Министерство науки и высшего образования РФ  
ФГБОУ ВО «Нижневартовский государственный университет»   
Факультет информационных технологий и математики

Кафедра информатики и методики преподавания информатики

Курсовой проект по дисциплине

«Основы научно-исследовательской деятельности»

АНАЛИЗ СТАТИСТИКИ СТРАХОВЫХ ВЫПЛАТ СРЕДСТВАМИ ЯЗЫКА PYTHON

Исполнитель:

студент группы 3353

Горидько Иван Александрович

Руководитель:

кандидат технических наук,

доцент кафедры информатики и методики

преподавания информатики

Катермина Татьяна Сергеевна

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

(подпись)

Нижневартовск, 2025

Оглавление:

[Глава 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПРОГНОЗИРОВАНИЯ СТРАХОВЫХ ВЫПЛАТ СРЕДСТВАМИ МАШИННОГО ОБУЧЕНИЯ 5](#__RefHeading___4)

[1.1. Коротко про саму задачу предсказания страховых выплат 5](#__RefHeading___8)

[1.2. Уточненная постановка задачи 5](#__RefHeading___7)

[1.3. Описать библиотеки, язык и среду разработки 5](#__RefHeading___6)

[1.4. Первичный разведочный анализ данных (EDA) 6](#__RefHeading___5)

**Анализ статистики страховых выплат средствами языка Python**

**Введение**

**Актуальность исследований.** В современном мире страхование является неотъемлемой частью финансовой стабильности как отдельных людей, так и целых предприятий. Страховые выплаты составляют значительную долю в расходах компаний, и их точное прогнозирование напрямую влияет на прибыльность и устойчивость страхового бизнеса. Машинное обучение и интеллектуальный анализ данных стали мощными инструментами для решения этой задачи, позволяя выявлять сложные, неочевидные для человека взаимосвязи в исторических данных. Язык Python с его богатой экосистемой библиотек, таких как Scikit-learn, Pandas и NumPy, является одной из самых популярных и эффективных сред для реализации подобных аналитических систем. Прогнозирование страховых выплат — это сложная и многогранная задача. Для ее успешного решения необходимо разбить процесс на ключевые этапы: от сбора и предобработки данных до построения, валидации и интерпретации прогнозных моделей.

**Объект исследования:** Формирование размеров страховых выплат на основе статистических данных.

**Процесс исследования:** Формирование размеров страховых выплат средствами интеллектуального анализа данных.

**Цель исследования:** Реализовать предсказания страховых выплат средствами языка python.

**Задачи исследования:**

* Провести анализ и предобработку предоставленного набора данных о страховых случаях.
* Реализовать и обучить модель линейной регрессии для прогнозирования размера страховых выплат.
* Реализовать и обучить модель регрессии на основе алгоритма "Случайный лес".
* Провести сравнительный анализ эффективности построенных моделей и выбрать оптимальную.

# Глава 1. **ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПРОГНОЗИРОВАНИЯ СТРАХОВЫХ ВЫПЛАТ СРЕДСТВАМИ МАШИННОГО ОБУЧЕНИЯ**

## 1.1. Коротко про саму задачу предсказания страховых выплат

Задача предсказания страховых выплат относится к классу задач прогнозной аналитики и машинного обучения с учителем, а именно — к задачам регрессии.

Её основная цель заключается в построении математической модели, способной на основе исторических данных о клиентах (таких как возраст, индекс массы тела, вредные привычки, регион проживания и другие демографические признаки) прогнозировать непрерывную числовую величину — ожидаемый размер страховой выплаты.

Точное прогнозирование позволяет страховым компаниям оптимизировать тарифную политику, более эффективно управлять рисками и финансовыми резервами, а также выявлять скрытые закономерности, влияющие на стоимость страховых случаев.

## 1.2. Уточненная постановка задачи

В рамках данной работы решается задача множественной линейной регрессии, где целевой переменной (зависимой переменной) является столбец charges (размер страховых выплат), а признаками (независимыми переменными) — все остальные столбцы в наборе данных: age, sex, bmi, children, smoker, region.

Категориальные признаки (такие как sex, smoker, region, а также производный признак weight\_category) перед обучением модели подвергаются процедуре One-Hot Encoding.

Процесс решения задачи включает в себя следующие этапы: первичный разведочный анализ данных (EDA), предобработка данных (обработка пропусков, кодирование категориальных переменных, нормализация числовых признаков), разделение данных на обучающую и тестовую выборки, обучение и валидация моделей, а также сравнительный анализ их эффективности по метрикам MSE, MAE и R².

## 1.3. Описать библиотеки, язык и среду разработки

Работа выполнена на языке программирования Python версии 3.x, который был выбран благодаря его широкой распространённости в задачах анализа данных и машинного обучения, а также наличию мощного стека специализированных библиотек.

В качестве основной среды разработки использовалась Jupyter Notebook, что позволяет интерактивно выполнять код, визуализировать данные и документировать ход исследования.

Для решения поставленных задач были применены следующие библиотеки: Pandas — для манипуляций с данными и их загрузки; NumPy — для выполнения численных операций; Matplotlib и Seaborn — для визуализации данных и построения графиков; Scikit-learn — для предобработки данных, реализации алгоритмов машинного обучения (Linear Regression, Random Forest), их обучения и оценки качества.

## 1.4. Первичный разведочный анализ данных (EDA)