DATA SCIENCE AND ADVANCED PYTHON CONCEPTS WORKSHOP FOR NEUROSCIENCE

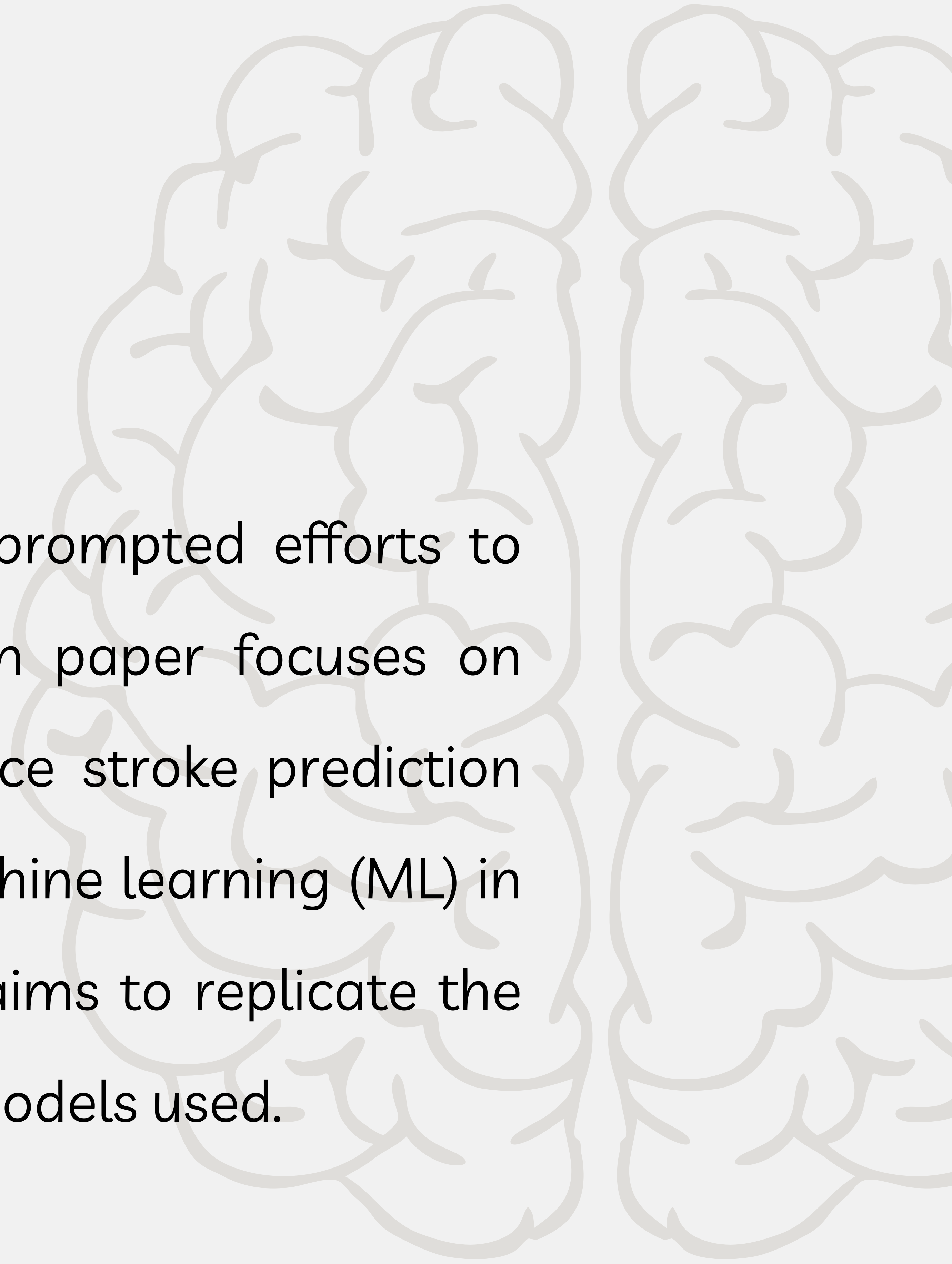
Stroke Prediction Using Machine Learning Models

AMIT DAVIDPUR & MESHI BEN OZ

30 JANUARY, 2025

Introduction

Stroke's significant negative impact on society has prompted efforts to improve its management and diagnosis. The chosen paper focuses on analyzing electronic health records (EHRs) to enhance stroke prediction and was selected for its innovative application of machine learning (ML) in addressing this critical healthcare issue. Our project aims to replicate the study's findings by implementing and testing the ML models used.



Methods

Preprocessing

- Checked for duplicates, missing values, and performed descriptive analysis.
- Visualized data with bar plots and histograms.
- Imputed missing BMI using group medians by gender and glucose level.

Feature Analysis

- Analyzed correlations (Pearson) and visualized with a heatmap.
- Used PCA to identify key components explaining variance.
- Balance the data.

Models

- Decision Tree, Random Forest, Neural Network, SVM, Gradient Boosting.
- Performed hyperparameter tuning with grid search.
- Evaluated on the balanced dataset with all features and PCA-reduced datasets (2 & 8 components), comparing performance metrics across different feature sets.

Results

Model	All features (Original Paper)	All features (Our Findings)	2 PCA (Original Paper)	2 PCA (Our Findings)	8 PCA (Original Paper)	8 PCA (Our Findings)
Decision Tree	0.74	0.66	0.73	0.57	0.73	0.55
Random Forest	0.74	0.73	0.69	0.66	0.72	0.73
Neural Network	0.74	0.73	0.74	0.69	0.75	0.75
SVM	0.68	0.73	-	0.62	-	0.73
Gradient Boosting	-	0.71	-	0.65	-	0.70