерения эффективности двух методов машинного обучения (многомерная линейная регрессия и случайный лес).

Ключевые слова: *машинное обучение, многомерная линейная прогрессия, случайный лес*.

Введение.

Еще с древних времен люди верили, что могут предсказывать погоду, но теперь возникла реальная необходимость делать прогноз заранее. Прогноз осадков имеет решающее значение для повышения производительности сельского хозяйства, что, в свою очередь, обеспечивает продовольствие и качественное водоснабжение граждан своей страны. Сельское хозяйство и качество воды зависят от количества осадков и воды на ежедневной и ежегодной основе [1]. Поэтому точное прогнозирование ежедневных осадков является сложной задачей для управления дождевой водой для сельского хозяйства и водоснабжения.

Различные исследователи провели исследования для улучшения прогнозирования ежедневных, ежемесячных и годовых объемов осадков с использованием метеорологических данных разных стран. Исследователи применяли методы интеллектуального анализа больших данных [2] и различные алгоритмы машинного обучения [3,4] для повышения точности ежедневного, ежемесячного и ежегодного прогнозирования осадков. Согласно результатам исследований, процесс прогнозирования в настоящее время перешел от методов интеллектуального анализа данных к методам машинного обучения. Ученые подтвердили, что алгоритмы машинного обучения лучше заменяют традиционный детерминированный метод прогнозирования погоды и осадков.

Несколько факторов окружающей среды влияют на наличие осадков и их интенсивность. Температура, относительная влажность, солнечный свет, давление, испарение и т. д. являются одними из факторов, которые прямо или косвенно влияют на наличие осадков и его интенсивность. В связи с данными факторами исследование было направлено на выявление соответствующих атмосферных особенностей, вызывающих осадки, и прогнозирование интенсивности ежедневных осадков с помощью методов машинного обучения. [5] Исходные данные собираются из региональной метеорологии и предварительно обрабатываются, чтобы сделать их пригодными для эксперимента. Каждая особенность предварительно обработанных данных коррелирует с переменной осадков для выявления соответствующих особенностей с помощью корреляции Пирсона. Затем в исследовании были проведены эксперименты с алгоритмами машинного обучения Radnom forest (RF) и MLR.

Многомерная линейная прогрессия.

Линейная регрессия может быть многомерной, которая имеет несколько независимых переменных, используемых в качестве входных функций, и простую линейную регрессию, которая имеет только одну независимую или входную функцию. Обе линейные регрессии имеют одну зависимую переменную, которую можно прогнозировать или предсказать на основе входных функций. В этом документе представлена многомерная линейная регрессия, потому что для прогнозирования зависимой переменной, называемой ежедневным количеством осадков, использовалось несколько переменных или особенностей окружающей среды. [6] Линейная регрессия – это контролируемый метод машинного обучения, используемый для прогнозирования неизвестного ежедневного количества осадков с использованием известных переменных окружающей среды. Многомерная линейная регрессия использовала несколько пояснительных или независимых переменных (X) и одну зависимую или выходную переменную, обозначаемую Y.

Общее многомерное уравнение линейной регрессии этой статьи приведено как:

$$ЕКО = (год * \beta1) + - (месяц * \beta2) + - (день * \beta3) + - (МахТ * \beta4) + - (МинТ * \beta5) + - (Вл * \beta6) + -(Исп * \beta7) + - (СлСв * \beta8) + - (СкСв * \beta9) + - \xi_я$$

Формула 1. Уравнение линейной регрессии

где ЕКО – ежедневное количество осадков, Y – год, M - , D - , МахТ – максиамльная температура, МинТ – минимальная температура, Вл – влажность воздуха, Исп – испарение, СлСв – солнечный свет, СкСв – скорость света, β – коэффициент регрессии, $\xi_я$ – это термин ошибки или шум

Случайный Лес (Random Forest).

Модель случайной лесной регрессии мощна и точна. Обычно он отлично работает во многих проблемах, включая функции с нелинейными отношениями. Случайная лесная регрессия (СЛ) – это контролируемый алгоритм машинного обучения, который использует метод ансамбля для регрессии. СЛ работает, создавая несколько деревьев решений во время обучения и выводя среднее значение классов в качестве прогноза всех деревьев. Алгоритм RF работает на следующих этапах:

1. Возьмите случайные точки данных p из тренировочного набора
2. Создайте дерево решений, связанное с этими точками данных p
3. Возьмите число N деревьев, чтобы построить и повторить шаги 1 и 2
4. Для новой точки данных заставьте каждое из N деревьев дерева предсказать значение y для точки данных и назначить новую точку данных среднему из всех прогнозируемых значений y.

Алгоритм случайного леса является одним из контролируемых алгоритмов машинного обучения, которые выбираются в качестве прогнозной модели для ежедневного прогнозирования осадков с использованием входных переменных или функций среды. Случайная лесная регрессия управляется путем построения множества деревьев решений во время обучения и вывода класса, который является способом среднего прогнозирования или регрессии отдельных деревьев. Согласно, СЛ-алгоритм эффективен для больших наборов данных, и хороший экспериментальный результат получается с использованием больших наборов данных, в которых отсутствует доля данных. [7]

Сбор данных

Для этого исследования необработанные данные были собраны с региональной метеорологической станции в городе Минск, Беларусь. Были включены десять функций данных, таких как год, месяц, дата, испарение, солнце, максимальная температура, минимальная температура, влажность, скорость ветра и количество осадков.

Измерение производительности

Корреляция Пирсона использовалась для измерения силы взаимосвязи между двумя переменными. Две переменные могут быть положительно или отрицательно коррелированы и не связаны между двумя переменными, если коэффициент корреляции Пирсона равен нулю. Модель коэффициентов корреляции Пирсона математически описывается как:

$$P_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 * (y_i - \bar{y})^2}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Формула 1. Корреляция Пирсона

где P_{xy} – коэффициент корреляции Пирсона, {(x1, y1), (x2, y2), ..., (xn, yn)} –парные данные, состоящие из n пар, $\bar{x}$ и $\bar{y}$ являются средним значением x и y соответственно.

Алгоритмы машинного обучения используют функции входных данных, которые выбираются с использованием коэффициента корреляции Пирсона в качестве соответствующих функций.

Заключение.

Чтобы выбрать переменные окружающей среды, которые коррелируют с осадками, корреляция Пирсона была проанализирована на переменных окружающей среды, представленных в таблице 1. Поскольку набор данных велик, переменные, которые коррелируют более 0,20 с осадками, рассматривались в качестве экологических особенностей эксперимента по прогнозированию осадков. Следовательно, для прогнозирования количества ежедневных осадков результаты экологических атрибутов, относящихся к ежедневному прогнозированию осадков, таких как испарение, относительная влажность, солнечный свет, максимальная суточная температура и минимальная суточная температура, показаны в таблице 1.

Таблица 1. Экологические особенности и их значение коэффициента Пирсона

Метрики	**Значение**
Год	0.012
Месяц	0.101
День	0.017
Испарение	0.279
Влажность	0.401
Максимальная дневная температура	0.296
Минимальная дневная температура	0.204
Солнечный свет	0.351
Скорость ветра	0.046
Дневные осадки	1.000

Список использованных источников

[1] Збигнев В.К. Изменение климата и водные ресурсы 3-е издание. 2008. – С. 219.

[2] A Data-Driven Approach for Accurate Rainfall Prediction. Shilpa Manandhar,Yee Hui Lee, Yu Song Meng. 2019 год

[3] Прикладной регрессивный анализ, 3-е издание. Norman Draper, Harry Smith. 2016 год

[4] Прогнозное моделирование в IBM SPSS Statistics, R и Python. Метод деревьев решений и случайный лес. Артем Груздев. 2018 год.

[5] Нестеренков, С.Н. Адаптивный поиск вариантов расписания с использованием модифицированного генетического алгоритма / С.Н. Нестеренков // Вести Института современных знаний. - 2015. - N 2. - С. 67-74.

[6] Нестеренков, С.Н. Функциональная модель процедур планирования и управления образовательным процессом как основа построения информационной среды учреждения высшего образования / С.Н. Нестеренков, Н.В. Лапицкая // Вести Института современных знаний. - 2018. - N 1. - С. 97-105.

[7] Нестеренков, С.Н. Сетевая модель и алгоритм составления расписания учебных занятий на основе данных прошлых периодов / С.Н. Нестеренков, Н.В. Лапицкая, О.О. Шатилова // Вести Института современных знаний. - 2018. - № 4. - С. 85-92.

FORECASTING DAILY PRECITATION BY MACHINE LEARNING

V.T. KUCHERENKO
Student of BSUIR, software engineer

S.N.NESTERENKOV
PhD, Associate Professor Dean of the Faculty of Computer Systems and Networks.

I.V.SHILOV
Student of BSUIR, software engineer

A.N. MARKOV
Senior lecturer of the department, Deputy head of the Center for Informatization and Innovative Developments

Center for Informatization and Development of the Belarusian University of State Informatics and Radioelectronics, Republic of Belarus.
Belarusian State University of Informatics and Radioelectronics, Republic of Belarus.
E-mail: vova.kucherenko.00@mail.ru, s.nesterenkov@bsuir.by, Ilyashilov@mail.ru

Abstract. The forecast of daily rainfall plays an important role in agricultural productivity and ensures food and water supplies to keep citizens healthy. But the country's erratic distribution of rainfall affects agriculture, on which the country's economy depends. The wise use of rainwater should be planned and practiced in countries to minimize the problem of drought and floods. The main goal of this study is to identify the relevant atmospheric features that cause precipitation and predict the intensity of daily precipitation using machine learning methods. Several types of research have been conducted to predict rainfall using data mining and machine learning methods on environmental datasets. The Pearson correlation method was used to select appropriate environmental variables that were used as input to the machine learning model. The dataset was collected at the local meteorological office in Minsk, Belarus to measure the performance of two machine learning methods (multivariate linear regression and random forest).

Keywords: machine learning, Multivariate Linear Regression, Random Forest.