

ASNI-MedHome

- Dataset
- Dataframe
- Exploratory Data
- Feature Engineering
- Pipeline
- Model Building
- EDA Reports
- Clinical Reports
- Scientific Reports
- Statistics Samples
- ML Samples
- User private codes

Asfendijarov Kazakh National Medical University

Автоматизированная Система Научных Исследований в медицине и здравоохранении «АСНИ-МЕД»

Система для любознательных и настойчивых



Добро пожаловать в АСНИ-МЕД!

АСНИ-МЕД - это система с открытым исходным кодом, помогающее пользователям применять избранные методы машинного обучения такие как поиск данных, предварительная обработка и построение моделей, создание научных отчетов и отчетов о проведении анализа клинических исследований (Clinical trials), другие процедуры. Самая важная в этой системе заключается в том, что вы можете делать всё это БЕЗ программирования!

Информация об АСНИ-МЕД

Краткое описание системы

Содержание

[Препамбула](#)


[Цель создания системы](#)


[Назначение системы](#)


[Область применения системы2](#)


×


ASNI-MedHome


 Dataset


 Dataframe


 Exploratory Data


 Feature Engineering


 Pipeline


 Model Building


 EDA Reports

 Clinical Reports

 Scientific Reports

 Statistics Samples

 ML Samples

 User private codes

Информация об АСНИ-МЕД

▼

Собрание избранных статей из области биостатистики и DataScience

▲

Собрание избранных статей из области биостатистики и Data Science

Внимание: Статьи окончательно не упорядочены по рубрикам!

Эти статьи дают общие сведения о вышеназванных областях и могут быть хорошим справочным материалом для различных категорий пользователей.

Приведенный в статьях программный код может быть адаптирован и использован в учебных целях и для реализации научных пароектов.

Если в статье есть ссылка на GitHub Repositorium, то было бы полезно клонировать этот Repositorium на персональный компьютер и использовать его в своих учебных и научных целях.

Basic Fundamentals of Statistics and Data Science

- [Basic Fundamentals of Statistics for Data Science](#)
- [Mastering Statistical Analysis with Statsmodel Library \(with Code\)](#)
- [Einstieg in das Maschinelle Lernen mit Python\(x,y\)](#)

Probability

- [Understanding Different Probability Distributions with Real-World Examples](#)
- [Probability Theory: Bayes Theorem](#)
- [From theory to practice: Harnessing probability for effective data science](#)
- [Use Python to Calculate Probabilities](#)
- [P-value Explained Clearly — Regression, PDF, Discrete](#)
- [Understanding Probability: 7 Essential Concepts for Data Science with Python Examples](#)













Correlation

- [Correlation & Covariance: Measure of Association](#)
- [Cohort Analysis](#)
- [Understanding Hypothesis Testing: One-sample, Two-sample, and Paired t-tests with Real-world Examples](#)
- [+++ Mastering Linear Regression with Statsmodels](#)
- [Linear Regression: Multicollinearity, they taught you wrong](#)
- [Multicollinearity Problems in Linear Regression. Clearly Explained!](#)
- [Case Study Interview Questions on Statistics for Data Science](#)
- [Detecting Trends in Time Series Data using Python](#)

ROC



ASNI-MedHome

-  Dataset
-  Dataframe
-  Exploratory Data
-  Feature Engineering
-  Pipeline
-  Model Building
-  EDA Reports
-  Clinical Reports
-  Scientific Reports
-  Statistics Samples
-  ML Samples
-  User private codes

[seaborn.html](#)

Multivariate Analysis

- [Air Quality Data- Statistical Analysis](#)
- [Data Science Metrics Explained With Strawberries](#)
- [Библиотека на Python, которая поможет вам с гипотезами и статистическими моделями](#)
- [Model Selection for Linear Regression](#)
- [How to run and interpret MULTIPLE regression models in R: quick guide with real-world economic data](#)
- [Linear Regression: Multicollinearity, they taught you wrong](#)
- [Optimizing Logistic Regression: Finding the Optimal Decision Threshold for Enhanced Predictions in Fintech](#)
- [Power Consumption forecasting with time series data — End-to-end Machine Learning Project using Python](#)
- [10 Most Common Machine Learning Algorithms Explained -2023](#)
- [Best Tips and Tricks: When and Why to Use Logarithmic Transformations in Statistical Analysis](#)
- [Customer Segmentation using RFM Analysis and K-Means Clustering in Python](#)
- [Learn Logistic Regression for Classification with Python: 10 Practical Examples](#)
- [The Stratified Cox Proportional Hazards Regression Model](#)
- [Modelling Binary Logistic Regression using Tidymodels Library in R \(Part-1\)](#)
- [Practical Process to Implement Hierarchical Clustering with Python](#)
- [Unraveling the Mysteries of Quantile Regression: A Comprehensive Analysis and Python Implementation](#)
- [Introduction to Linear Regression Model in Exploratory](#)
- [Multicollinearity Problems in Linear Regression, Clearly Explained!](#)
- [Statistical functions \(scipy.stats\)](#)
- [Statistical concepts that every Data Scientist should know](#)
- [Understanding Hypothesis Testing: One-sample, Two-sample, and Paired t-tests with Real-world Examples<https://medium.com/@dancerworld60/understanding-hypothesis-testing-one-sample-two-sample-and-paired-t-tests-with-real-world-75d99bc9cc95>](#)
- [From theory to practice: Harnessing probability for effective data science](#)
- [Ways to improve K-means clustering](#)
- [Multivariate Analysis using SAS](#)

Machine Learning

- [10 Machine Learning Algorithms that will DOMINATE 2023](#)
- [The Complete Guide to Time Series Models](#)
- [Linear Discriminant Analysis \(LDA\)](#)
- [An Introduction to Classification in Machine Learning](#)

Graphs

- [Generate Publication-Ready Plots Using Seaborn Library Part-1](#)
- [Plotly: Data Visualization Comprehensive Guide](#)
- [Adventures in Plotly: Scatter Plots](#)

ASNI-MedHome

Dataset

Dataframe

Exploratory Data

Feature Engineering

Pipeline

Model Building

EDA Reports

Clinical Reports

Scientific Reports

Statistics Samples

ML Samples

User private codes

Что такое исследовательский анализ данных и зачем он нам нужен?

Что такое исследовательский анализ данных и зачем он нам нужен?

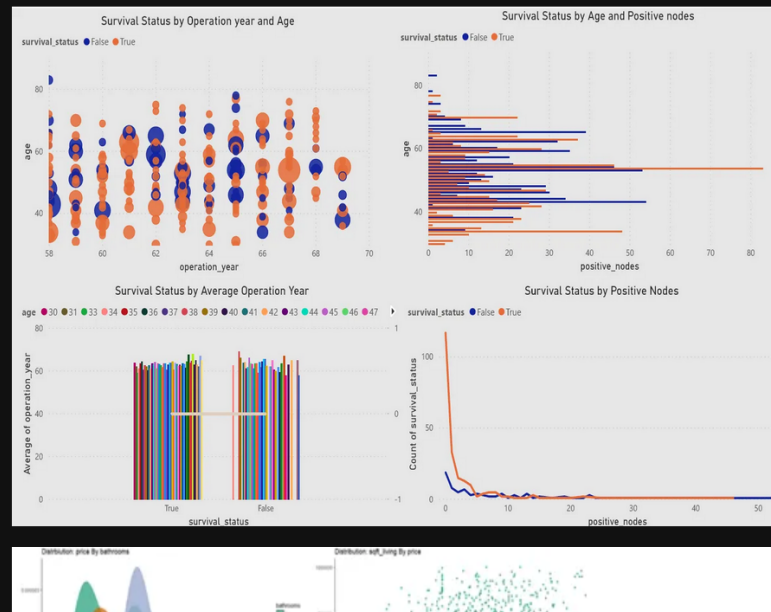
В чем заключается концепция?

Исследовательский анализ данных - это набор методов, которые были в основном разработаны Джоном Тьюки, Джоном Уайлдером с 1970 года. Философия этого подхода заключается в изучении данных перед применением конкретной вероятностной модели. По словам Джона Тьюки и Джона Уайлдера, исследовательский анализ данных похож на детективную работу.

Исследовательский анализ данных (EDA) был предложен Джоном Тьюки, чтобы побудить статистиков изучить данные и, возможно, сформулировать гипотезы, которые могли бы привести к сбору новых данных и экспериментам.

“Величайшая ценность картины – это когда она заставляет нас заметить то, чего мы никогда не ожидали увидеть” – Джон Тьюки

EDA отличается от анализа исходных данных (IDA), который более узко фокусируется на проверке допущений, необходимых для подгонки модели и проверки гипотез, а также на обработке недостающих значений и выполнении преобразований переменных по мере необходимости. EDA включает в себя IDA.





ASNI-MedHome

Dataset

Dataframe

Exploratory Data

Feature Engineering

Pipeline

Model Building

EDA Reports

Clinical Reports

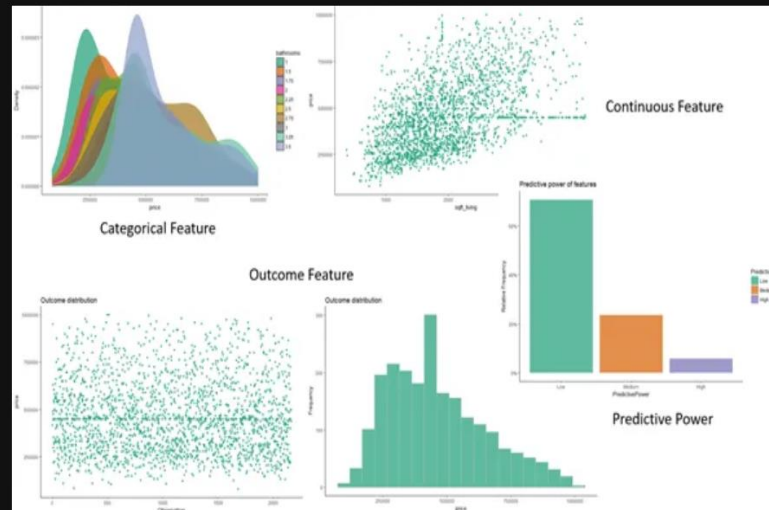
Scientific Reports

Statistics Samples

ML Samples

User private codes

Deploy



“Это моя любимая часть аналитики: брать скучные плоские данные и воплощать их в жизнь с помощью визуализации “ - Джон Тьюки

Потребность в Исследовательском анализе данных:

- 1. Визуально проанализируйте набор данных, чтобы получить представление о данных
- 2. Использование множества методов для того, чтобы сделать определенные выводы из полученных данных
- 3. Помогает нам в выборе функций, поскольку мы можем определить важные функции
- 4. Затем эти функции будут использованы для построения моделей машинного обучения
- 5. Цель состоит в том, чтобы определить, является ли прогнозирующая модель жизнеспособным аналитическим инструментом для конкретной бизнес-задачи, и если да, то какой тип моделирования наиболее подходит
- 6. Это помогает при передаче результатов нетехническому специалисту
- 7. Пропуск шага EDA может привести к неточному созданию модели

В качестве примера EDA нами используется набор данных Хабермана о выживаемости раковых пациентов

Информация о наборе данных:

Описание: Набор данных содержит случаи из исследования, которое проводилось в период с 1958 по 1970 год в больнице Биллингса Чикагского университета по изучению выживаемости пациенток, перенесших операцию по поводу рака молочной железы.

Ссылка на набор данных: <https://www.kaggle.com/gilsousa/habermans-survival-data-set> Artikel: <https://aniketpatilvashi.medium.com/what-is-exploratory-data-analysis-and-why-we-need->



ASNI-MedHome

Dataset

Dataframe

Exploratory Data

Feature Engineering

Pipeline

Model Building

EDA Reports

Clinical Reports

Scientific Reports

Statistics Samples

ML Samples

User private codes

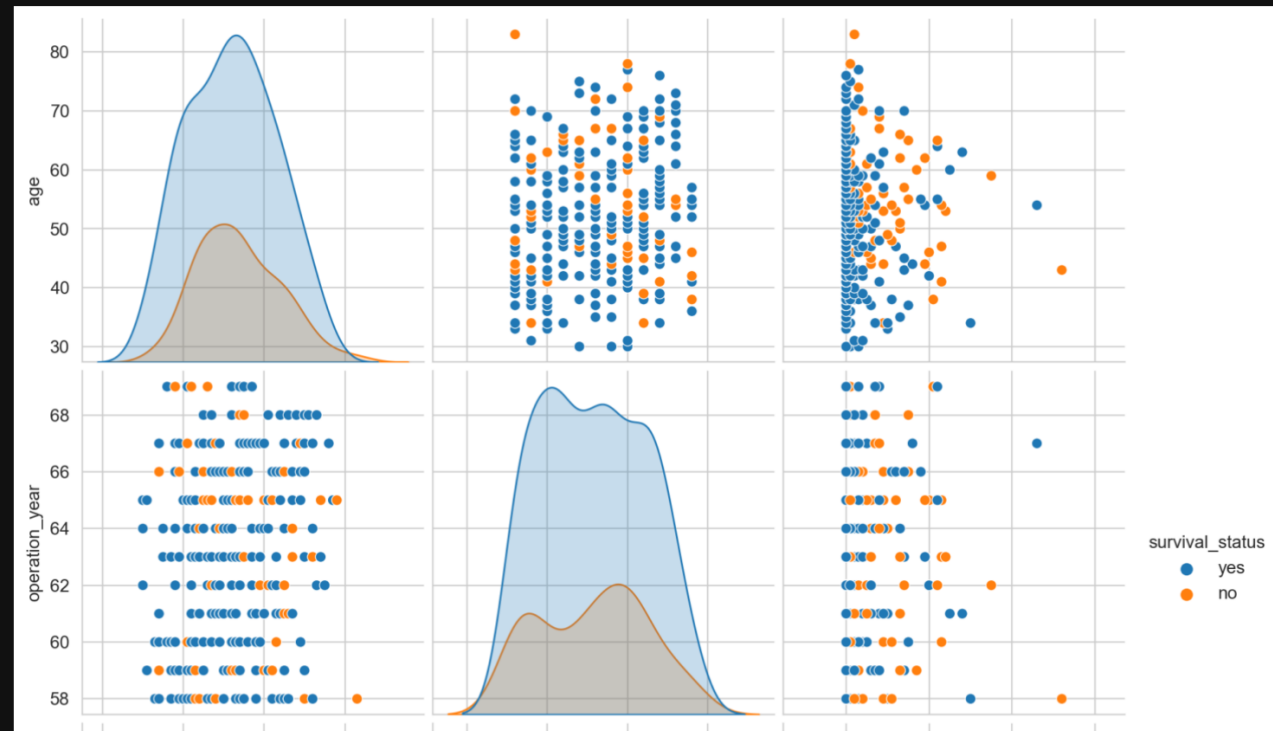
Существуют различные возрастные группы, которым проводится операция

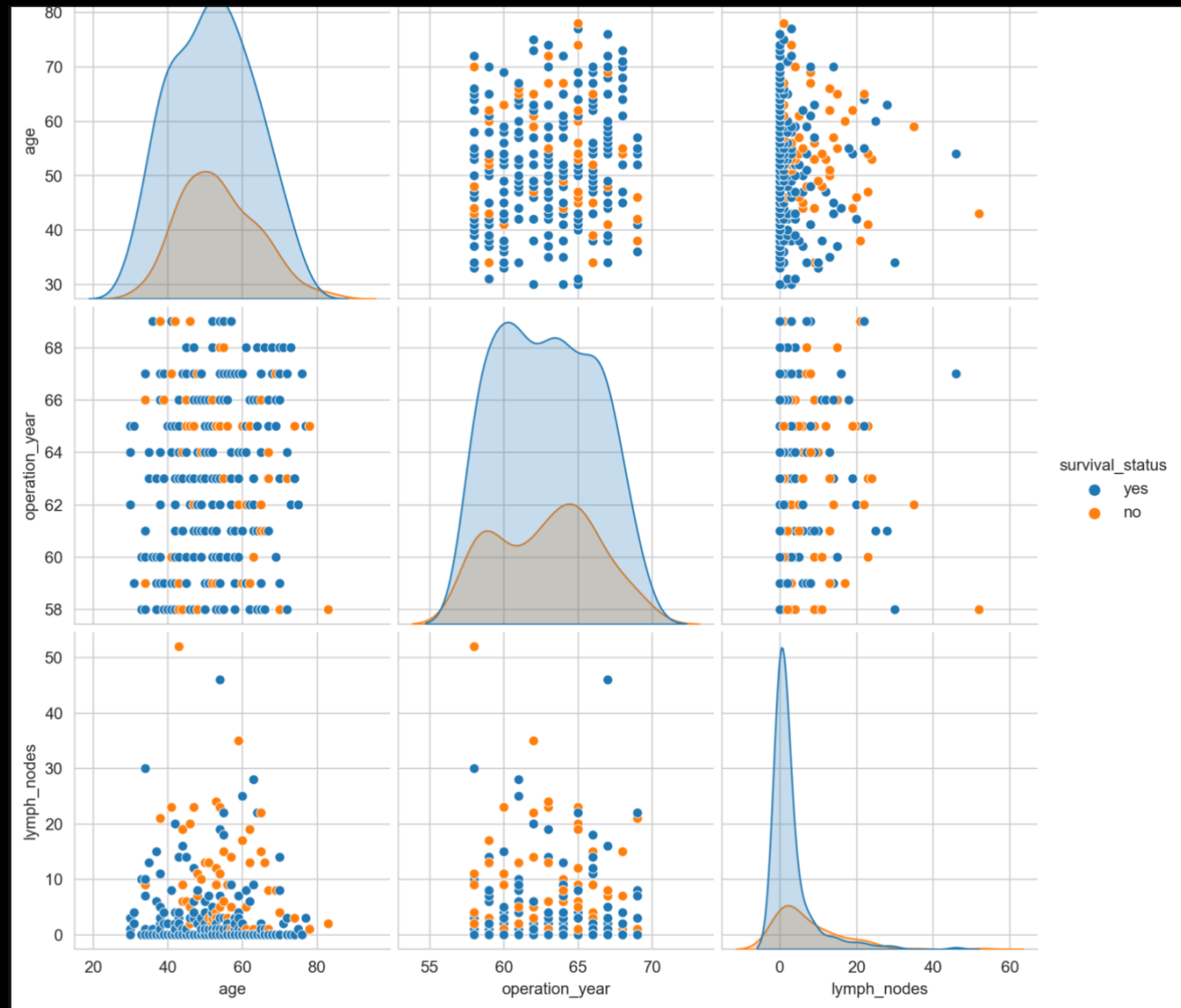
большинство операций было проведено в возрастной группе 52 года со стандартным отклонением в 3 года

МНОГОМЕРНЫЙ АНАЛИЗ (MULTIVARIATE ANALYSIS)

По сути, многомерный анализ - это инструмент для поиска закономерностей и взаимосвязей между несколькими переменными одновременно. Он позволяет нам предсказать, какое влияние изменение одной переменной окажет на другие переменные. Это дает многомерному анализу решающее преимущество перед другими формами анализа.

```
# First lets do multivariate analysis which will help us in univariate analysis
# Pair plots
sns.set_style("whitegrid");
sns.pairplot(df, hue="survival_status", height=3);
plt.show()
```





Медиана положительных значений для выживших пациентов равна нулю. Это центральная величина, которую мы выполняем EDA, удаляя выбросы для выживших пациентов.

Violin Plots

Violin Графики - это комбинация Box Plots и функции плотности вероятности (CDF).

Замечание:

Пациенты с более чем 1 узлом имеют меньшие шансы на выживание. Чем больше узлов, тем меньше шансов на выживание.

У большого процента выживших пациентов было 0 узлов. Тем не менее, небольшой процент пациентов, у которых не было положительных подмышечных узлов, умерли в течение 5 лет после операции, таким образом, отсутствие положительных подмышечных узлов не всегда гарантирует выживание.

Было сравнительно больше людей, которые были прооперированы в 1965 году и не прожили более 5 лет.

В возрастной группе от 45 до 65 лет было сравнительно больше людей, которые не выжили. Сам по себе возраст пациента не является важным параметром при определении выживаемости пациента.

Графики box и violin для параметров age и year дают аналогичные результаты со значительным перекрытием точек данных. Перекрытие на графике прямоугольника и графике скрипки узлов меньше по сравнению с другими характеристиками, но перекрытие все еще существует, и поэтому трудно установить пороговое значение для классификации обоих классов пациентов. 3D-график: 2D График плотности, Контурный график:

Заключительные замечания:

Нормальное распределение не может быть эффективно использовано для классификации, поскольку большее или меньшее количество пациенток из групп того же года пережили или не пережили операцию на молочной железе. Мы можем наблюдать пик смертности в 1964 году. Мы также можем видеть, что выживаемость со временем снижается.

Давайте сравним приведенный выше пункт с "уровнем выживаемости" или "смертностью", поскольку мы не можем зависеть только от данных о выживаемости или смертности:

Коэффициент выживаемости имеет незначительную тенденцию к снижению. Является ли это следствием несбалансированности данных? Нет, он нормализован. Что может быть причиной этого? Не наша цель.

Positive_nodes определенно указывает на наличие рака, Но это подразумевает, что операции на молочной железе не очень полезны.

Это не может быть сильно несбалансированным набором данных, поскольку отсутствие выживаемости составляет более половины показателя выживаемости

Поскольку это не несбалансированный набор данных, из этого набора данных можно сделать различные выводы как с помощью EDA, так и с помощью IDA Заключение:

Вы можете диагностировать рак у пациентов, используя набор данных Хабермана, применяя различные методы анализа данных и используя различные библиотеки Python.

