CHANAKYA UNIVERSITY

Heart disease prediction using machine learning algorithms

Exploring the role of ML in heart disease prediction

-Harshit Jindal

Presented by: Dilli Krishna

Sreeya Vydyam

Shaik Mahin Basha

CHANAKYA UNIVERSITY

Contents

Introduction	Brief overview of the presentation objectives and structure.
Background and motivation	Understand the significance and reasons behind utilizing machine learning for heart disease prediction.
Data source and preprocessing	Details on the data origins and the steps involved in preparing the data for analysis.
Machine learning algorithms	Overview of the specific algorithms employed in the prediction model.
Results and analysis	Presentation of the findings and insights derived from the predictive model outcomes.

Introduction to the Study

CHANAKYA UNIVERSITY

Addressing Heart Disease Prediction Challenges

Increase in heart disease cases

Need for accurate prediction methods

Overview of the prediction system developed

Use of machine learning algorithms for prediction

Background and Motivation

Understanding the Global Impact of Cardiovascular Diseases

CARDIOVASCULAR DISEASES (CVDS)

GLOBAL CVD DEATHS

SIGNIFICANCE OF EARLY DIAGNOSIS

PRIOR RESEARCH LIMITATIONS

Objective of the Project

Enhancing Heart Disease Prediction Through Machine Learning

Predictive Modelling

Efficiency Improvement

Accuracy Enhancement

CHANAKYA UNIVERSITY

Data Source

Insights from Dataset on Heart Disease Prediction

Dataset Origin

The dataset was sourced from the UCI repository, containing medical history details of 270 patients.

Attributes Information

Comprises 13 medical attributes including age, gender, chest pain type, fasting blood sugar levels, etc.

Data Split

The data was divided into training and testing sets to develop and evaluate machine learning models effectively.

CHANAKYA UNIVERSITY

Machine Learning Algorithms Used

K-Nearest Neighbours (KNN)
A non-parametric method used for classification and regression that works based on the similarity of data points.

Distance Metrics
to measure the distance or similarity between two points in each space, such as Euclidean, Manhattan Etc.,,.

Normalisation

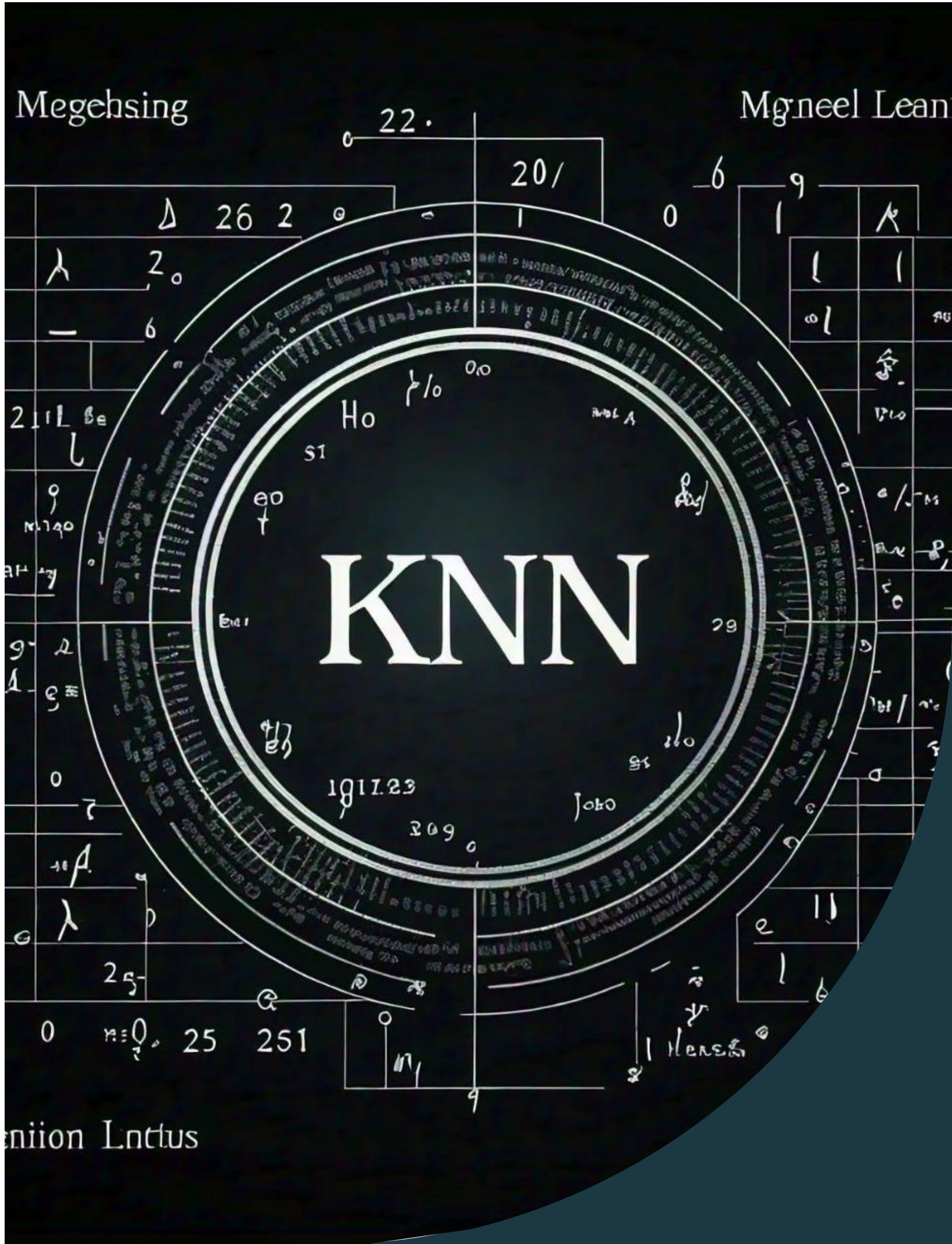
Ensuring all features contribute equally. This improves model performance and stability by accelerating optimization.

Logistic Regression
A statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.

CHANAKYA UNIVERSITY

Understanding K-Nearest Neighbors (KNN) Algorithm

- KNN is an instance-based learning algorithm used for classification and regression.
- It classifies a data point based on the majority class among its k nearest neighbors.
- Select the number of neighbors (k).
- Calculate the distance between the new data point and all existing data points.
- Identify the k nearest neighbors.



Distance Metrics

Distance metrics measure the similarity or dissimilarity between data points.

Crucial for algorithms like KNN, clustering, etc.

- Euclidean Distance:
    - Straight-line distance.
    - Sensitive to the scale of data.
- Manhattan Distance:
    - Sum of absolute differences.
    - Suitable for grid-like data.
- Choice of distance metric impacts algorithm performance.
- Euclidean is best for continuous data; Manhattan for discrete or grid-based data.

CHANAKYA
UNIVERSITY

Normalisation

Normalization is the process of adjusting data values to a common scale or range, making them easier to compare and analyse.

Certainly! Here are the normalization techniques that we have used :

Min-Max Scaler	Robust Scaler	standard scaler

CHANAKYA UNIVERSITY

Min-Max Scaler

Purpose	Purpose: Scales data to a fixed range, typically 0 to 1.
Identify	Identify the Minimum and Maximum: • Find the minimum value (X_{min}) and the maximum value (X_{max}) in the data.
Apply	Apply the Transformation: • For each value xxx in the data, apply the formula: • $x' = (x - X_{min}) / (X_{max} - X_{min})$. • This scales the data such that the smallest value becomes 0 and the largest value becomes 1.

CHANAKYA UNIVERSITY

Robust Scaler

Purpose	Purpose: Scales data using statistics that are robust to outliers, typically the median and the interquartile range (IQR).
Compute	Compute the Median and IQR: •Find the median and the IQR (difference between the 75th and 25th percentiles).
Apply	Apply the Transformation: •For each value xxx in the data, apply the formula: •x′=(x−median)/IQR •This centers the data around the median and scales it according to the IQR.

CHANAKYA UNIVERSITY

Standard Scaler

Purpose: Standardizes features by removing the mean and scaling to unit variance.

Steps:

Compute the Mean and Standard Deviation:

Find the mean (μ) and the standard deviation (σ) of the data.

Apply the Transformation:

For each value x in the data, apply the formula:

$x' = (x - \mu)/\sigma$

This centers the data around 0 with a standard deviation of 1.

CHANAKYA UNIVERSITY

Logistic Regression

Logistic regression is a method used in machine learning to predict whether something belongs to one of two categories. For example, it can help determine if an email is spam or not spam, or if a customer will buy a product or not.

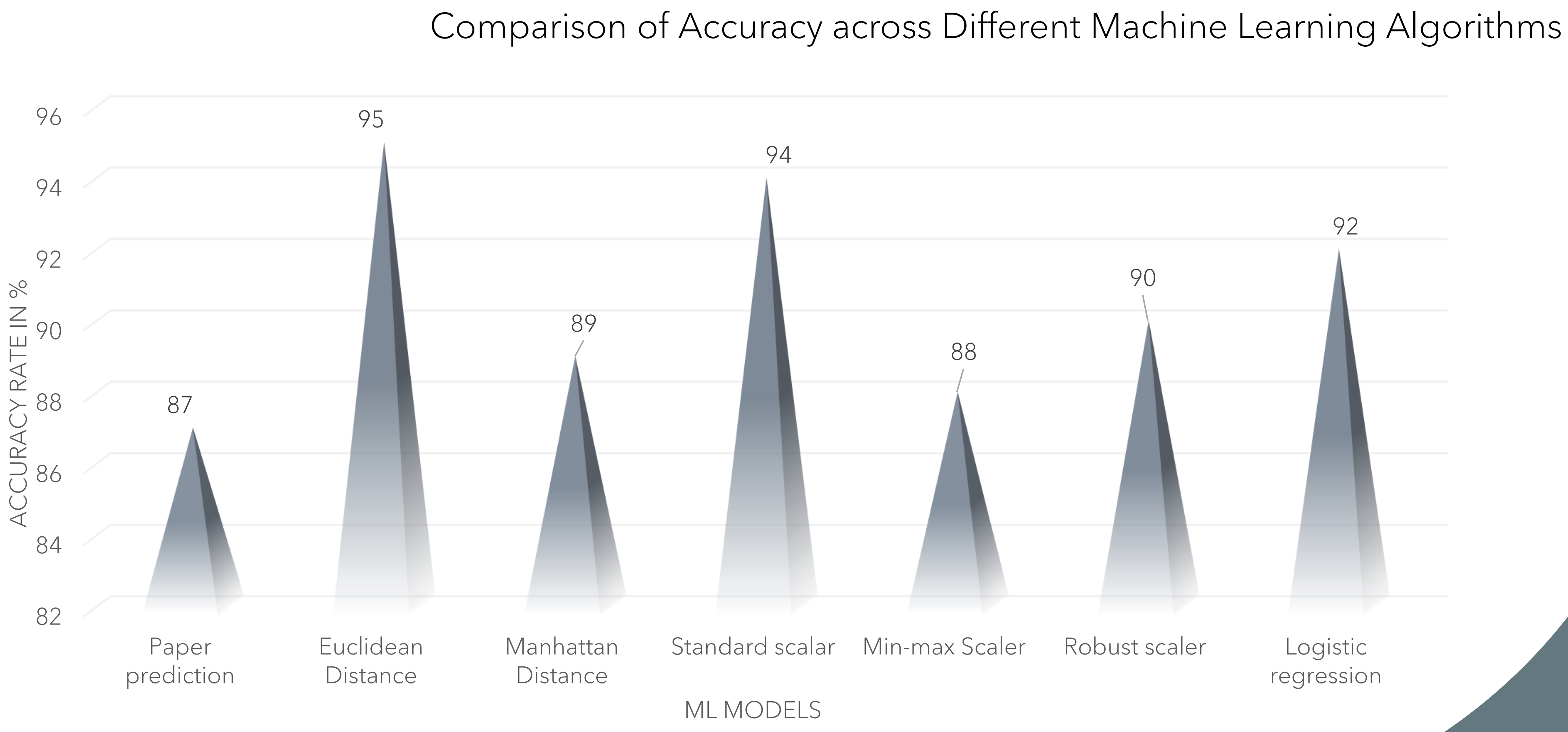
Input Data: You start with some data that includes various features (characteristics) and a binary outcome (yes/no, true/false, 0/1).

Model Training: You use this data to train the logistic regression model. The model learns the relationship between the features and the binary outcome.

Probability Prediction: The model outputs a probability score between 0 and 1. This score indicates the likelihood that a particular instance belongs to the positive class (e.g., "yes" or "1").

Threshold Decision: You set a threshold (usually 0.5). If the probability score is above the threshold, the model predicts the positive class. If it's below, it predicts the negative class.

Accuracy Comparison on test data



Findings from our models

Development of Cardiovascular Disease Detection Model

Utilized Logistic Regression, KNN with distance metric and Normalization Techniques

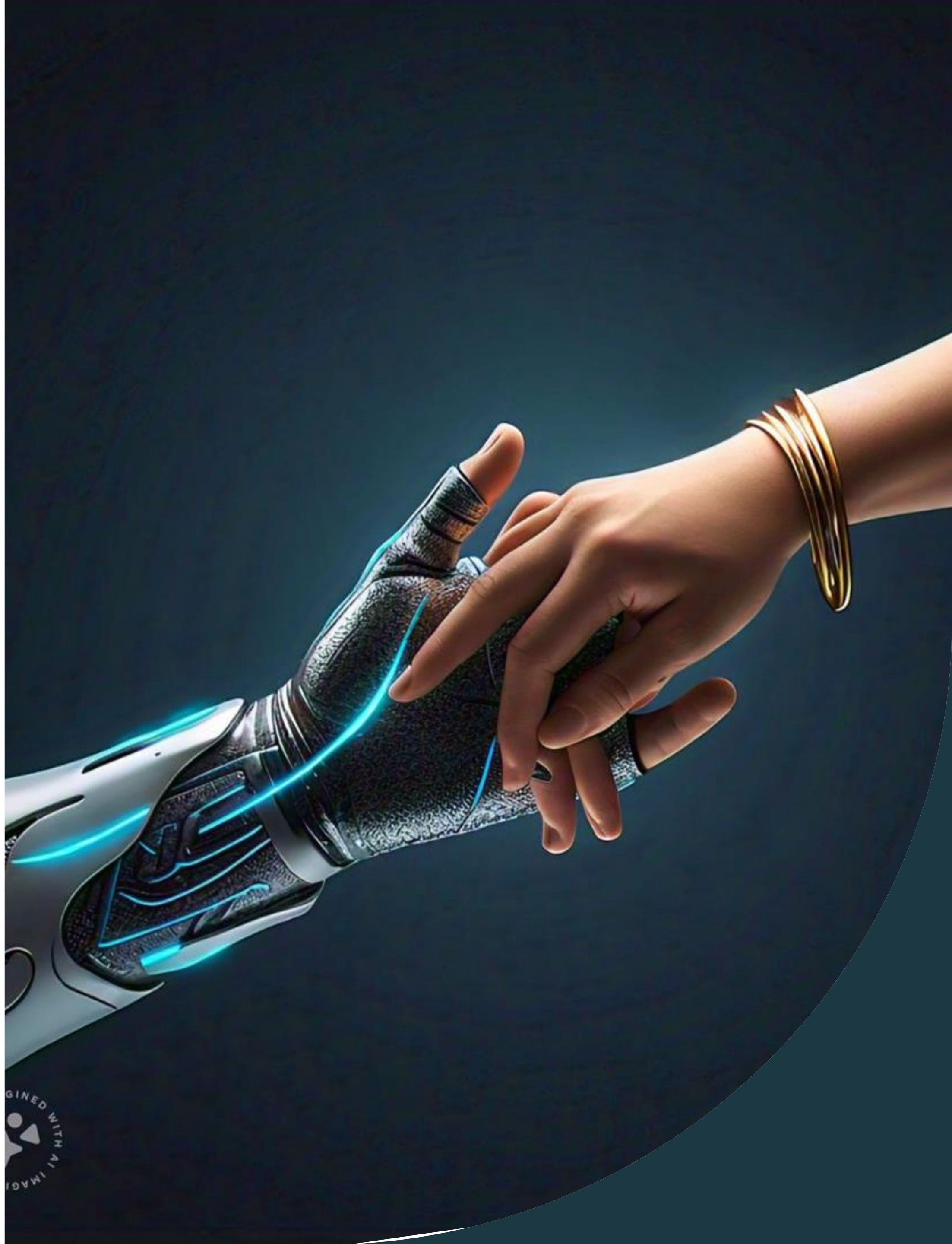
Achieved Accuracy of 92.75% in average among all the models

KNN with Euclidean metrics as Most Efficient Algorithm

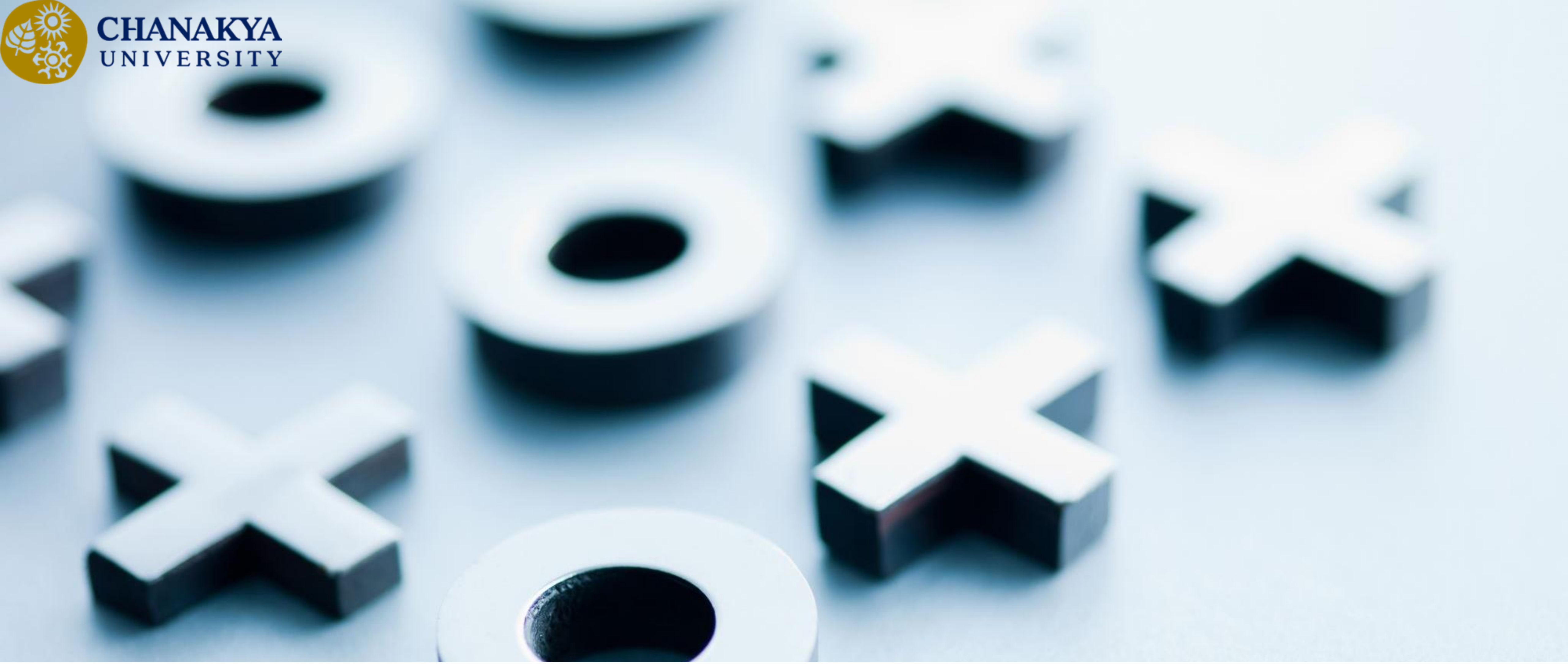
Outperformed other algorithms in terms of performance and predictive capabilities.

Potential for Improved Medical Care and Cost Reduction

Implementation of these models can lead to better healthcare outcomes and reduced expenses.



CHANAKYA UNIVERSITY



Thank you for your patience