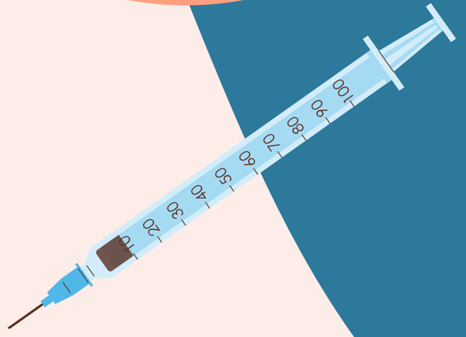
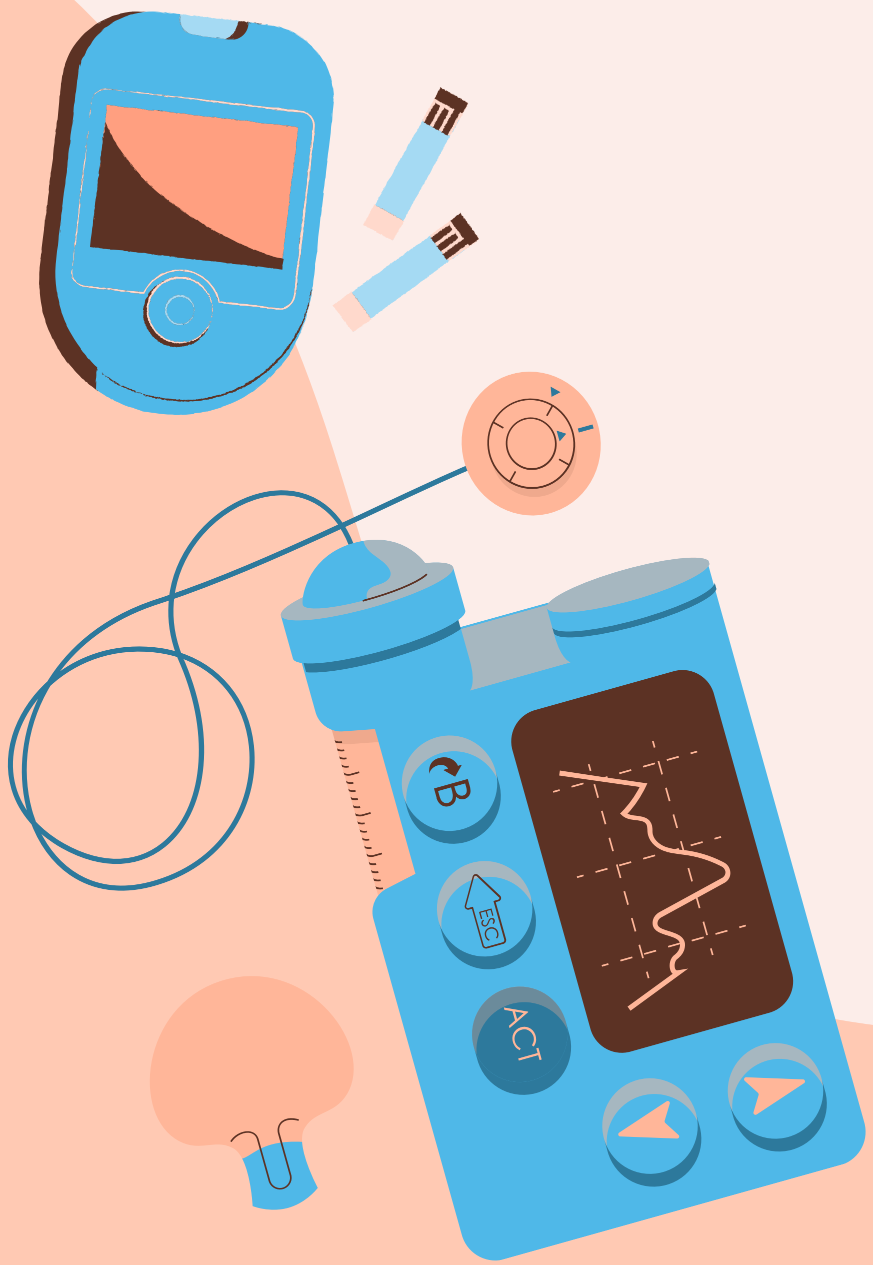
DIABETES PREDICTION

Group 4





The team

Nguyễn Lan Nhung

Trần Danh Tường

Trần Thị Thu Trang

Nguyễn Thị Thanh Hiền

Part A
Literature

01. Data Science in Healthcare

02. Medical theory about diabetes

03. Methods and models

Part B
Dataset

01. Overview

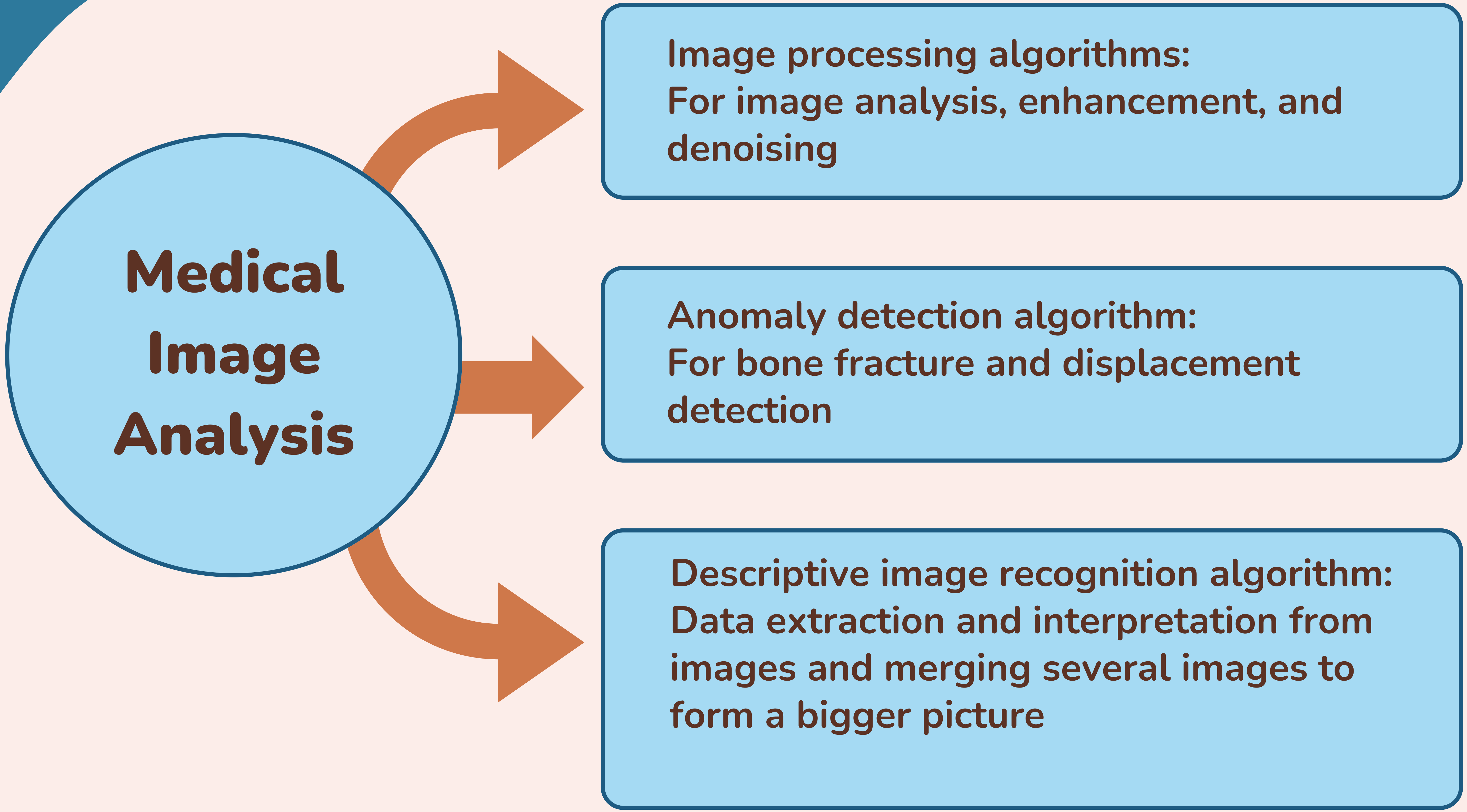
02. Cleaning

03. Modeling

04. Comparing and conclusion

DATA SCIENCE IN HEALTHCARE

→



Medical Image Analysis

Image processing algorithms:
For image analysis, enhancement, and denoising

Anomaly detection algorithm:
For bone fracture and displacement detection

Descriptive image recognition algorithm:
Data extraction and interpretation from images and merging several images to form a bigger picture

Genomic Data Science

Genomic data science plays a crucial role in advancing precision medicine, where healthcare interventions are tailored to an individual's unique genetic makeup

Discovering Drugs

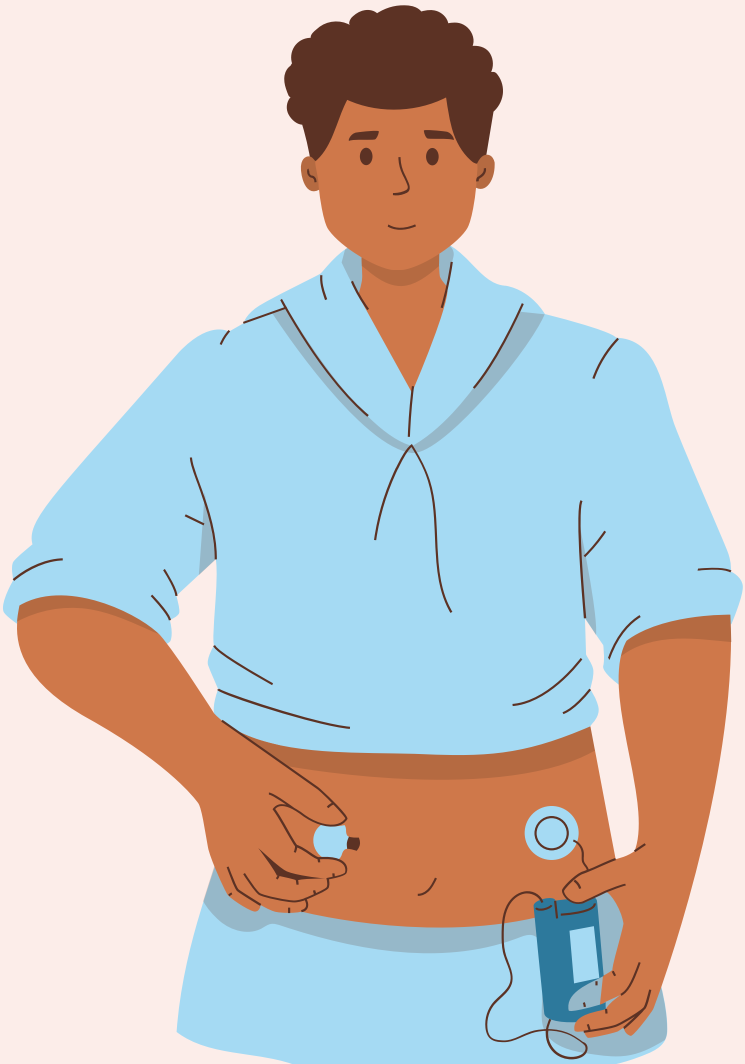
Providing the groundwork for the synthesis of drugs using Artificial Intelligence

Predictive Analytics in Healthcare

Predictive models in Data Science correlate and associate every data point to symptoms, habits, and diseases

- The identification of a disease's stage
- The extent of damage
- An appropriate treatment measure

- Manage chronic diseases
- Monitor and analyze the demand for pharmaceutical logistics
- Predict future patient crises
- Deliver faster hospital data documentation



Tracking Patient Health

- Developing wearable devices for patients that allow doctors to collect most of this data like heart rate, sleep patterns, blood glucose, stress levels, and even brain activity

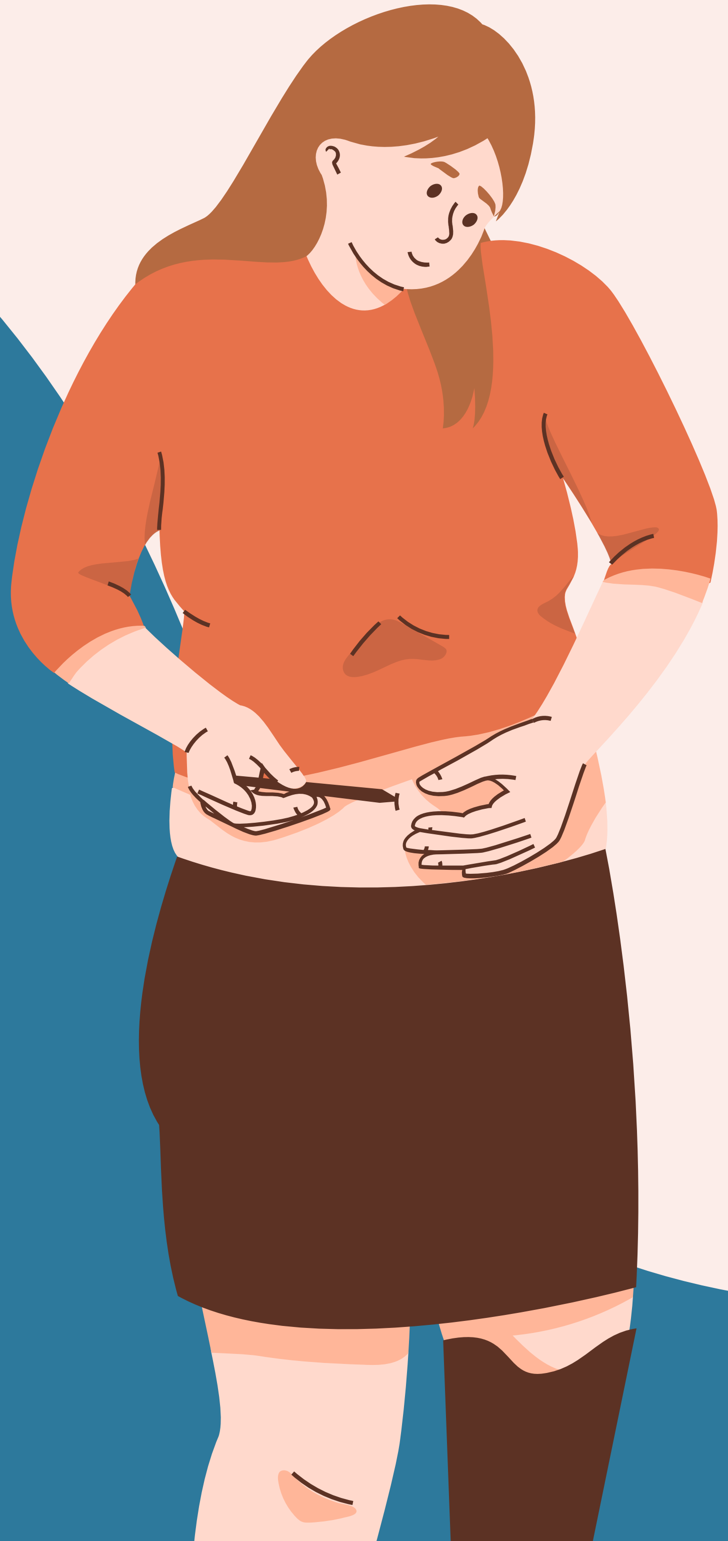
- Doctors can detect and track common conditions, like cardiac or respiratory diseases

- Detecting the slightest changes in the patient's health indicators and predict possible disorders

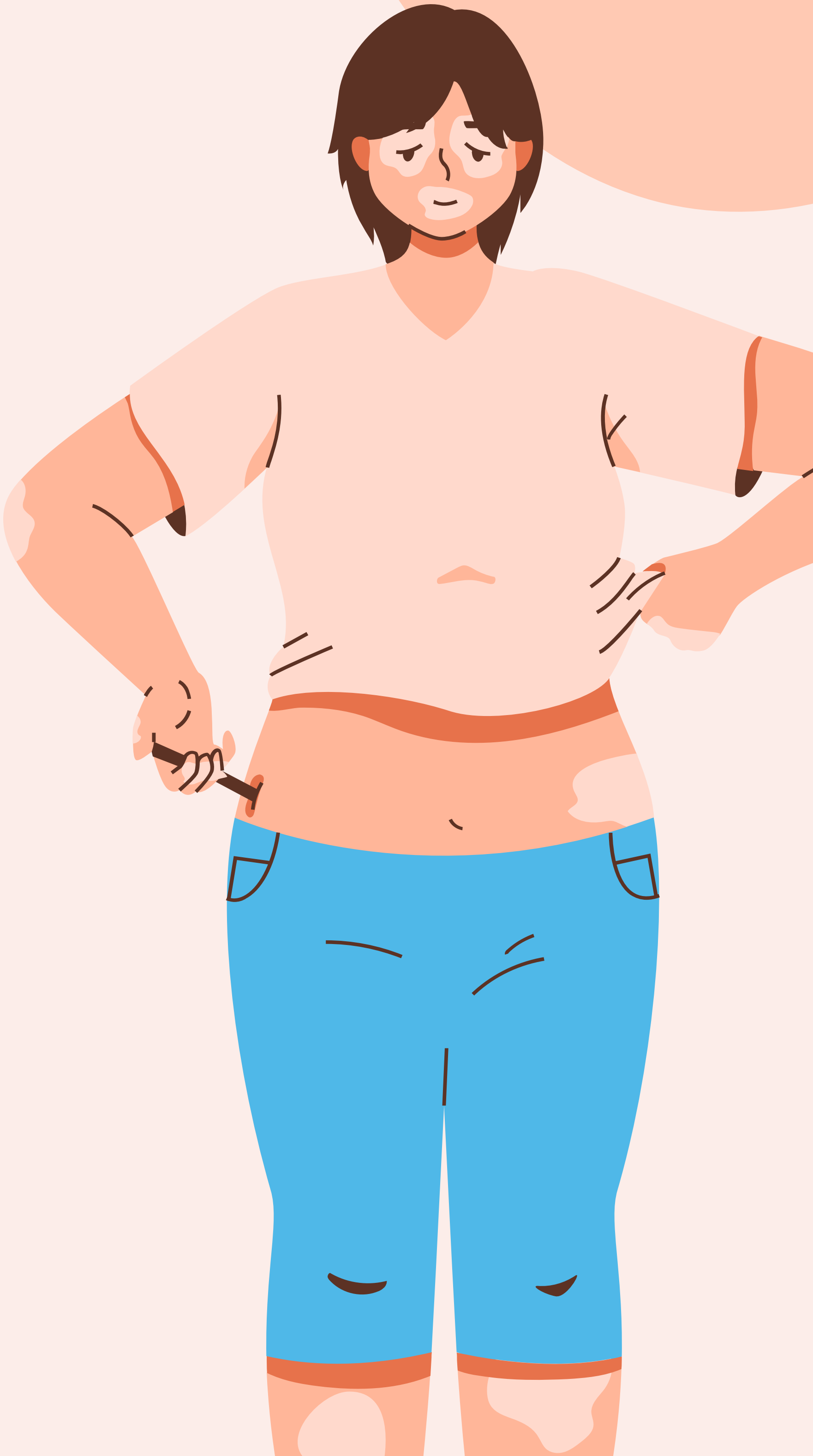
MEDICAL THEORY ABOUT DIABETES

→

WHAT IS DIABETES?



Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood glucose.



Type 1

- The body's immune system attacks and destroys the insulin-producing beta cells in the pancreas => Little/no insulin
- Often diagnosed in children and young adults
- Require lifelong insulin therapy

Type 2

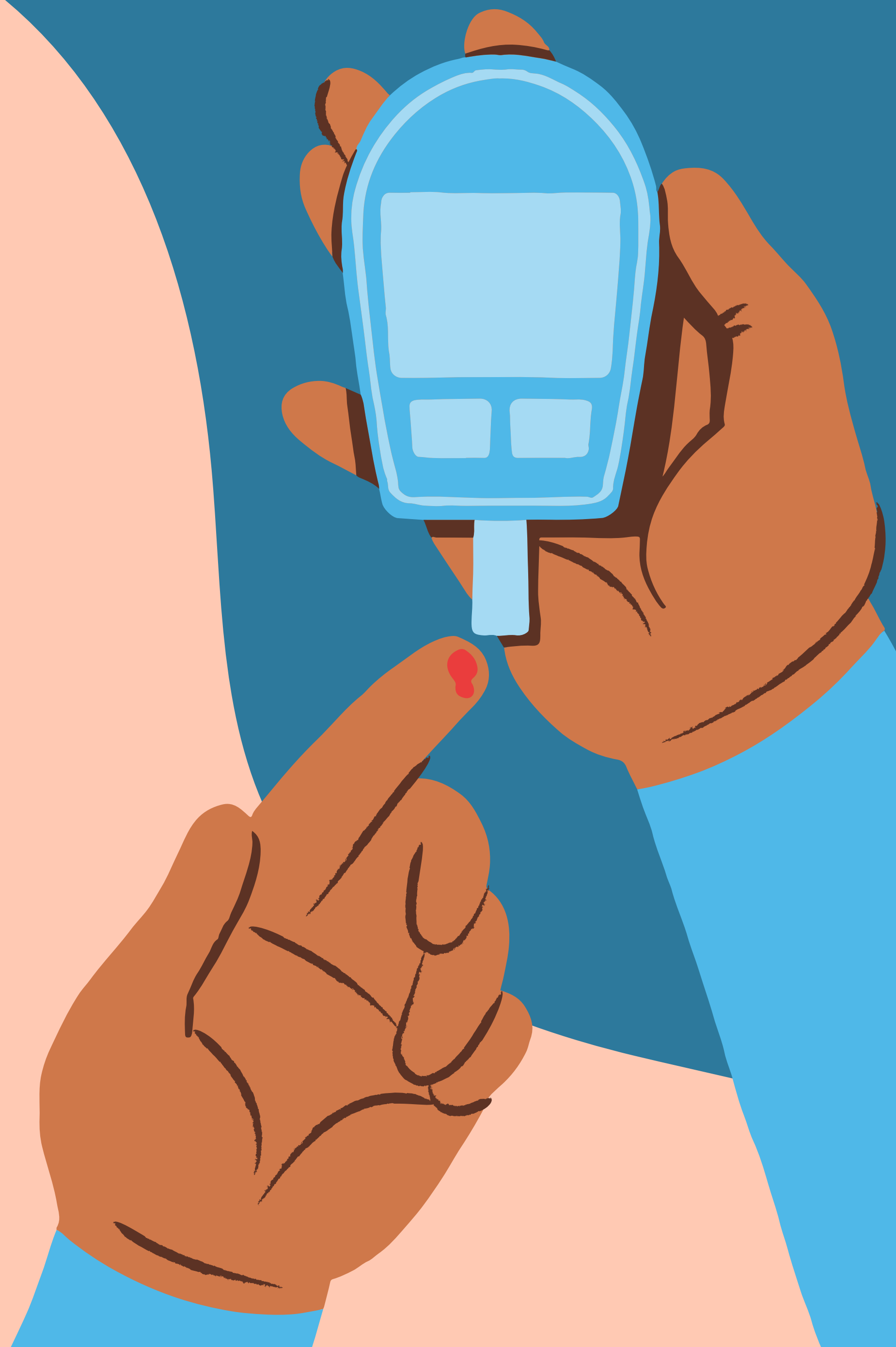
- Insulin resistance: where the body's cells do not respond effectively to insulin, not enough insulin to maintain normal blood glucose levels
- Can be influenced by genetics, lifestyle, obesity
- Often diagnosed in adults, but now more prevalent in children and adolescents

Gestational Diabetes

- Hyperglycemia with blood glucose values above normal but below the diagnostic of diabetes during pregnancy
- Usually resolves after delivery
- Increases the risk of developing Type 2 diabetes later in life for both the mother and child

WHAT ARE THE SIGNS AND SYMPTOMS OF DIABETES?

- Increased thirst
- Frequent urination
- Unexplained weight loss
- Fatigue
- Blurred vision
- Slow healing of wounds

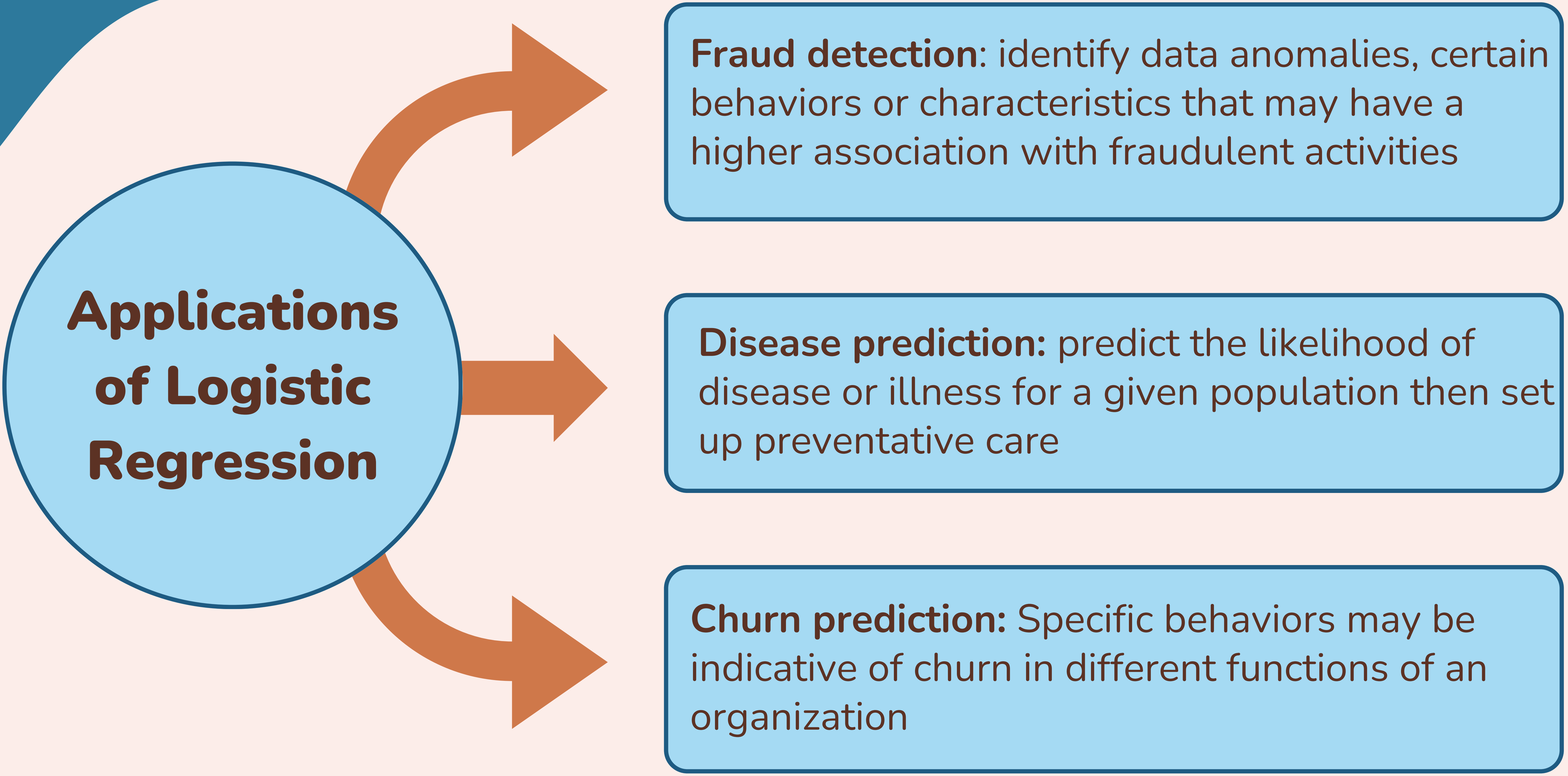


METHODS AND MODELS

LOGISTIC REGRESSION

Logistic regression estimates the probability of an event occurring based on a given dataset of independent variables.

In logistic regression, a logit transformation is applied to the odds—that is, the probability of success divided by the probability of failure.



Applications of Logistic Regression

Fraud detection: identify data anomalies, certain behaviors or characteristics that may have a higher association with fraudulent activities

Disease prediction: predict the likelihood of disease or illness for a given population then set up preventative care

Churn prediction: Specific behaviors may be indicative of churn in different functions of an organization

KNN

K-Nearest Neighbors Classifier

The KNeighborsClassifier is a classification algorithm in machine learning that belongs to the category of instance-based or lazy learning models. It makes predictions based on the majority class of the k-nearest neighbors of a data point.

KNN

Pros and Cons

Pros of KNN

- Doesn't require an explicit training phase.
- Suitable for both classification and regression tasks.
- Easy to understand and implement. The underlying principle of classifying based on proximity is straightforward.

Cons of KNN

- Highly dependent on the choice of the hyperparameter 'k.'
- Calculating distances between the new data point and all training instances can be computationally expensive, especially for large datasets.
- Sensitive to irrelevant or redundant features.

Naive Bayes Classifier

- The Naïve Bayes classifier is a popular supervised machine learning algorithm used for classification tasks such as text classification, statistics.
- This approach is based on the assumption that the features of the input data are conditionally independent given the class, allowing the algorithm to make predictions quickly and accurately.

Application of Naive Bayes Classifier

01 Real-time Prediction: Naive Bayesian classifier is an eager learning classifier and it is super fast.

02 Multi-class Prediction: This algorithm is also well known for its multi-class prediction feature.

03 Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayesian classifiers have a higher success rate as compared to other algorithms.

04 Recommendation System: Naive Bayes Classifier and Collaborative Filtering together build a Recommendation System to filter unseen information and predict whether a user would like a given resource or not.

SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised machine learning algorithm used for both classification and regression

The main objective is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space

Advantages

- High stability due to dependency on support vectors and not the data points.
- Does not get influenced by Outliers.
- No assumptions were made about the datasets.
- Numeric prediction problems can be dealt with by SVM.

Disadvantages

- Blackbox method.
- Inclined to the overfitting method.
- Very rigorous computation.

DECISION TREE

Decision tree is a flowchart-like tree structure where each internal node denote the feature, branches denote the rules and the leaf nodes denote the result of the algorithm.

- Can be prone to overfitting, especially when the tree grows deep and becomes overly complex => pruning, limiting tree depth
- Might not perform as well as more sophisticated models on complex datasets, particularly when dealing with high-dimensional data or situations with intricate decision boundaries

They serve as a fundamental building block in many machine learning algorithms and are valuable for their transparency and ease of implementation.

RANDOM FOREST

Random Forest is an ensemble learning method based on decision tree classifiers. It operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

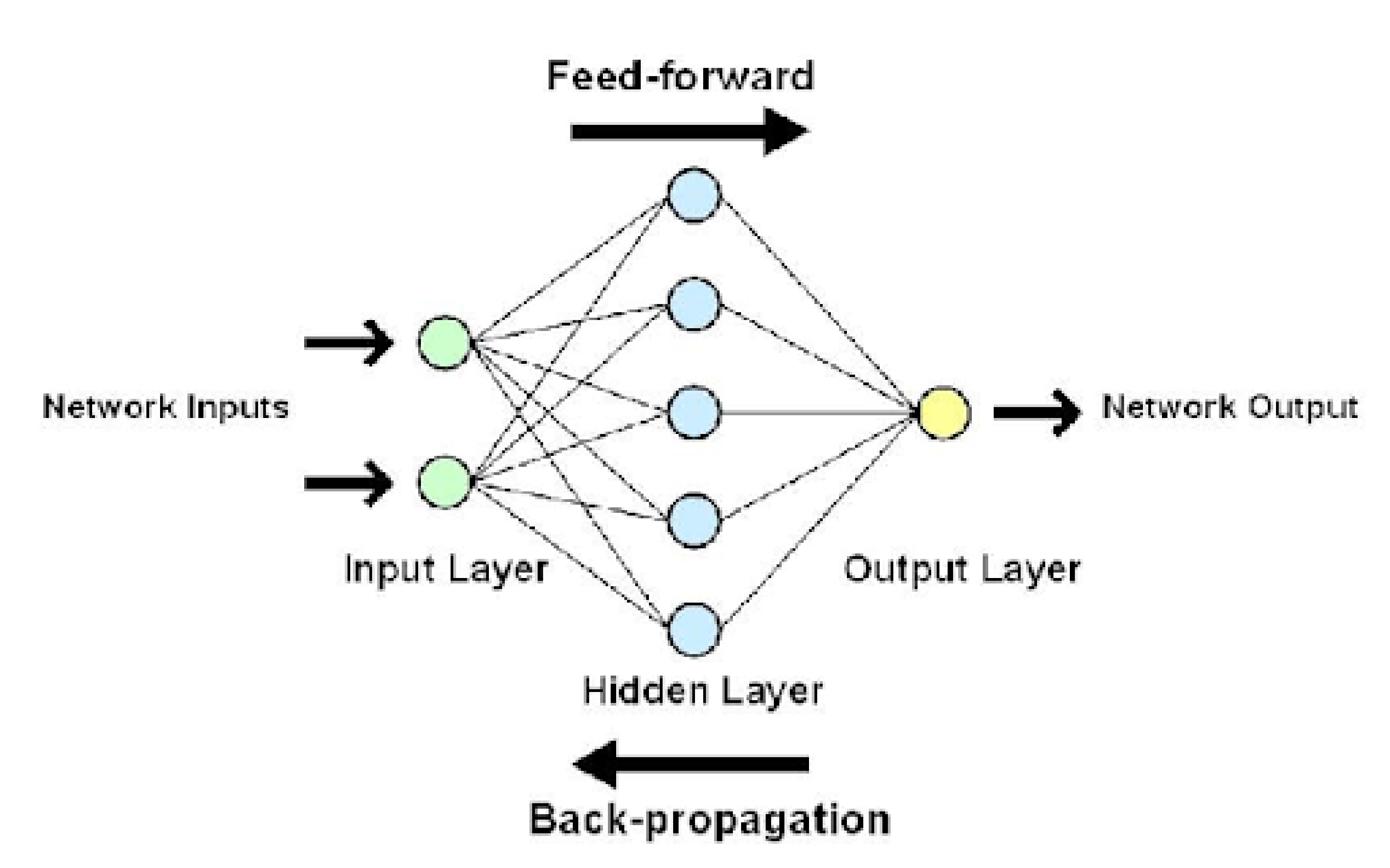
Widely used in practice due to their versatility and ability to deliver high-quality predictions across a range of applications, including but not limited to finance, healthcare, and image classification

ARTIFICIAL NEURAL NETWORK (ANN)

ANN consists of interconnected nodes organized into layers

Information flows through these nodes, and the network adjusts the connection strengths during training to learn from the data

There are three layers in the network architecture: the input layer, the hidden layer (more than one), and the output layer



Application of ANN

Image Processing and Character Recognition

ANNs play a significant part in picture and character recognition because of their capacity to take in many inputs, process them, and infer hidden and complicated, non-linear correlations. Character recognition, such as handwriting recognition, has many applications in fraud detection (for example, bank fraud) and even national security assessments

Forecasting

Forecasting is widely used in everyday company decisions (sales, the financial allocation between goods, and capacity utilization), economic and monetary policy, finance, and the stock market.

DATASET

OVERVIEW

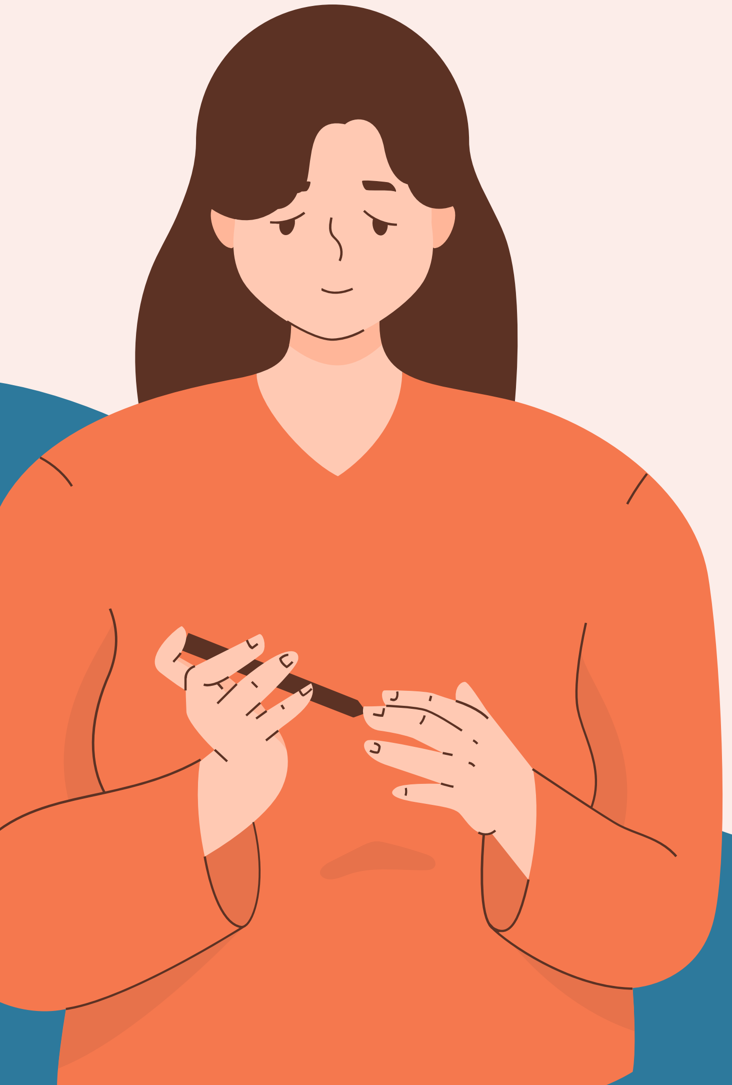
No	Column Name	Meaning
1	Precnancies	Number of pregnancies
2	Glucose	Plasma glucose concentration from oral glucose tolerane test
3	BloofPressure	Diastolic blood pressure (mn Hg)
4	SkinThickness	Triceps skin fold thickness (mn)
5	Insulin	Serum insulin level in blood
6	BMI	Body mass index (weight in kg/ (height in m) ^2
7	DiabetesPredigreeFunstion	Score or the likelihood of diabetes basec on family history
8	Age	Age in years
9	Outcome	Final result (1: Yes, the individual has diabetes 0: No, the individual does not have diabetes)

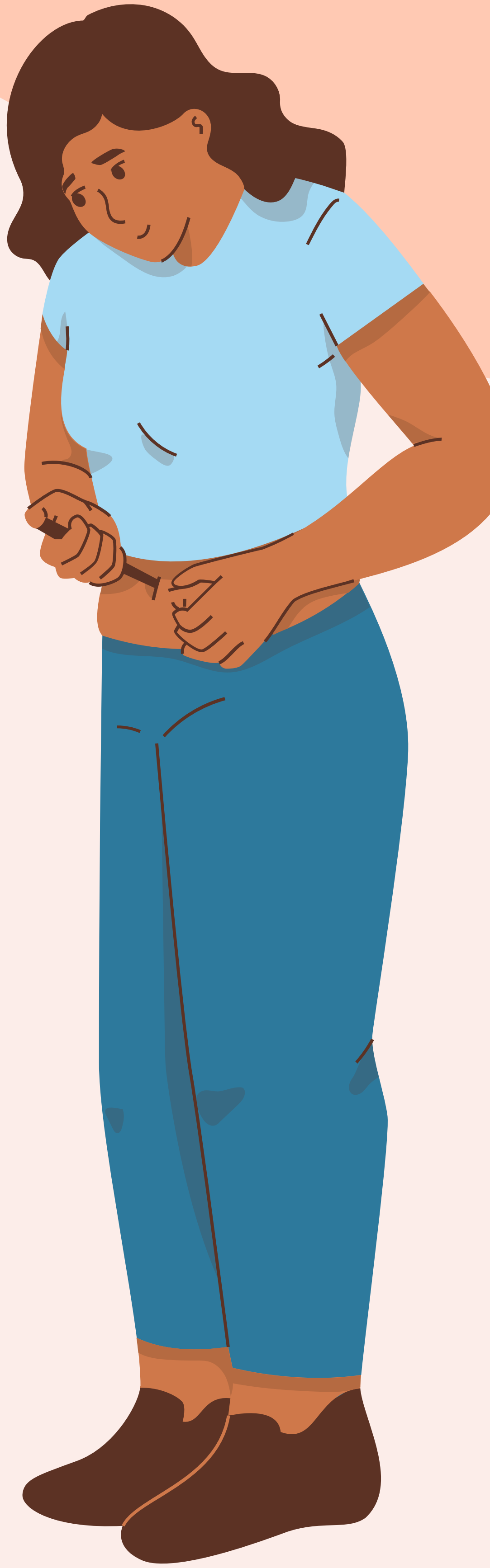
Preprocessing

- Replacing missing values
- Standard Scaler
- Adding new features
- GridSearch

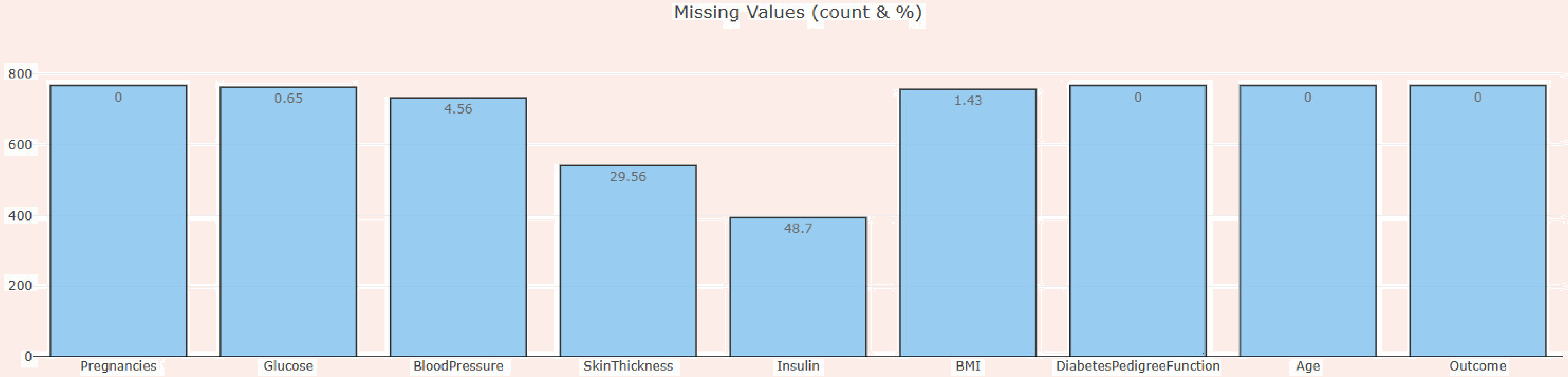
Model

- Logistic Regression
- KNN
- Naives Bayes
- Support Vector Machine
- Decision Tree
- Random Forrest
- ANN





HANDLING MISSING VALUE



ADDING NEW FEATURES

- Age <= 30 & Glucose <= 120
- BMI <= 30
- BloodPressure <= 80
- Glucose/ DiabetesPedigreeFunction
- Age / Insulin ...
- GridSearch

BASIC MODEL EVALUATION

	Accuracy	ROC AUC score
Logistic Regression	77.27	0.7327
KNeighbors	74.67	0.7121
Naives Bayes	74.02	0.7178
Support Vector Machine	83.11	0.7936
Decision Tree	80.51	0.7911
Random Forrest	81.81	0.7327
ANN	75.32	0.7329

RANDOM SEARCH & LIGHTGMB

FOLD	ACCURACY	PRECISION	RECALL	F1 SCORE	ROC AUC
1	0.903	0.915	0.789	0.851	0.945
2	0.864	0.789	0.833	0.811	0.926
3	0.896	0.865	0.833	0.849	0.949
4	0.889	0.846	0.83	0.838	0.944
5	0.928	0.875	0.925	0.899	0.972
mean	0.896	0.858	0.844	0.85	0.947
std	0.021	0.041	0.043	0.029	0.015

Thank you!

FOR YOUR LISTENING