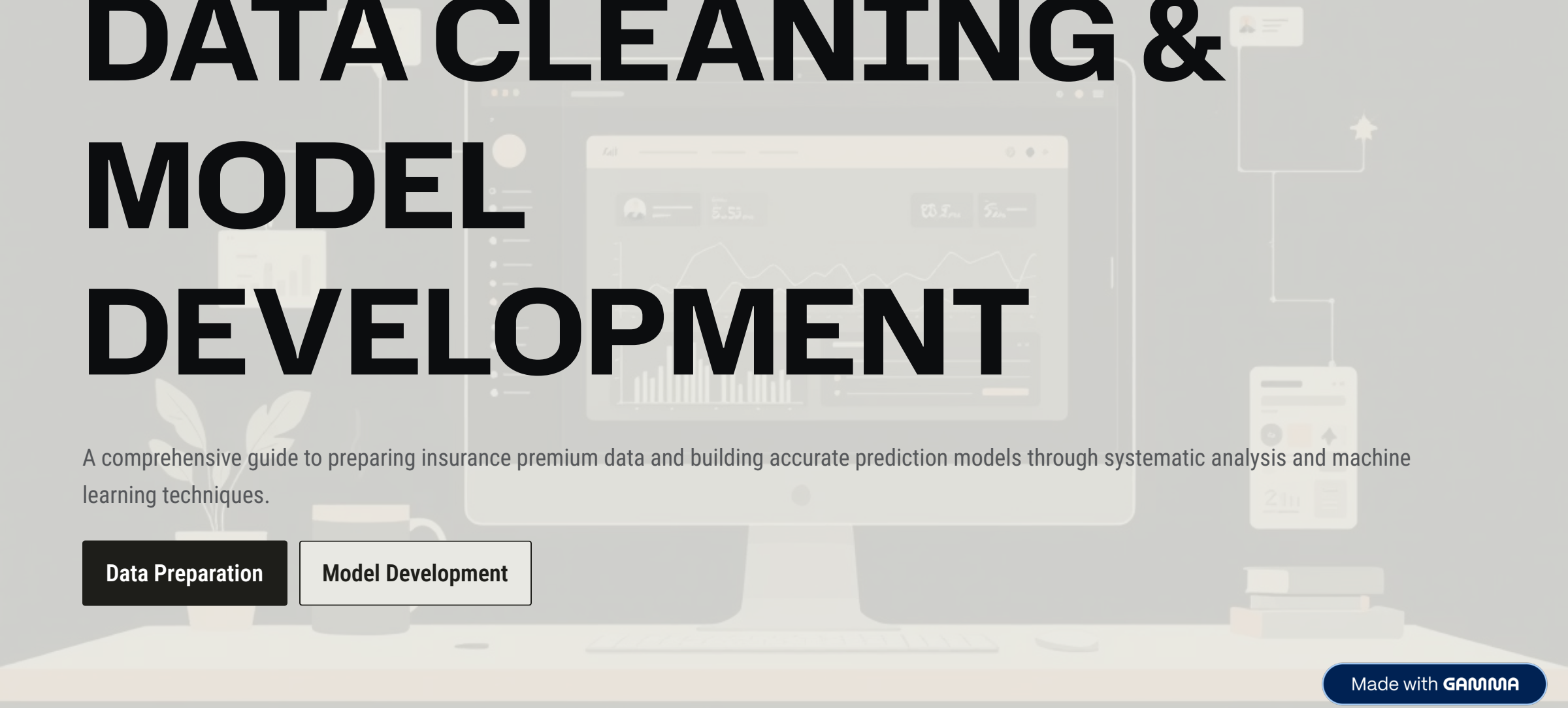
DATA CLEANING & MODEL DEVELOPMENT



A comprehensive guide to preparing insurance premium data and building accurate prediction models through systematic analysis and machine learning techniques.

Data Preparation | **Model Development**

DATA IMPORT AND INITIAL ANALYSIS

FOUNDATION STEPS

Dataset loaded using Pandas library. Initial exploration revealed dataset dimensions through .shape method. Data types verified using .info() function. First five rows examined with .head() to understand structure and content patterns.

01

LOAD DATASET

Import data using Pandas

02

CHECK DIMENSIONS

Verify rows and columns count

03

INSPECT TYPES

Review feature data types

04

PREVIEW DATA

Examine initial rows

MISSING VALUES AND OUTLIERS HANDLING

IDENTIFY MISSING VALUES

Used .isnull().sum() to detect missing data across all columns and quantify gaps in dataset.

FILL OR REMOVE

Applied mean or mode imputation strategies to handle missing values appropriately based on feature type.

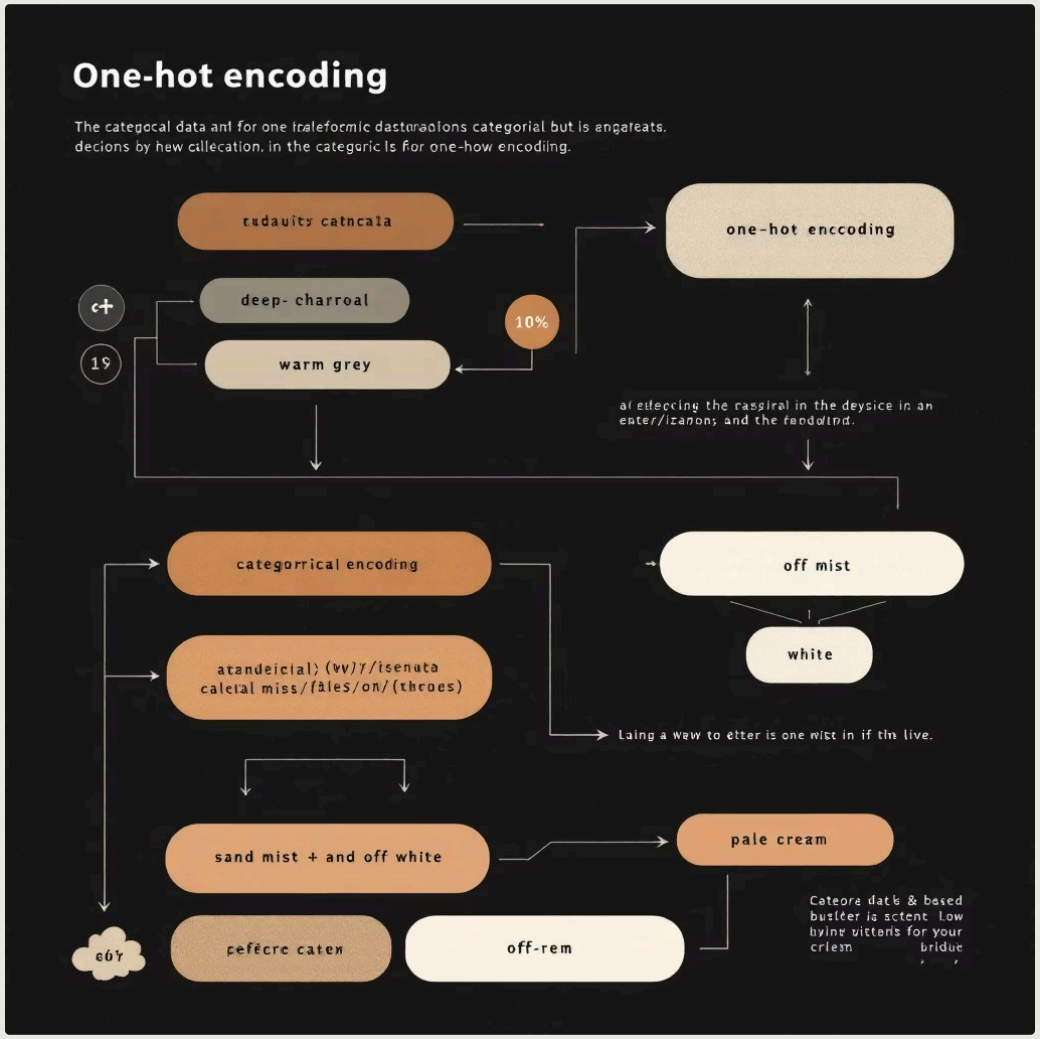
DETECT OUTLIERS

Identified extreme values that could skew model performance using statistical methods and visualization.

CORRECT OUTLIERS

Applied capping or removal techniques to handle outliers and ensure data quality for modeling.

CATEGORICAL DATA ENCODING



TRANSFORMING TEXT TO NUMBERS

Identified categorical features like gender and smoker status. Converted these text-based features into numerical values using encoding techniques such as One-Hot Encoding, making them compatible with machine learning algorithms.

1

TEXT DATA

Gender, Smoker

2

ENCODED VALUES

0s and 1s

DATA PREPARATION COMPLETE

Final cleaned dataset fully prepared for modeling phase with all missing values handled, outliers corrected, and categorical features properly encoded.

QUALITY ASSURED

No missing values, outliers managed, all features numerical

READY FOR ML

Dataset structured optimally for machine learning algorithms

EXPLORATORY DATA ANALYSIS

UNDERSTANDING RELATIONSHIPS

Analyzed relationships between features and premium amounts using Matplotlib and Seaborn visualizations. Examined correlation patterns to identify which features most strongly influence insurance premiums and how variables interact with each other.

VISUALIZATIONS

Created graphs showing feature relationships

CORRELATIONS

Checked feature interdependencies

FEATURE ENGINEERING AND DATA SPLITTING

FEATURE CREATION

Engineered new features to enhance model performance based on domain knowledge

1

DATA SPLITTING

Divided dataset into Training Set (for learning) and Testing Set (for validation)

3

2

FEATURE SELECTION

Removed irrelevant or redundant features that didn't contribute to predictions

MODEL SELECTION AND TRAINING

LINEAR REGRESSION

Simple baseline model for establishing performance benchmarks

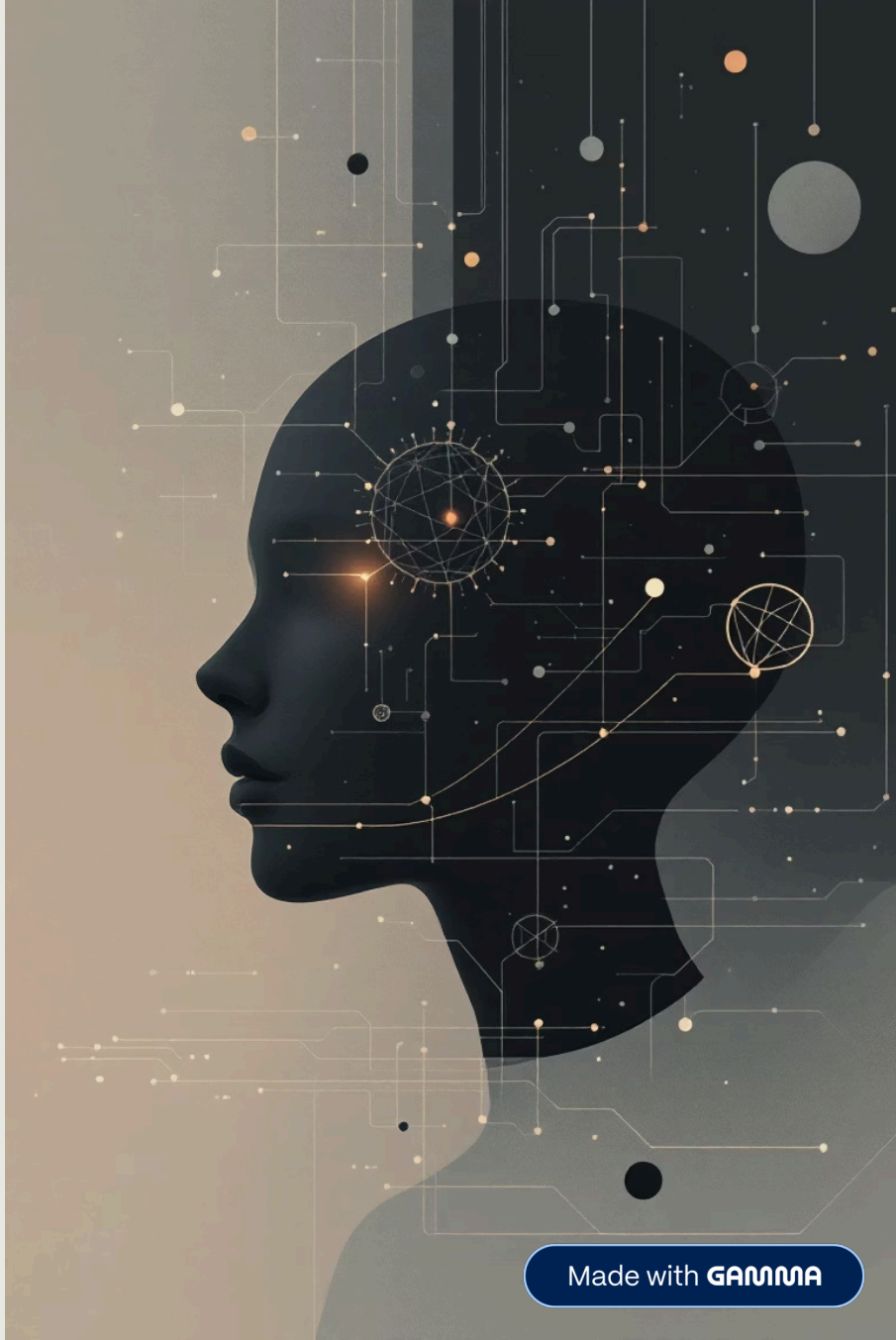
RANDOM FOREST

Ensemble method using multiple decision trees for robust predictions

GRADIENT BOOSTING

Advanced technique building models sequentially to minimize errors

Each model trained on Training Set using fit method. Multiple algorithms tested to identify best performer for insurance premium prediction task.



MODEL EVALUATION AND OPTIMIZATION

PERFORMANCE METRICS

Evaluated each model on Testing Set using RMSE, MAE, and R-squared metrics. Selected best-performing model and applied Hyperparameter Tuning through Grid Search and Random Search to optimize performance further.

87%

TRAINING ACCURACY

Model performance on training data

86%

TESTING ACCURACY

Validated performance on unseen data

FINAL MODEL DEPLOYMENT

BEST MODEL SELECTED

Identified top performer after comprehensive evaluation

HYPERPARAMETERS TUNED

Optimized model configuration for maximum accuracy

MODEL SAVED

Preserved final model as pickle file for future deployment

> The optimized model achieved 87% training accuracy and 86% testing accuracy, demonstrating strong predictive capability with minimal overfitting. Ready for production deployment.