

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Министерство образования Московской области

Автономная некоммерческая общеобразовательная организация
«Физтех-лицей» имени П. Л. Капицы

Принята на заседании
педагогического совета
Протокол № _____
от «_____» _____ 202__ г.

Утверждена
Приказом № _____ от
«_____» _____ 202__ г.
Директор
Машкова М.Г.

**Дополнительная общеобразовательная общеразвивающая программа
«Методы анализа данных и машинного обучения»**

Направленность программы: научно-техническая
Уровень программы: углубленный

Возраст обучающихся: 16-18 лет

Срок реализации программы: 1 год

Руководитель направления:

Райгородский Андрей Михайлович
Доктор физико-математических наук,
директор ФПМИ МФТИ

Авторы-составители:

Благодарный Евгений Владимирович
заведующий учебно-методической
лабораторией инноватики
Иванова Анастасия Сергеевна
руководитель проектов учебно-методической
лаборатории _____ инноватики
Тихонов Илья, аспирант кафедры ЭВМ МФТИ

г.
2023

Долгопрудный

Обратная сторона титульного листа:

Программа прошла внутреннюю экспертизу и рекомендована к реализации в АНОО «Физтех-лицей» имени П. Л. Капицы

Экспертное заключение (рецензия) № _____ от « ____ » _____ 202__ г.

Эксперт (ФИО, должность)

Раздел 1. Пояснительная записка

1.1. Нормативно-правовые основы разработки ДООП.

1.2. Направленность ДООП научная.

Дополнительная общеразвивающая программа «Методы анализа данных и машинного обучения» — научно-технической направленности.

1.3. Актуальность программы.

Согласно Указу о национальных целях развития России до 2030 года (от 21 июля 2020 года) одними из задач являются возможности для самореализации и развития талантов, цифровая трансформация. И Указу Президента РФ от 1 декабря 2016 г. N 642 «О Стратегии научно-технологического развития Российской Федерации» одним из перспективных направлений является переход к передовым цифровым, интеллектуальным производственным технологиям, роботизированным системам, новым материалам и способам конструирования, создание систем обработки больших объемов данных, машинного обучения и искусственного интеллекта.

Отличительные особенности программы — формирование эффективной системы выявления, поддержки и развития способностей и талантов у детей и молодежи, основанной на принципах справедливости, всеобщности и направленной на самоопределение и профессиональную ориентацию всех обучающихся.

Уровень освоения программы – углубленный в соответствии с Методическими рекомендациями по проектированию дополнительных общеразвивающих программ (включая разноуровневые программы) министерства образования и науки РФ (письмо от 18 ноября 2015 г. № 09-3242).

1.5. Цель и задачи программы.

Цель программы — получение навыков, необходимых для первичной профессиональной пробы в машинном обучении и анализе данных.

Целевые ориентиры дополнительного образования детей определены в документах:

- Федеральный закон «Об образовании в Российской Федерации» от 29.12.2012 N 273-ФЗ (ст.2 п.1, 2, 3, 14; ст. 75);
- Приказ Министерства просвещения РФ от 27 июля 2022 г. N 629 «Об утверждении Порядка организации и осуществления образовательной деятельности по дополнительным общеобразовательным программам»;
- Концепция развития дополнительного образования детей (утверждена распоряжением Правительства РФ от 31 марта 2022 года N 678-р).

Задачи	обучения:
Участники	будут владеть навыками:
- обработки данных из разных источников, анализа показателей и принятия стратегических решений на основе данных;	
- анализа больших данных с использованием существующей в организации методологической и технологической инфраструктуры;	
- выявления истинных бизнес-проблем или бизнес-возможностей;	
- сбора информации о бизнес-проблемах или бизнес-возможностях.	

Задачи	воспитания:
- подготовить к самостоятельному выбору профессиональных интересов с помощью реализации разных проектных задач;	
- создать условия для социализации и самореализации обучающихся.	

Задачи	развития:
- развивать мотивационные качества учащихся, мотивация к деятельности и решению проблем-задач;	
- развивать интеллектуальные качества учащихся, познавательный интерес и способности, используя данные о применении изучаемых явлений в окружающей жизни, о новостях науки и техники, учебные дискуссии.	

1.6. Категория обучающихся

Обучающийся:

- 16-18 лет;
- владеет навыками программирования на любом языке, знание математики за пределами школьной программы;
- проходит два этапа отбора: тестирование и защита работ. Допускается дополнительный набор обучающихся на следующий год обучения на основании результатов тестирования.

1.7. Сроки реализации и объем программы.

Срок реализации программы – 1 год. Объем программы — 144 часа.

1.8. Формы организации образовательной деятельности и режим занятий.

Формы организации образовательной деятельности: индивидуальные, групповые. Виды занятий: лекции, практические и семинарские занятия, лабораторные работы, круглые столы, мастер-классы, мастерские, деловые и ролевые игры, тренинги, выездные тематические занятия, выполнение самостоятельной работы, концерты, выставки, творческие отчеты, соревнования и другие виды учебных занятий и учебных работ.

Указывается продолжительность и режим занятий в неделю, количество учебных часов в год. Режим занятий должен соответствовать заявленному уровню программы и требованиям Санитарно-эпидемиологических правил и нормативов СанПиН 2.4.4.3172-14.

Численный состав объединения определяется локальным актом образовательной организации в соответствии с приказом Министерства просвещения РФ от 27 июля 2022 г. N 629 «Об утверждении Порядка организации и осуществления образовательной деятельности по дополнительным общеобразовательным программам».

Пример. Занятия проводятся в разновозрастных группах, численный состав группы – 15 человек.

Формы организации образовательной деятельности – групповые, индивидуальные.

Виды занятий: теоретические и практические занятия, деловые и ролевые игры, выставки, творческие отчеты.

Режим занятий: 1-й год обучения — занятия проводятся 2 раза в неделю по 2 часа, 2-й год обучения – 3 раза в неделю по 2 часа.

Продолжительность занятий – 45 минут, перерыв 10 минут.

1.9. Планируемые результаты освоения программы.

Метапредметные результаты освоения программы:

Обучающиеся будут:

- ориентироваться в методах анализа данных и в принципах машинного обучения;
- уметь оценивать результат своей деятельности и достигать результата и извлекать пользу даже из отрицательного результата.

Личностные результаты освоения программы:

Вид деятельности	Освоение
Деятельность по выявлению бизнес-проблем, выяснению потребностей заинтересованных сторон, обоснованию решений и обеспечению проведения изменений в организации	Сбор информации о бизнес-проблемах или бизнес-возможностях
	Выявление истинных бизнес-проблем или бизнес-возможностей
Создание и применение технологий больших данных	Анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры

-

Результаты воспитания:

У обучающихся будут сформированы:

- устойчивый познавательный интерес к профессии разработчик и работа с большими данными;
- ориентация на достижение успеха и перевод проблемы в задачу;
- готовность к самосовершенствованию, духовному саморазвитию.

Раздел 2. Содержание программы.

Содержание программы соотносится с целью и планируемыми результатами ее освоения.

Нормативные ориентиры к отбору содержания дополнительных общеразвивающих программ определены в документах:

- Приказ Министерства просвещения РФ от 27 июля 2022 г. N 629 «Об утверждении Порядка организации и осуществления образовательной деятельности по дополнительным общеобразовательным программам»;
- Концепция развития дополнительного образования детей (утверждена распоряжением Правительства РФ от 31 марта 2022 года N 678-р).

№ п/ п	Наименование модулей	Всего, час.	в том числе:				
			лек ции	практиче ские занятия (семинар ы)	самостоятел ьная работа	контрол ьные задания	Форма контроля
1	Базовые алгоритмы	60	20	30	9	1	тестирование
2	Глубокое обучение	70	20	40	8	2	тестирование
3	Итоговая аттестация	14	6	0	8	0	проект
	Итого:	144	34	30	36	2	

2.1. Учебный (тематический) план на 1 год обучения:

2.2. Содержание учебного плана:

№ п/п	Содержание обучения, наименование и тематика практических занятий (вебинаров), самостоятельных работы	Объе м, ак.час .
	Модуль 1. Базовые алгоритмы.	74
1	Введение в ML	
	Чем заниматься в ML?	2

	Постановка задачи ML, типы разных постановок задач	2
	Типы данных	2
	Конвейер обработки данных, Crispr-dm	2
	Метрики качества и функция потерь	2
	Метрики качества задачи классификации и регрессии (их получение и интерпретация)	2
2	Предобработка	
	Проблемы и варианты загрузки	4
	Пример EDA	2
	Подготовка: стандартизация, полиномиальные признаки (что такое kernel trick)	2
	Метрики качества задачи классификации алгоритма KNN, перебор параметров алгоритмов с помощью sklearn	2
	Интерпретация метрик качества классификации	2
	Метрики задачи регрессии, перебор параметров алгоритма для максимизации метрик регрессии	4
3	Линейная регрессия	2
	Постановка задачи линейной регрессии	2
	- Постановка задачи линейной регрессии - Аналитическое решение -- переопределенная система линейных уравнений -- множественная регрессия -- линейные трансформации пространства -- визуализации матричных перемножений -- нормальные уравнения и визуализация аналитического решения -- реализация на numpy линейной и полиномиальной регрессии (проблема некорректной постановки задачи) и через sklearn	2

	- Функция правдоподобия -- Максимизация функции правдоподобия и условия теоремы Маркова-Гаусса -- Компромисс дисперсии-смещения -- Вывод аналитического решения через максимизацию функции правдоподобия (получение MSE) - Градиентный спуск на примере линейной регрессии -- Предел, секущая и касательная, метод конечных разностей, правила дифференцирования на простом примере, исследование функции на экстремумы. -- Вывод шагов градиентного спуска из MSE -- Реализация на numpy (стахостического) градиентного спуска	4
4.	Регуляризация и логистическая регрессия	
	- Компромисс смещения-разброса - Регуляризация -- Некорректно поставленная задача: регуляризация логистической регрессии - l1 и l2 регуляризации. -- Вывод с шагами градиентного спуска. -- Реализация на numpy и через sklearn. -- Отбор признаков с l1 регуляризацией	2
	- Логистическая регрессия -- Кросс-энтропия. Вывод шагов спуска логистической регрессии (связь с линейной регрессией). -- Визуализация отступа и logistic loss, связь с кросс-энтропией -- Принцип максимального правдоподобия и логистическая регрессия, сравнение с zero-one-loss -- Реализация на numpy логистической регрессии и перебор параметров через sklearn -- Регуляризация l1, l2 логистической регрессии, перебор параметров с помощью sklearn -- Полиномиальные признаки и логистическая регрессия	2
5.	SVM, работа с представлением, наивный Байес	

	- SVM -- Постановка задачи SVM, визуализация -- Вывод шагов hard-margin svm -- Hinge loss: регуляризация SVM. Связь с logloss и ReLU -- Ядра в SVM (теорема Мерсера) -- Реализация на numpy SVM, перебор параметров через sklearn	2
	- Что такое обучение представлению в ML, отбор признаков -- Варианты подготовки текста, TF-IDF -- Алгоритмы аугментации - Smote, adasyn, totem links -- Кривые обучения - как обнаружить нерепрезентативные данные -- Методы поиска выбросов	2
	- Теория вероятностей и наивный Байес -- Случайная величина, понятия вероятности и условной вероятности, (не)зависимые события, теорема Байеса -- Алгоритм наивного байеса, постановка задачи, проблемы и применения -- Варианты наивного Байеса и реализация на numpy. Модификации наивного байеса (tf-idf + фильтрация предлогов + bow с парами слов), перебор параметров на sklearn. Поддержка программы чтения с экрана включена.	2
6.	Методы понижения размерности	
	-Зачем нужны методы понижения размерности - Разложения -- Спектральное разложение -- Сингулярное разложение -- Оценка ранга и сжатие без потерь -- Проксимация матрицей меньшего ранга, норма Фробениуса -- Преобразование признаков -- Сингулярное разложение и низкоранговое приближение	2

	PCA -- Постановка задачи -- Реализация через матрицу ковариации -- Реализация через сингулярное разложение -- Выбор числа компонент -- Реализация через sklearn на примере понижения размерности tf-idf -- Недостатки PCA	2
	T-SNE -- Постановка задачи и отличие от PCA -- Функция потерь и перплексия. -- Недостатки T-SNE	2
	UMAP -- Сравнение с T-SNE -- Анализ топологических данных и упрощенные комплексы -- Низкоранговое представление -- Efficient K-Nearest Neighbor Graph Construction for Generic Similarity Measures	2
	- Работа с tensorboard projector	2
7.	Кластеризация	

	- Постановка задачи кластеризаций и сферы применения - Нормы - Типы алгоритмов - Метрики задачи кластеризации - Алгоритмы -- k-means / k-means++ / k-medians -- Mean shift -- Агломеративная кластеризация -- DBSCAN -- Affinity Propagation -- Спектральная кластеризация	2
	- Приложения алгоритмов кластеризации на примере FAISS - Реализация кластеризации через sklearn, сравнение алгоритмов	2
8.	Деревья и градиентный бустинг	
	- Дерево решений, отличие от остальных алгоритмов -- Принципы построения в задачах регрессии и классификации -- Работа с количественными признаками -- Комбинаторная энтропия, вывод энтропии по Шеннону, прирост информации. Реализация с нуля на питоне на игрушечном примере. -- Реализация через sklearn -- Недостатки алгоритма дерева	2
	- Ансамблевые методы -- Бэггинг -- Случайный лес, сверхслучайные деревья -- Оценка важности признаков -- Недостатки деревьев	2

	Градиентные методы -- Что было до: адабуст -- Алгоритм Фридмана. Градиентный бустинг для задачи регрессии, вывод шагов обновления для различных функций потерь -- Градиентный бустинг для задачи классификации, вывод шагов обновления для различных функций потерь	2
	- Работа с catboost и перебор моделей с tensorboard	2
	Модуль 2. Глубокое обучение	70
1	Введение в Pytorch	
	— тензоры и операции с ними, вычислительный граф ИНС и обратное распространение ошибки — полносвязная нейронная сеть, обратное распространение ошибки, загрузка данных, автодифференцирование на примере логистической регрессии Универсальные способы оптимизации моделей — оптимизаторы, батчинг, нормализации, функции активации, работа с моделями, pytorch lightning Оптимизация, часть 2 — Дистилляция, прунинг, квантизация, onnx	14
2	Машинное зрение, введение	

	— Основы CNN и типичные операции в сверточных сетях Машинное зрение, архитектуры — Виды архитектур и их реализация, дообучение Машинное зрение - детекция 2 — faster r-cnn, реализация Машинное зрение - детекция 1 — Yolo подобные системы, реализация Машинное зрение - сегментация — U-net Машинное зрение - GAN — stylegan 2, normalizing flow Машинное зрение в 3d — 3d ml, point cloud	14
3	Обработка текста, введение	
	— дистрибутивная семантика, w2v, lstm Обработка текста, lstm+seq2seq — задачи Seq2seq на примере задач перевода Обработка текста, attention, attention block — Механизм внимания, реализация на pytorch, визуализация Обработка текста, transformer — Реализация трансформер модели на pytorch Обработка текста, bert — Трансформеры как универсальная модель машинного обучения, трансформеры и Bert-ология Обработка текста, SOTA — huggin face, deepPavlov на примере разных задач NLP	14
4	Обработка временных рядов, классический подход	

	— эконометрические подходы Обработка временных рядов, глубокое обучение — lstm, cnn, transformer Обработка звука, sota — tacotron2, huggin face, asteroid	14
5	Обучение с подкреплением, введение	
	— online/offline learning, q-learning, policy-gradients Обучение с подкреплением, policy gradients — gym, road to PPO и RND Обучение с подкреплением, model-based — Alpha zero Обучение с подкреплением, transformer — upside-down rl, dicision transformer	14
	Итого	144

Раздел 3. Методическое обеспечение ДООП.

3.1. Формы аттестации и оценочные материалы:

Оценка качества освоения программы проводится по двухбалльной системе: «зачтено», «не зачтено» по результатам промежуточного контроля (тестирование, проверочные задания на взаимную оценку), контроля посещаемости практических занятий и результатам аттестации.

Обучающий считается аттестованным в случае положительных результатов работы (не менее 70% баллов от итоговой оценки) в процессе обучения и успешной защиты. При этом баллы за защиту проекта начисляются только при достижении 50% порога при прохождении каждого проверочного этапа. После аттестации школьник получает оценку “отлично”, если набрано не меньше 80% баллов от возможного максимума, “хорошо”, если набрано не меньше 65%, “удовлетворительно”, если набрано не меньше 40%, в противном программа не пройдена.

Результат тестирования, решения проверочных заданий и написания кода проверяется автоматически системой на образовательной платформе. Проектная работа проверяется преподавателем и ассистентами.

Составляющие процесса обучения, которые оцениваются в ходе обучения, и их вклад в итоговую оценку представлены в таблице 1.

Таблица 1 – Составляющие процесса обучения

	Основные показатели оценки	Вклад в итоговую оценку
1	Посещения занятий	50%
2	Практические занятия	10%
3	Итоговая аттестация	40%

Оценочные материалы

Пример тестового задания

Из чего можно вывести основные метрики задачи бинарной классификации?

F1

Recall (Полнота)

Precision (Точность)

Confusion matrix (Матрица ошибок)

Ответ: Confusion matrix

3.2. Материально-техническое обеспечение программы.

Оборудованные учебные кабинеты, лаборатории, актовые залы и коворкинги, конференц-залы. Технические средства, используемые в образовательном процессе: компьютерное и мультимедийное оборудование, видео-, аудиовизуальные средства, специальное профессиональное оборудование для реализации учебных проектов.

Наименование специализированных аудиторий, кабинетов, лабораторий	Вид занятий	Наименование оборудования, программного обеспечения
--	--------------------	--

Аудитории	Лекции	Электронная доска, компьютеры.
Информационно-коммуникационная платформа дистанционных семинаров	Практические занятия	Слушателю необходимо наличие доступа в сеть интернет, компьютер. Преподавателю курса необходимо оборудование.
Система дистанционного обучения провайдера массовых открытых онлайн курсов	Самостоятельная работа	Наличие компьютера и доступа в сеть интернет.

3.3. Кадровое обеспечение программы.

Программа может быть реализована педагогом дополнительного образования с уровнем образования и квалификации, соответствующим Профессиональному стандарту «Педагог дополнительного образования детей и взрослых».

3.4. Список литературы:

Основная литература

1. Прикладная и компьютерная лингвистика / Под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. — М.: URSS, 2016. — 320 с.

Дополнительная литература

1. Murphy K.P. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
2. Sheldon A. Linear Algebra Done Right. Springer, 2015.
3. DasGupta A. Probability for statistics and machine learning. Springer, 2011.
4. Умнов. Аналитическая геометрия и линейная алгебра (2011) — МФТИ.
5. Нестеров. Методы выпуклой оптимизации (2010)

6. Diez, Barr, Çetinkaya-Rundel, Dorazio. Advanced High School Statistics (2015)
7. DasGupta. Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics (2011)
8. Vance Martin, Stan Hurn, David Harris. Econometric Modelling with Time Series.
9. Specification, Estimation and Testing. Cambridge University Press, 2013.
10. Lütkepohl, H., 2005, New introduction to multiple time series analysis. Springer Science & Business Media.
11. Box, G.E.P., Jenkins G.M. and G.C. Reinsel, 2008, Time Series Analysis: Forecasting and Control (Wiley Series in Probability and Statistics), Wiley Publ., 4th ed.
12. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
13. Введение в информационный поиск / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. — М.: Вильямс, 2011. — 528 с.
1. Natural Language Processing for the Working Programmer / Daniël de Kok, 2011.