Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Механико-математический факультет

УТВЕРЖДАЮ:
Декан
Л. В.Гензе

Рабочая программа дисциплины

# Технологии машинного обучения

по направлению подготовки
**01.03.01 Математика**
**02.03.01 Математика и компьютерные науки**

Направленность (профиль) подготовки/ специализация:
**Современная математика и математическое моделирование**
**Вычислительная математика и компьютерное моделирование**

Форма обучения
**Очная**

Квалификация
**Математик. Преподаватель / Математик. Аналитик / Математик. Исследователь**
**Математик. Преподаватель / Математик. Вычислитель /**
**Исследователь в области математики и компьютерных наук**

**Год приема**
**2024, 2025**

СОГЛАСОВАНО:
Руководитель ОП
**Л.В.Гензе**

Председатель УМК
**Е.А.Тарасов**

Томск – 2024

**1. Цель и планируемые результаты освоения дисциплины**

Целью освоения дисциплины является формирование следующих компетенций:

РОПК-1.1 Знаком с решенными и не решенными задачами в области своих научных интересов, знаком с методами решения научных задач в области своих научных интересов.

РОПК-1.2 Умеет: - Понимать цели и задачи исследования, предмет и объект исследований, актуальность и значимость проводимых исследований

- Анализировать методы и способы решения исследовательских задач
- Проводить информационный поиск (собирать и обрабатывать научную и научно техническую информацию) для решения исследовательских задач
- Использовать цифровые и информационные ресурсы, научную, опытно-экспериментальную и приборную базы по тематике проводимых исследований и (или) разработок
- Проводить исследования, эксперименты, наблюдения, измерения в рамках решаемых задач
- Интерпретировать научные (научно-технические) результаты, полученные в ходе решения исследовательских задач

РОПК-2.1 Знаком с отечественными и зарубежными базами данных и системами учета научных (научно-технических) результатов

РОПК-2.2 Умеет: - Использовать в профессиональной деятельности отечественные и зарубежные базы данных и системы учета научных (научно-технических) результатов

- Информировать научную общественность о своих результатах полученных в ходе проведенных исследований, экспериментов, наблюдений, измерений на научных (научно-практических) мероприятиях
- Участвовать в научных дискуссиях по тематике своей исследовательской работы на научных (научно-практических) мероприятиях
- Представлять научные (научно-технические) результаты в форме публикаций в рецензируемых научных изданиях
- Представлять научные (научно-технические) результаты в отечественных и зарубежных базах данных и системах учета

**2. Задачи освоения дисциплины**

– Сформировать теоретические знания по основам машинного обучения для построения формальных математических моделей и интерпретации результатов моделирования;

– Выработать умения и навыки использования библиотек языка Python для разработки алгоритмов машинного обучения;

– Развитие и использование математических и информационных средств на основе методов машинного обучения в научной и практической деятельности.

**3. Место дисциплины в структуре образовательной программы**

Дисциплина относится к Блоку 2 «Дисциплина (модули)».

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор.

**4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине**

Девятый семестр, зачет.

**5. Входные требования для освоения дисциплины**

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: программирование, алгебра, математический анализ, дифференциальные уравнения, математическая логика, дискретная математика, теория вероятностей и математическая статистика, практика Python для научных исследований.

**6. Язык реализации**

Русский

**7. Объем дисциплины**

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:
-лекции: 24 ч.
-практические занятия: 24 ч.
в том числе практическая подготовка: 24 ч.
Объем самостоятельной работы студента определен учебным планом.

**8. Содержание дисциплины, структурированное по темам**

Тема 1. Введение в машинное обучение.

Основные этапы решения задачи машинного обучения. Примеры прикладных задач. Виды обучения: с учителем, без учителя, с подкреплением. Основные типы задач. Основные проблемы машинного обучения: недостаточный объем обучающей выборки, пропуски в данных, переобучение/недообучение.

Тема 2. Решение задачи регрессии.

Метод наименьших квадратов. Измерение ошибки в задачах регрессии. Линейная регрессия. Решение проблемы переобучения: L1- регуляризация (Lasso), L2-регуляризация (гребневая регрессия), эластичная сеть (L1+L2). Настройка гиперпараметров алгоритма с помощью перекрестной проверки.

Тема 3. Решение задачи классификации.

Линейная модель классификации. Логистическая регрессия как бинарный классификатор. Функция потерь (ошибок классификации). Confusion matrix (матрица ошибок классификации). Метрики качества классификации: accuracy (доля правильных ответов), precision (точность), recall (полнота), F1-мера. AUC-ROC – площадь под кривой ошибок. Метрическая классификация - метод ближайших соседей (kNN).

Тема 4. Древовидные модели и анмамбли.

Решающие деревья. Случайный лес. Случайный лес как пример бэггинга. AdaBoost. Градиентный бустинг в задачах регрессии и классификации. Стэкинг.

**9. Текущий контроль по дисциплине**

Текущий контроль по дисциплине проводится путем контроля выполнения домашних и индивидуальных заданий и фиксируется в форме контрольной точки не менее одного раза в семестр.

**10. Порядок проведения и критерии оценивания промежуточной аттестации**

Результаты зачета определяются оценками «зачтено», «незачтено».

При оценке выполнения индивидуальных и домашних заданий учитывается правильность, оригинальность и сроки выполнения.

Ближе к концу семестра каждый студент получает итоговое индивидуальное задание. Оно включает в себя некий результирующий итог по освоению материала курса. Работа оформляется в виде отчёта, который студенту необходимо защитить: рассказать о ходе выполнения работы и ответить на дополнительные вопросы по теории выбранной для решения задачи.

По результатам защиты итогового индивидуального задания и по результатам сданной практики определяется оценка.

**Пример задачи**

В ходе разведки месторождений нефти специалисты производят пробные бурения скважин и осуществляют анализ получаемых в ходе этого технических, геологических и геофизических данных. Целью этого является обнаружение нефтенасыщенных пластов, то есть пластов, содержащих в себе нефть и способных ее отдавать.

Перед вами стоит задача разработать алгоритм интеллектуального анализа реальных данных, позволяющий наиболее качественно определять наличие или отсутствие нефтяных пластов на тех или иных глубинах залегания скважин.

Метрикой качества выступает точность нахождения нефтенасыщенного пласта

$$Accuracy = \frac{samples\_true}{sampels\_all},$$

Где $samples\_true$ - количество правильных предсказаний наличия/отсутствия нефтяного пласта,

$sampels\_all$ - общее количество записей в таблице

**Формат ввода**

data_train.csv — файл с обучающими табличными данными

X_data_predict.csv — файл с данными, для которых необходимо предсказать целевую переменную

Файл с тренировочными табличными данными содержит информацию по 600 скважинам, для каждой из которых имеется различная техническая, геологическая и геофизическая информация в виде следующих полей:

- **MD** — относительная глубина скважины (относительно поверхности бурения), всегда является положительной величиной, используется для привязки глубин внутри скважины, но не может выступать в роли какого-то признака при прогнозе (по крайней мере с физической точки зрения).
- **TVDSS** — глубина скважины относительно уровня моря, всегда является положительной величиной, может отражать поверхность геологического пласта или уровень водонефтяного контакта.
- **Layer** — название пласта, геологическая принадлежность интервала, качественная характеристика, выдаваемая геологом на основе его понимания геометрических характеристик целевого пласта, служащая для сопоставления пластов из различных скважин между собой.
- **GK** — гамма-каротаж, измеряет естественную радиоактивность пород, различные минералы имеют разное содержание радиоактивных материалов, как правило, чем выше — тем больше глинистая составляющая и меньше песчанистая, может измеряться в единицах API или мкр/ч.
- **NNKT_big** — нейтронный каротаж, регистрирует относительное водородосодержание, что может говорить о количестве пор в горных породах (они

не могут быть пустыми и всегда содержат какой-то флюид, который в значительном объеме содержит в себе водород). Меньшие значения отвечают за более высокое флюидосодержание.

- **PS** — каротаж естественной поляризации, последняя возникает при фильтрации флюида через породу, уменьшение значений говорит о наличии проницаемого интервала. Единица измерения — милливольты, может иметь совершенно разный масштаб в разных скважинах.
- **IK** — индукционный каротаж, отражает электрическую проводимость горных пород, величину, обратную сопротивлению. Поскольку нефть является диэлектриком, а вода проводником, высокие показания отражают водонасыщенные пласты, а низкие — интервалы, вмещающие нефть. С другой стороны, плотные породы, не содержащие в себе пор, также имеют высокое сопротивление, поскольку не имеют в себе флюида, который способен проводить ток.
- **BK** — боковой зонд, отражает сопротивление горной породы, интерпретируется схожим образом с кривой индукционного каротажа, но уже наоборот, повышенные значения — нефть или плотные породы, пониженные — вода или глина.
- **PZ** — потенциал-зонд, отражает сопротивление горной породы, интерпретируется схожим образом с кривой индукционного каротажа, но уже наоборот, повышенные значения — нефть или плотные породы, пониженные — вода или глина. Схож с боковым зондом (BK), но имеет другую глубинность исследования.
- **Grad_zond** — другая группа зондов, отвечающих за сопротивление горных пород, в зависимости от числа в названии определяется глубинность метода. При бурении буровой раствор попадает в пласт и может изменить содержание того или иного флюида, поэтому, в теории, пониженые сопротивления в затронутой части пласта и повышенные в глубинной могут быть признаком наличия углеводородов.
- **target_collector** — бинарная характеристика, выдаваемая специалистом по интерпретации каротажных данных, отвечающая за то, является ли тот или иной интервал коллекторским пластом, то есть пластом, способным принимать и отдавать флюид.
- **target_oil** — бинарная характеристика, выдаваемая специалистом по интерпретации каротажных данных, отвечающая за то, является ли тот или иной интервал коллекторским нефтенасыщенным пластом.
- **Well** — номер скважины.

В качестве целевой переменной выступает **target_oil**, которая при значении 1 говорит о наличии нефтенасыщенного пласта, а при значении 0 — о его отсутствии.

**Формат вывода**

В файл submission.csv необходимо записать одну колонку, в которой для каждой скважины из тестовой выборки стоит классифицирующая ее метка.

### 11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://lms.tsu.ru/course/view.php?id=37875

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

в) Самостоятельная работа студентов включает в себя: теоретическое освоение лекционного курса, практическое выполнение заданий и индивидуальных заданий, подготовку к зачету с оценкой. Для выполнения самостоятельной работы обеспечивается доступ к информационным ресурсам курса:

- материалы лекций;
- список вопросов для самостоятельной проверки знаний и подготовки к зачёту.
- список литературы, включающий учебники и книги по изучаемым в курсе вопросам.

Все лабораторные работы и индивидуальные задания подобраны так, чтобы максимально стимулировать психологическую установку студентов-математиков на формирование связи между математической теорией и ее практическим применением. Отчет по каждой лабораторной работе включает теоретическую часть, выполненное практическое задание и анализ полученных результатов.

г) Для успешного освоения материала студентам необходимо посещать занятия, а во время самостоятельной работы пользоваться основной и дополнительной литературой, базами данных и информационно-справочными системами, которые представлены в списке литературы. Самостоятельная работа студентов состоит в повторении материала с практических занятий и самостоятельного изучения дополнительных вопросов, более глубокого анализа темы с помощью литературы.

## 12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

1. Любанович Б. Простой Python. Современный стиль программирования. 2-е. изд. – СПб.: Питер, 2021. – 592 с.
2. Самое полное руководство по разработке на Python в примерах от сообщества Stack Overflow. — Москва : Издательство АСТ, 2024. — 672 с.
3. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. - СПб.: Питер, 2016. – 392 с.
4. Грас Дж. Data Science. Наука о данных с нуля. 2-е. изд., перераб. и доп. – СПб.: БХВ-Петербург, 2021 – 416с.
5. Бурков А. Машинное обучение без лишних слов. — СПб.: Питер, 2020. — 192 с.

б) дополнительная литература:

6. Вандер П. Дж. Python для сложных задач: наука о данных и машинное обу-чение. СПб.: Питер, 2020. – 576с.
7. Элбон К. Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. / К. Галлатин, К. Элбон. - 2-е изд., перераб. и доп. - Астана: АЛИСТ, 2024. - 448 с.
8. Уатт Дж. Машинное обучение: основы, алгоритмы и практика применения. СПб.: БХВ-Петербург, 2022 – 640с.

## 13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– операционная система Windows 7 или Windows 10 https://www.microsoft.com/ru-ru/software-download/windows10

– python (дистрибутив python) https://www.python.org/?downloads

– Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index

– ЭБС Лань – http://e.lanbook.com/

## 14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

Для проведения лабораторных работ и самостоятельной работы используются ауди-тории учебно-вычислительной лаборатории ММФ При выполнении индивидуальных заданий, самостоятельных и лабораторных работ используется свободное и лицензионное программное обеспечение:

– офисный пакет Microsoft Office 2010 (составление отчетов);

– IDE для python (программа для организации работы на языке python).

## 15. Информация о разработчиках

Стребкова Екатерина Александровна, ст. преподаватель кафедры вычислительной математики и компьютерного моделирования ММФ ТГУ