УДК 004.946: 004.85

КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ НА ОСНОВЕ АЛГОРИТМОВ И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В СРЕДЕ PYTHON

[1]И.Л. Касимов, [1]Н.И. Юнусов, [2]Ш.Ш. Зияев

[1]Таджикский национальный университет

[2]Таджикский технический университет имени академика М.С. Осими

В данной статье рассматриваются основные алгоритмы и методы машинного обучения, включая деревья решений и их ансамблевые модификации, такие как случайный лес и градиентный бустинг. Деревья решений являются мощным инструментом для анализа данных, позволяя моделировать сложные зависимости между признаками и предсказывать результаты. В статье анализируются принципы построения деревьев решений, критерии разбиения, методы уменьшения переобучения, а также преимущества и недостатки данного подхода.

Рассматриваются примеры применения деревьев решений в различных областях. Приведено математическая постановка задачи установления диагноза заболеваний верхних дыхательных путей и результаты решения задачи в виде графа.

Ключевые слова: *искусственный интеллект, машинное обучение, нейронные сети, деревья решений, случайный лес, градиентный бустинг, кластеризация, ансамблевые методы, анализ данных, предсказательные модели, оптимизация решений.*

МОДЕЛСОЗИИ КОМПЮТЕРИИ ҲАЛЛИ ДАРАХТШАКЛ ДАР АСОСИ АЛГОРИТМҲО ВА УСУЛҲОИ ОМӮЗИШИ МОШИНӢ ДАР МУҲИТИ PYTHON

И.Л. Қосимов, Н.И. Юнусов, Ш.Ш. Зиёев

Дар ин мақола алгоритмҳо ва усулҳои асосии омӯзиши мошинҳо, аз ҷумла дарахтони қарорҳо ва тағйиротҳои ансамбли онҳо, ба монанди ҷангалҳои тасодуфӣ ва градиент бустинг. Дарахтони қарорҳо як воситаи пуриқтидор барои таҳлили додаҳо мебошанд, ки имкон медиҳанд вобастагии мураккаби байни хусусиятҳоро моделонед ва натиҷаҳоро пешгӯӣ кунед. Дар мақола принсипҳои сохтани дарахтони қарорҳо, меъёрҳои тақсимкунӣ, усулҳои кам кардани фишурдани зиёдатӣ, инчунин афзалиятҳо ва нуқсонҳои ин равиш таҳлил карда мешаванд. Намунаҳои истифодаи дарахтони қарор дар соҳаҳои гуногун баррасӣ карда мешаванд. Тартиби математикии масъалаи мукаррар намудани ташхиси беморихои роххои болоии нафас ва натичахои халли масъала дар шакли график оварда шудаанд.

Калидвожаҳо: *зеҳни сунъӣ, омӯзиши мошинӣ, шабакаҳои нейронӣ, ҳалли дархтшакл, ҷангали тасодуфӣ, градиент бустинг, синфгузорӣ, усулҳои ансамблӣ, таҳлили додаҳо, моделҳои пешгӯӣ, оптималгардонии қарорҳо.*

COMPUTATIONAL MODELING OF DECISION TREES BASED ON MACHINE LEARNING ALGORITHMS AND METHODS IN THE PYTHON ENVIRONMENT

I.L. Qosimov, N.I. Yunusov, Sh.Sh. Ziyoev

This article discusses the main algorithms and methods of machine learning, including decision trees and their ensemble modifications, such as random forest and gradient boosting. Decision trees are a powerful tool for data analysis, allowing you to model complex dependencies between features and predict results. The article analyzes the principles of constructing decision trees, partitioning criteria, methods for reducing overfitting, as well as the advantages and disadvantages of this approach. Examples of using decision trees in various fields are considered. A mathematical formulation of the problem of establishing a diagnosis of upper respiratory tract diseases and the results of solving the problem in the form of a graph are given.

Keywords: *artificial intelligence, machine learning, neural networks, decision trees, random forest, gradient boosting, clustering, ensemble methods, data analysis, predictive models, decision optimization.*

Введение

Одним из важнейших направлений искусственного интеллекта (ИИ), является машинное обучение (МО), которое позволяет компьютерам обучаться на основе данных и делать прогнозы, а также принимать решения. В статье анализируются основные типы машинного обучения, а также методы и алгоритмы, используемые в этой области. На рисунке 2 показан граф дерева решений с более сложной структурой. Графы деревьев решений с более сложной структурой применяются в различных областях, где необходимо принимать решения на основе множества факторов или признаков. Чем сложнее структура дерева, тем больше вариантов решений можно учесть, что делает такие. [3, 6].

Существуют три основных типа машинного обучения:

Обучение с учителем (Supervised Learning). В представленных **д**анных, как правило содержатся входные признаки и соответствующие им выходные значения (метки). Основная цель – обучение модели на размеченных данных для последующего предсказания новых значений. Примерами таких алгоритмов, являются линейная регрессия, логистическая регрессия, деревья решений, случайный лес (Random Forest), метод опорных векторов (SVM), нейронные сети. Применение, прогнозирование цен, диагностика заболеваний, системы рекомендаций.

Обучение без учителя (Unsupervised Learning). Данные не содержат заранее размеченных меток. Основная цель – выявление скрытых закономерностей, кластеризация или уменьшение размерности. Примерами таких алгоритмов являются k-means, DBSCAN, метод главных компонентов (PCA) и автоэнкодеры. Применяются для анализа поведения клиентов, выявление аномалий и сегментация рынка сбыта.

Обучение с подкреплением (Reinforcement Learning, RL). Применяется принцип поощрения и наказаний. При этом агент, воздействуя на объект, получает поощрение за успешные действия и учится оптимальной стратегии. В качестве примера можно привести алгоритмы Q-learning, Deep Q-Networks

(DQN), метод градиента политики (Policy Gradient). Области применения, автономные системы, робототехника, игровые системы с искусственным интеллектом. [3,11,12,13].

Таблица 1 – Алгоритмы и методы деревьев решений

Алгоритмы деревьев решений	Примеры алгоритмов	Области применения
Линейная регрессия	Стандартный метод наименьших квадратов (Ordinary Least Squares, OLS) — используется для нахождения коэффициентов линейной модели, минимизируя ошибку между предсказанными и фактическими значениями.	Прогнозирование , оценка показателей объектов, моделирование технологических процессов
Логистическая регрессия	Стандартный метод логистической регрессии, который использует сигмоидальную функцию для преобразования линейной комбинации признаков в вероятность принадлежности к классу.	Классификация изображений, спам-фильтры, предсказание вероятности заболевания или оттока клиентов.

Деревья решений и ансамблевые методы.

Дерево решений — это графическое представление алгоритма, который используется для принятия решений или классификации на основе определённых признаков данных. Это иерархическая структура, состоящая из узлов (вершин), которые представляют собой условия или признаки, и рёбер (ветвей), которые показывают переходы между этими условиями. Конечные узлы (листья дерева) содержат результат или решение, которое получается на основе заданных условий. Основные элементы дерева решений:

Корень — начальный узел, который представляет собой всю выборку данных.

Внутренние узлы — представляют собой признаки (или условия), по которым происходит разделение данных.

Ветки — соединяют узлы и представляют собой решения или результаты, получаемые на основании сравнения признаков.

Листья — конечные узлы, которые содержат результат (классификацию или прогноз), который мы хотим получить.

Ансамблевые методы — это техники машинного обучения, которые объединяют несколько моделей для получения более точных и надежных предсказаний. Они часто применяются для улучшения результатов деревьев решений. [12, 13].

Приведём математическая формулировка дерева решений.

Предположим, у нас есть набор данных $D=\{(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)\}$, где x_i — вектор признаков, а y_i — метка (для задачи классификации) или значение (для задачи регрессии). В процессе построения дерева решений происходит разбиение данных на основе условий, которые минимизируют ошибку в предсказаниях.

Основные шаги для построения модели дерева решений:

Выбор признака и порога для разбиения: для каждого внутреннего узла нужно выбрать оптимальный признак x_j и пороговое значение t_j, по которым данные будут разделяться. Например, для задачи классификации это может быть решение вида $x_j \leq t_j$.

В контексте классификации, выбор признака и порога может быть основан на критерии «информационной выгоды» (например, с использованием показателей энтропии или критерия Джини).

Критерий разделения: для классификации часто используется энтропия или индекс Джини:

Энтропия для набора данных D:

$$H(D) = -\sum_{k=1}^{K} p_k \log(p_k) \quad (1)$$

где p_k — вероятность того, что объект принадлежит классу k.

Индекс Джини:

$$Gini(D) = 1 - \sum_{k=1}^{k} p_k^2 \quad (2)$$

где p_k — вероятность принадлежности к объекту класса k.

При построении регрессионной модели, эти критерии заменяются на *среднеквадратическую ошибку*.

Разбиение данных D производится на каждом узле и разделяются на два подмножества D1 и D2, которые соответствуют ветвям дерева.

Рекурсивное деление: Этот процесс повторяется рекурсивно, пока не будет выполнено условие остановки. Это условие может быть:

- ✓ Достигнут максимальный уровень дерева.
- ✓ Размер подмножества становится слишком малым.
- ✓ Качество разбиения не улучшает точность модели.

Предсказания: после построения дерева, для нового объекта с признаками x процесс предсказания будет состоять в прохождении от корня дерева к листу, следуя по ветвям в зависимости от значений признаков.

Пример. Предположим, что у нас есть набор данных с двумя признаками x_1 и x_2 для задачи классификации.

- ✓ На корне дерева мы выбираем условие $x_1 \leq t_1$, которое делит данные на два подмножества D_1 и D_2.
- ✓ На следующем уровне для каждого из подмножеств можем выбрать новые признаки и условия (например, $x_2 \leq t_2$ для D_1).
- ✓ В конечных листьях дерева мы получаем классы или значения для предсказания.

Таким образом, математическая модель дерева решений включает в себя выбор признаков и порогов на каждом уровне дерева с использованием критериев разбиения (например, энтропии или индекса Джини) и рекурсивное деление данных на подмножества до достижения условий остановки. Используя методы машинного обучения, приведём несколько примеров деревьев решений с использованием компьютерных моделей, код которых написан на языке Python. [4, 12, 13].

Деревья решений – последовательные разветвления по признакам данных.

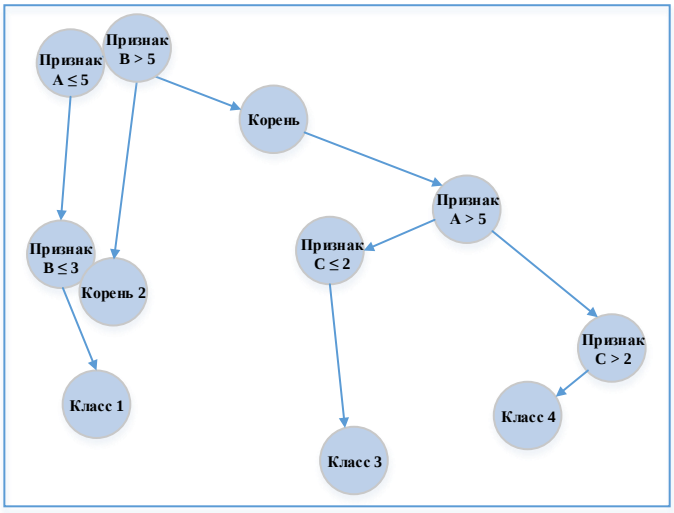

Рисунок 1 – Дерево решений, построенное на основе признаков данных

На рисунке 1 представлен граф дерева решений, где каждый узел представляет признак или конечное решение (класс).

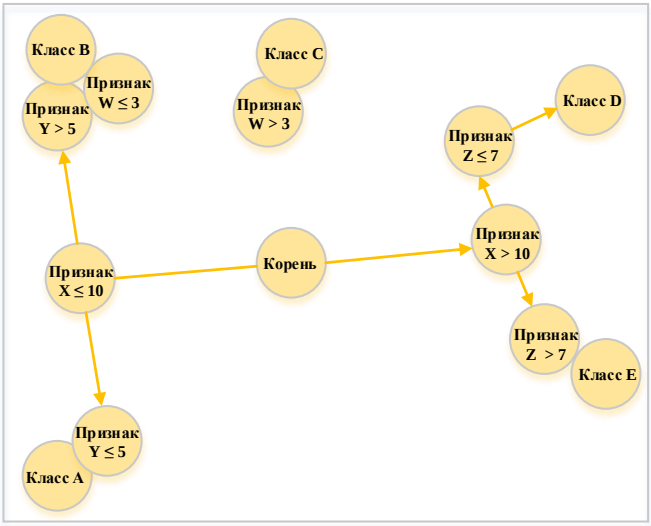

Рисунок 2 – Граф дерева решений с усложнённой структурой

Графы деревьев решений с более сложной структурой (рисунке 2) применяются в различных областях, где необходимо принимать решения на основе множества факторов или признаков. Чем сложнее структура дерева, тем больше вариантов решений можно учесть, что делает такие модели более гибкими и точными в определённых задачах. Области, где используются такие деревья: медицина, финансовый сектор, управление рисками, промышленность и производство, транспорт и логистика и так далее.

Случайный лес (Random Forest) – Случайный лес (Random Forest) — это ансамблевый метод, который использует множество деревьев решений. Ансамбль деревьев решений, улучшающий точность предсказаний.

Он работает так:

1. Создаётся несколько деревьев решений (каждое строится на случайной под выборкой данных).
2. Деревья голосуют за окончательное решение (для классификации — большинством голосов, для регрессии — средним значением). [10, 12, 13].

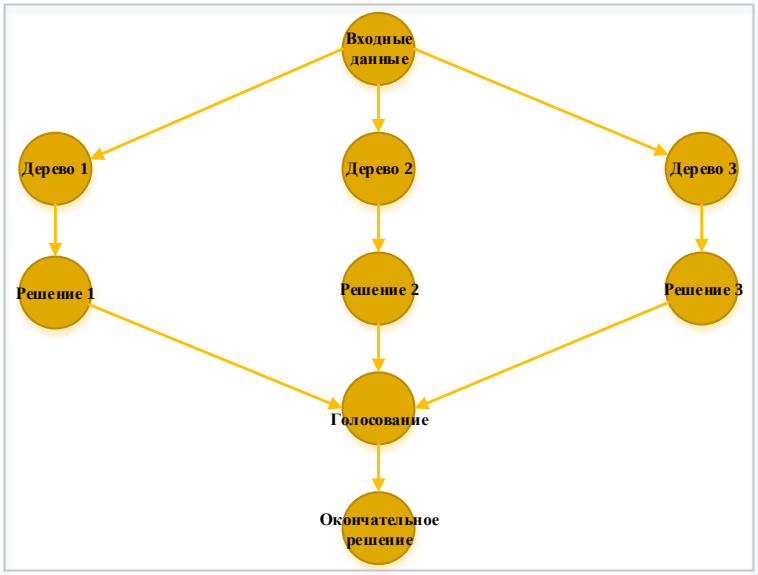

Рисунок 3 – Граф структуры случайного леса (Python)

Граф случайного леса: несколько деревьев решений обрабатывают входные данные, а затем их результаты объединяются через голосование для формирования результата.

Градиентный бустинг (XG Boost, Light GBM) – Это один методов машинного обучения, который используется в основном для задач построения регрессионной модели и классификации. Он основан на принципе последовательного добавления предикторов, обычно представленных в виде деревьев решений. Каждое дерево корректирует своего предшественника, основываясь на градиенте функции потерь, что позволяет модели улучшать свои предсказательные способности [1, 2, 7].

XG Boost (Extreme Gradient Boosting) — это реализация градиентного бустинга, известная своей скоростью и эффективностью. XGBoost обеспечивает высокую производительность за счёт оптимизации вычислений и поддерживает множество параметров для настройки [1, 2].

Light GBM (Light Gradient Boosting Machine) — это метод градиентного бустинга, который формирует сильного ученика путём последовательного добавления слабых учеников методом градиентного спуска. Light GBM отличается высокой скоростью обучения при работе с большими объёмами данных, благодаря оптимизации процесса построения деревьев [1, 2, 5].

Пример: выявление заболеваний на основе анализа связей между симптомами, диагнозами и пациентами с использованием графовой модели.

Для анализа выберем данные пациентов, заболевших инфекционными заболеваниями, которые возникают в результате поражения дыхательных путей различными вирусами. Чаще для обозначения этих инфекций используется термин ОРВИ (острая респираторная вирусная инфекция), часто используется термин ОРЗ (острое респираторное заболевание). ОРВИ обычно проявляться как простуда, острый синусит, острый фарингит, острый ларингит, конъюнктивит, отит, острый бронхит и вирусная пневмония.

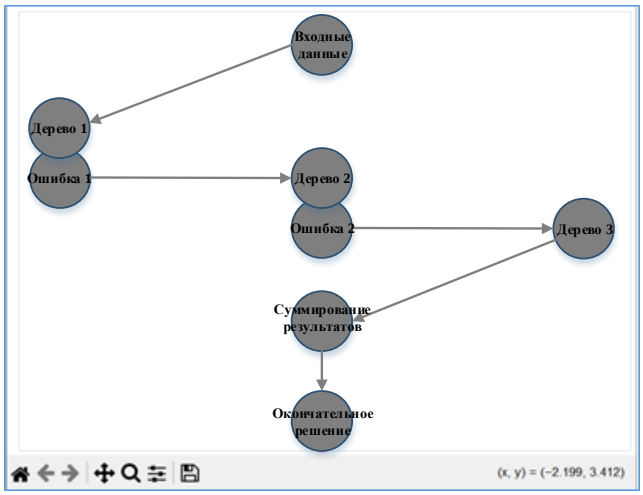

Рисунок 4 – Граф градиентного бустинга (Python)

Математическая формализация задачи

Пусть имеются три типа объектов:

Пациенты: $P = \{p_1, p_2, \dots, p_n\}$

Симптомы: $S = \{s_1, s, \dots, s_m\}$

Диагнозы (заболевания): $D = \{d_1, d, \dots, d_k\}$

Построим неориентированный граф:

$$G = (V, E),$$

где:

$V = P \cup S \cup D$ — множество вершин;

$E \subseteq V \times V$ — множество рёбер, представляющих связи между объектами.

Каждое ребро может иметь тип или вес, отражающий силу связи:

$(p_i, s_j) \in E$: у пациента p_i наблюдается симптом s_j;

$(p_i, d_l) \in E$: у пациента p_i диагностировано заболевание d_l;

$(s_j, d_l) \in E$: симптом s_j часто встречается при заболевании d_l.

Введем также функцию весов:

$$w: E \to [0, 1],$$

которая отражает степень вероятности или силу связи между объектами.

Алгоритм решения задачи включает в себе следующие шаги.

1. **Создание входных данных:** определение списков пациентов, симптомов и диагнозов.
2. **Построение графа:** формирование связей между пациентами, симптомами и диагнозами.
3. **Визуализация графа:** отображение структуры графа для наглядного анализа.
4. **Анализ структуры графа:** применение простой логики предсказания на основе связей между симптомами и диагнозами.

Данные исследования группы пациентов на предмет постановки диагноза заболеваний, приведены в таблице 1. Ниже представлен результат решения задачи с использованием языка Python.

Таблица 1 – Симптомы болезней ОРВИ пациентов

Пациенты P / Симптомы S	1	2	3
Температура высокая	1	1	1
Стекание слизи	1	1	0
Усталость	0	0	1
Кашель	1	0	1
Сыпь	0	0	0
Диагноз D	1	2	3
d_1-грипп, d_2-аллергия, d_3-ковид			

Возможные диагнозы устанавливаемые по симптомам болезни.

d_1- Если у пациента высокая температура, кашель и усталость— вероятно грипп.

d_2- Если температура нормальная и есть насморк — аллергия.

d_3- Если жар, кашель, но есть боль в горле и усталость — ковид.

d_4- Если высокая температура, кашля нет и боли в горле нет — простуда.

d_5- Если температура нормальная и насморка нет — простуда.

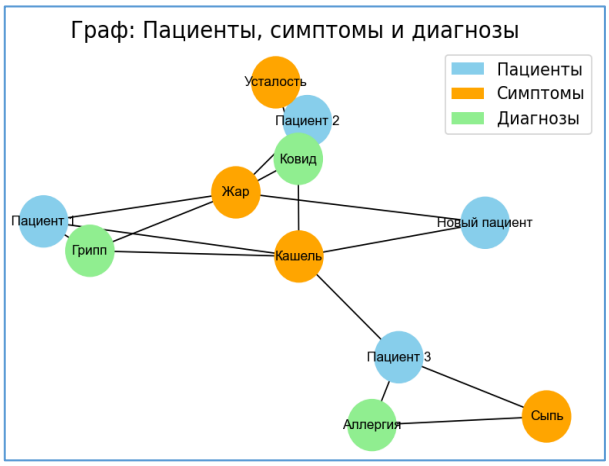

Рисунок 5 – Граф диагноза заболевания по симптомам (Python)

На приведенном графе узлы — это признаки или решения, ребра переходы по условиям, а листья конечные классы (диагнозы).[8, 9]

Заключение

В работе приведены теоретические основы и примеры математических моделей деревьев решений, включающих в себе выбор признаков и порогов на каждом уровне дерева с использованием различных критериев разбиения.

Рассмотрен пример решения конкретной задачи с использованием метода машинного обучения, построения деревьев решений с использованием компьютерных моделей. Граф решения показывает соответствие диагноза пациента признакам болезни. Основное преимущество деревьев решений как видно из приведенного результата, заключается в их интерпретируемости, что делает их подходящими для задач, требующих объяснимости решений. Перспективы применения деревьев решений связаны с их интеграцией в большие данные и использованием гибридных моделей, комбинирующих деревья с нейросетями для улучшения результатов в сложных задачах.

Рецензент: Гуломсафдаров А.Г.– к.т.н., и.о. доцент, зав. кафедрой «Программирование и компьютерная Инженерия» ТТУ имени академика М.С. Осими.

Литература

1. Электронных ресурс. https://www.mql5.com/ru/articles/14926 (дата обращение 10.04.2025).
2. Электронных ресурс. https://elar.urfu.ru/bitstream/10995/140528/1/m_th_ v.i.onufrienko_ 2024.pdf (дата обращение 10.03.2025).
3. Косимов И.Л. Теория и концепсия машиного обучения, искуственого интелекта и нейронных сетей. // Вестник института развития Академии образования Министерство оброзование и науки Республики Таджикистан, Душанбе, 2024, №4 (54) - С.190-196.
4. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
5. Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach. Pearson.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data
8. Mining, Inference, and Prediction. Springer.
9. Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall.
10. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
11. Косимов, И.Л. Решение некоторых сложных математических задач с помощью языка программирование высокого уровня Python // И.Л. Косимов, М.И. Косимова / Вестник Таджикский национальный университет. – 2023. №3. –С.24-34. ISSN: 2413-452X
12. Косимов, И.Л. Компьютерное моделирование физических процессов средствами языка программирования Python 3.10 / И.Л. Косимов // Вестник Таджикского национального университета. Серия естественных наук. – 2022. – №.3. – С.106-115.

13.Косимов, И.Л. Возможности языка программирования Python в решении системы линейных уравнений / И.Л. Косимов, Г.И. Рахматова / Вестник Института развития образования. - №1 (33). – 2021. – С.236-240. EDN: KUXMDN.

МАЪЛУМОТ ДАР БОРАИ МУАЛЛИФОН - СВЕДЕНИЯ ОБ АВТОРАХ -NFORMATION ABOUT AUTHORS

TJ	RU	EN
Қосимов Исмоил Латипович	Касимов Исмаил Латипович	Qasimov Ismail Latipovich
н.и.т., дотсент	к.т.н., доцент	PhD, Associate Professor
Донишгоҳи миллии Тоҷикистон	Таджикский национальный университет	Tajik National University
E-mail: qosismoil@yandex.ru		
TJ	RU	EN
Юнусов Низомуддин Исмоилович	Юнусов Низомуддин Исмаилович	Yunysov Nizomyddin Ismailovich
н.и.т., дотсент	к.т.н, доцент	PhD, Associate Professor
Донишгоҳи миллии Тоҷикистон	Таджикский национальный университет	Tajik National University
	E-mail: U.Nizomyddin@gmai.com	
TJ	RU	EN
Зиёев Шуҳрат Шарофидинович	Зияев Шухрат Шарофидинович	Ziyoev Shuhrat Sharofidinovich
Муаллими калон	Старший преподователь	Senior Lecturer
Донишгоҳи техникии Тоҷикистон ба номи акад. М.С. Осимӣ	Таджикский технический университет имени акад. М.С. Осими	Tajik Technical University named after academician M.S. Osimi
	E-mail: sh.ziyaev1986@gmail.com	