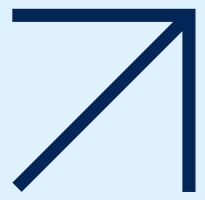


# DEFINING BUSINESS PROBLEM



## SITUATION

**NHS healthcare costs** are rising by **3.3%** per year, expected to reach **8.9%** of national income by 2033/34, This growth is **outpacing economic expansion**.

### Key drivers:

- **Growing & aging population** → More patients, longer treatments.
- **Rising chronic diseases** → Complex, long-term care needs.

## COMPLICATION

If costs remain **uncontrolled**, the NHS will face:

**1.** **Financial strain:**  
→ **Higher taxes** or national insurance contributions.

**2.** **Resource inefficiencies:**

- **Low-risk patients** may receive unnecessary care.
- **High-risk patients** may not receive timely, adequate treatment.

## QUESTION

**HOW CAN WE MANAGE THESE UNPREDICTABLE COSTS AND ENSURE EFFICIENT RESOURCE DISTRIBUTION**

**Healthcare costs are unpredictable**, making budget planning difficult.

**Inefficient resource allocation** leads to wasted funds & poor patient care.

**Data-driven decisions are key** to sustainability and efficiency.

## ANSWER

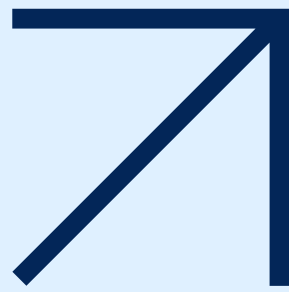
**By predicting billing amounts and leveraging data-driven insights, we can:**

**Financial Benefits**  
→ Optimize NHS budgets & prevent overspending.

**Operational Improvements**  
→ Allocate resources efficiently to high-risk patients.

**Patient Care Enhancements**  
→ improve patient care through targeted preventive measures.

# ABOUT THE HEALTHCARE DATASET



This dataset contains **55,500 rows and 15 columns**. It includes a mix of **categorical (object)**, **numerical (integer and float)**.

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number
55495	eLIZABeTH jaCkSOn	42	Female	O+	Asthma	2020-08-16	Joshua Jarvis	Jones-Thompson	Blue Cross	2650.714952	417
55496	KYle pEReZ	61	Female	AB-	Obesity	2020-01-23	Taylor Sullivan	Tucker-Moyer	Cigna	31457.797307	316
55497	HEATher WaNG	38	Female	B+	Hypertension	2020-07-13	Joe Jacobs DVM	and Mahoney Johnson Vasquez,	UnitedHealthcare	27620.764717	347
55498	JENniFER JOneS	43	Male	O-	Arthritis	2019-05-25	Kimberly Curry	Jackson Todd and Castro,	Medicare	32451.092358	321
55499	jAMES GARCIA	53	Female	O+	Arthritis	2024-04-02	Dennis Warren	Henry Sons and	Aetna	4010.134172	448

## 15 Features

*Name, Age, Gender, Blood Type, Medical Condition, Date of Admission, Doctor, Hospital, Insurance Provider, Billing Amount, Room Number, Admission Type, Discharge Date, Medication, Test Results*

## 55,500 Features

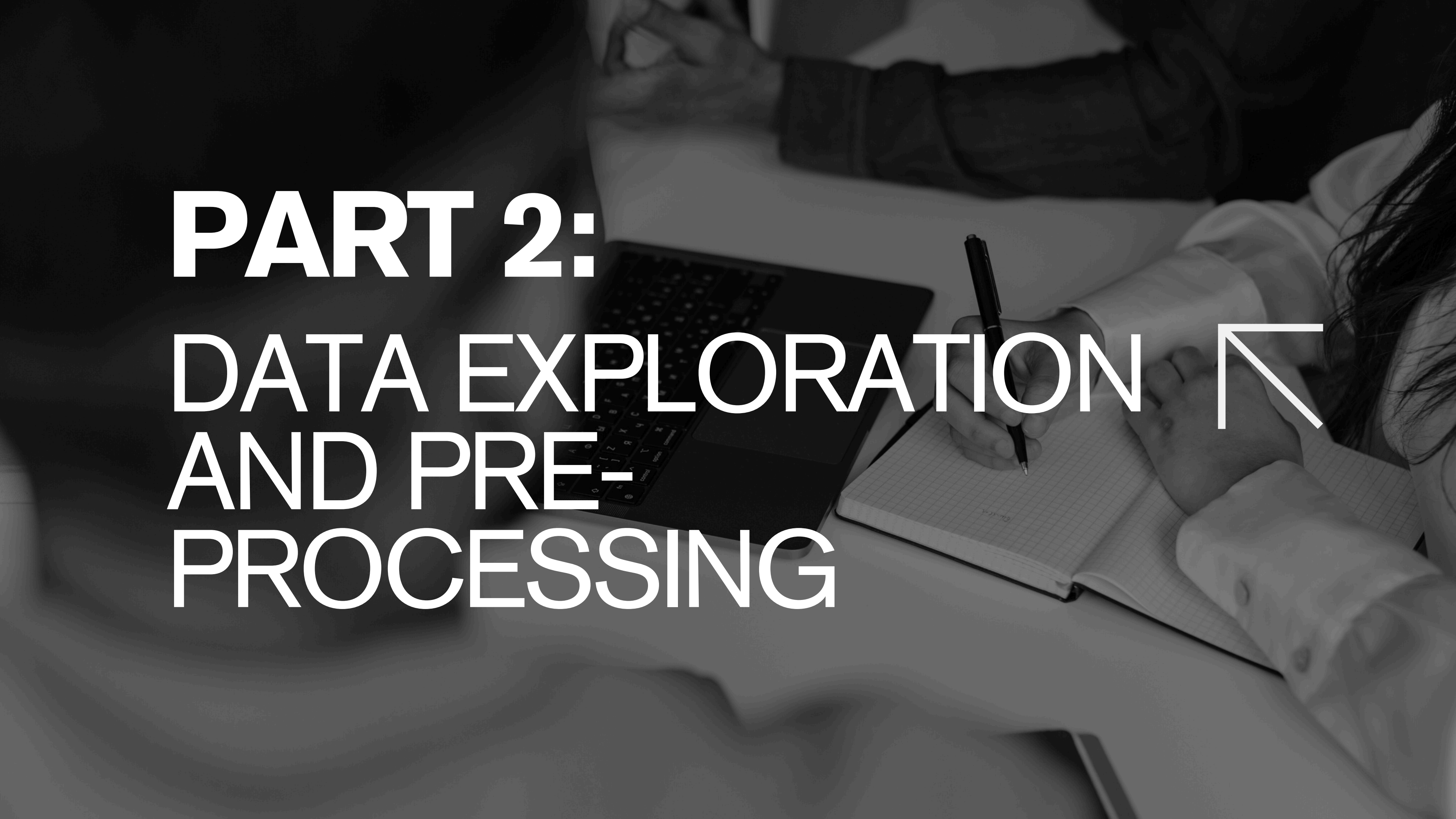
*Which provides specific information about the patient, their admission, and the healthcare services provided*

*Note: The Billing Amount is a continuous numerical variable, so predicting it falls under Regression, not Classification.*

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55500 entries, 0 to 55499
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype  
---  -
 0   Name                55500 non-null object  
 1   Age                 55500 non-null int64  
 2   Gender              55500 non-null object  
 3   Blood Type          55500 non-null object  
 4   Medical Condition    55500 non-null object  
 5   Date of Admission    55500 non-null object  
 6   Doctor              55500 non-null object  
 7   Hospital            55500 non-null object  
 8   Insurance Provider   55500 non-null object  
 9   Billing Amount        55500 non-null float64
10   Room Number         55500 non-null int64  
11   Admission Type       55500 non-null object  
12   Discharge Date       55500 non-null object  
13   Medication           55500 non-null object  
14   Test Results         55500 non-null object  
dtypes: float64(1), int64(2), object(12)
memory usage: 6.4+ MB
```

This dataset **already has no missing values** across all variables. Most categorical features are currently stored as text labels and **will require encoding** for machine learning.



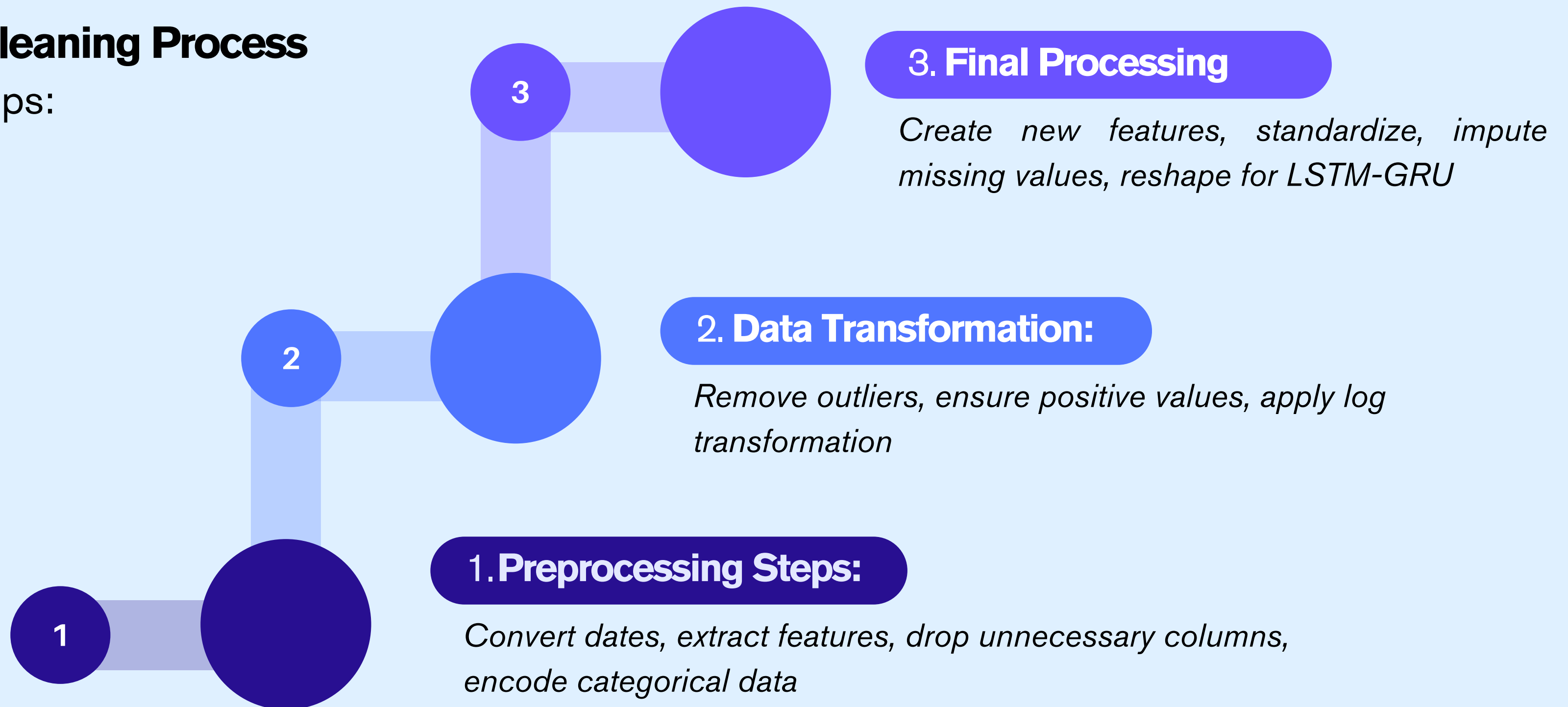
# **PART 2:** **DATA EXPLORATION** ↖ **AND PRE-** **PROCESSING**

# EXPLORATORY DATA ANALYSIS (EDA) & DATA PRE-PROCESSING



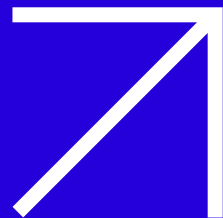
## Data Cleaning Process

Key Steps:





# BEFORE & AFTER DATA PREPROCESSING



Before:

Before:

Healthcare Dataset:															
	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
0	Bobby JacksOn	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol	Normal
1	LesLie TErRy	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen	Inconclusive
2	DaNnY sMitH	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin	Normal
3	andrEw waTtS	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen	Abnormal



**Feature Engineering**  
→ Created **Age\_Stay**, **Month\_Day** for improved modeling



**Outlier Removal**  
→ Applied **Z-score filtering** to remove extreme values



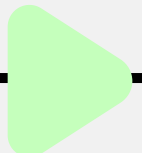
**Data Normalization**  
→ Performed **log transformation** on Billing Amount to reduce skewness



**Standardization**  
→ Applied **StandardScaler** for uniform feature scaling



**Missing Data Handling**  
→ Used **IterativeImputer** to reconstruct missing values



**Final Reshaping**  
→ Converted data to **3D format for LSTM-GRU**

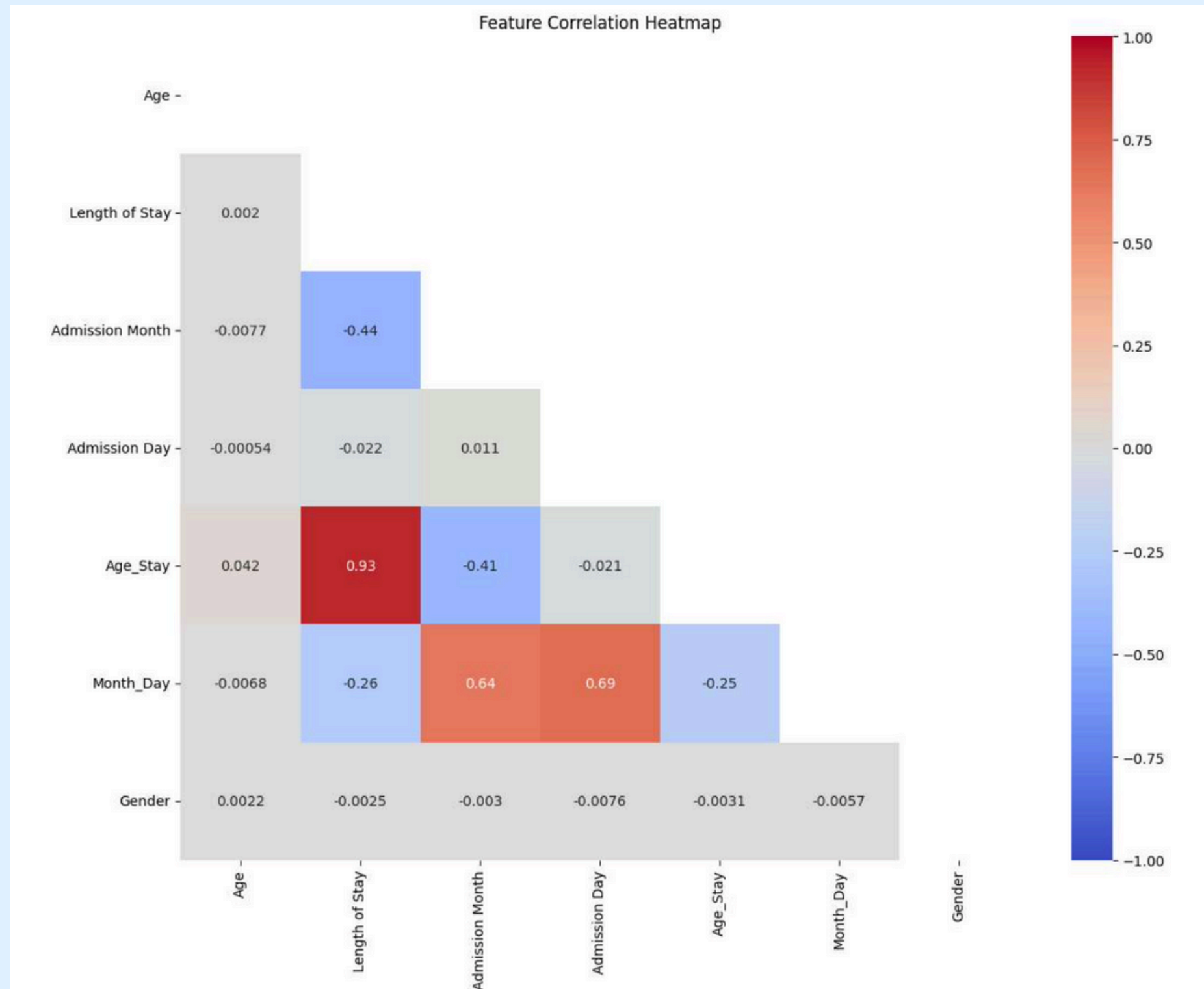
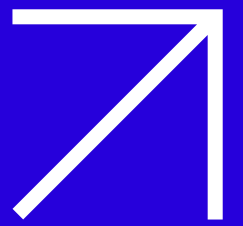
After:

Final Processed Dataset with Feature Engineering:

	Age	Gender	Medical Condition	Length of Stay	Admission Month	Admission Day	Age_Stay	Month_Day
0	30	0	Cancer	2.0	1	31	60.0	31
1	62	0	Obesity	NaN	8	20	NaN	160
2	76	1	Obesity	-74.0	9	22	-5624.0	198
3	28	1	Diabetes	NaN	11	18	NaN	198



# VISUALIZATION: CORRELATION MATRIX



## Key Insights:

The **Correlation Matrix** helps us identify **key relationships** between features,

guiding **feature selection**

(e.g., retaining *Age\_Stay* and testing *Month\_Day* for seasonality)

to ensure that our models learn the most relevant patterns for accurate **Billing Amount predictions**.

## Strong Positive Correlations:

Length of Stay & Age\_Stay (**0.93**), Month\_Day with Admission Day (**0.69**) and Admission Month (**0.64**)



# **PART 3:** **MODEL TRAINING** ↖ **& HYPERPARAMETER** **TUNING**

# MAPE

MAPE stands for Mean Absolute Percentage Error, which is a key evaluation metric used to assess the accuracy of the healthcare billing prediction models.

MAPE measures how far the model's predictions are from the actual billing amounts, expressed as a percentage. It's calculated by taking the average of the absolute percentage differences between the predicted values and actual values.

Why MAPE was chosen for this specific application:

Interpretability: MAPE provides an easy-to-understand percentage error that business stakeholders can quickly grasp. For example, a MAPE of 10% means predictions are off by 10% on average.

Business relevance: In healthcare finance, percentage errors are often more meaningful than absolute dollar amounts. A \$100 error on a \$1,000 bill (10%) is more significant than a \$100 error on a \$10,000 bill (1%).

Scale-independence: Since healthcare billing amounts can vary widely, MAPE normalizes the errors relative to the actual values, making it appropriate for evaluating predictions across different billing magnitudes.

In the code, MAPE was calculated for each model using scikit-learn's `mean_absolute_percentage_error` function and then multiplied by 100 to convert to a percentage:

```
results[name] = mean_absolute_percentage_error(y_test, preds) * 100
```

Lower MAPE values indicate better model performance, as they represent smaller percentage deviations from the actual billing amounts. The comparative MAPE scores shown in the bar plot visualization help determine which model provides the most accurate billing predictions for healthcare financial planning.

Comparison across models: MAPE allows for fair comparison between different modeling approaches (Random Forest, MLP, and LSTM-GRU) on the same scale.

THE FORMULA FOR MAPE IS:

$$\text{MAPE} = (1/N) * \sum |(\text{ACTUAL} - \text{PREDICTED})/\text{ACTUAL}| * 100\%$$



# OPTUNA | HYPERPARAMETER TUNNING

## Random Forest (RF)

Key Hyperparameters: Parameters like ``max_depth``, ``n_estimators``, ``max_features``, and ``min_samples_split`` significantly affect RF's predictive performance and computational efficiency.

Role of Optuna: Optuna efficiently explores the parameter space using techniques like Tree-structured Parzen Estimators (TPE), outperforming traditional grid or random search methods. This ensures better generalization and reduced overfitting.

## 2. Multi-Layer Neural Networks (MLNN)

Key Hyperparameters: Learning rate, number of layers, neurons per layer, activation functions, and dropout rates are critical for MLNN performance.

Role of Optuna: MLNNs require careful balancing to avoid underfitting or overfitting. Optuna automates this process by testing combinations of hyperparameters and identifying optimal configurations efficiently.

## 3. LSTM-GRU Architectures

- Key Hyperparameters: Sequence length, hidden units, learning rate, optimizer type, and dropout are crucial for time-series regression tasks.
- Role of Optuna: These architectures involve complex interactions between parameters. Optuna's Bayesian optimization effectively handles this complexity, leading to improved model accuracy and training stability.

## Advantages of Using Optuna

- Efficiency: Reduces the number of trials needed compared to exhaustive search methods.
- Dynamic Search: Adapts based on prior trials to focus on promising regions of the parameter space.
- \*\*Integration: Works seamlessly with frameworks like Scikit-learn, TensorFlow, and PyTorch.

In summary, Optuna's advanced optimization capabilities make it an invaluable tool for fine-tuning RF, MLNN, and LSTM-GRU models to achieve superior regression performance.

# MODEL IMPLEMENTATION

## RANDOM FOREST

Random Forest is an ensemble method constructs multiple decision trees during training and aggregates predictions.

### Why?

- **Aggregation for Comprehensive Billing Insights:**

Multiple decision trees combine outputs to **capture diverse variables**, as collecting opinions to form **robust billing forecasts**.

- **Voting Mechanism to Mitigate Noise:**

Reduces the influence of noisy data, ensure predictions **remain reliable** with the **variable** healthcare billing records.

- **Diverse splits on high-dimensional billing features:**

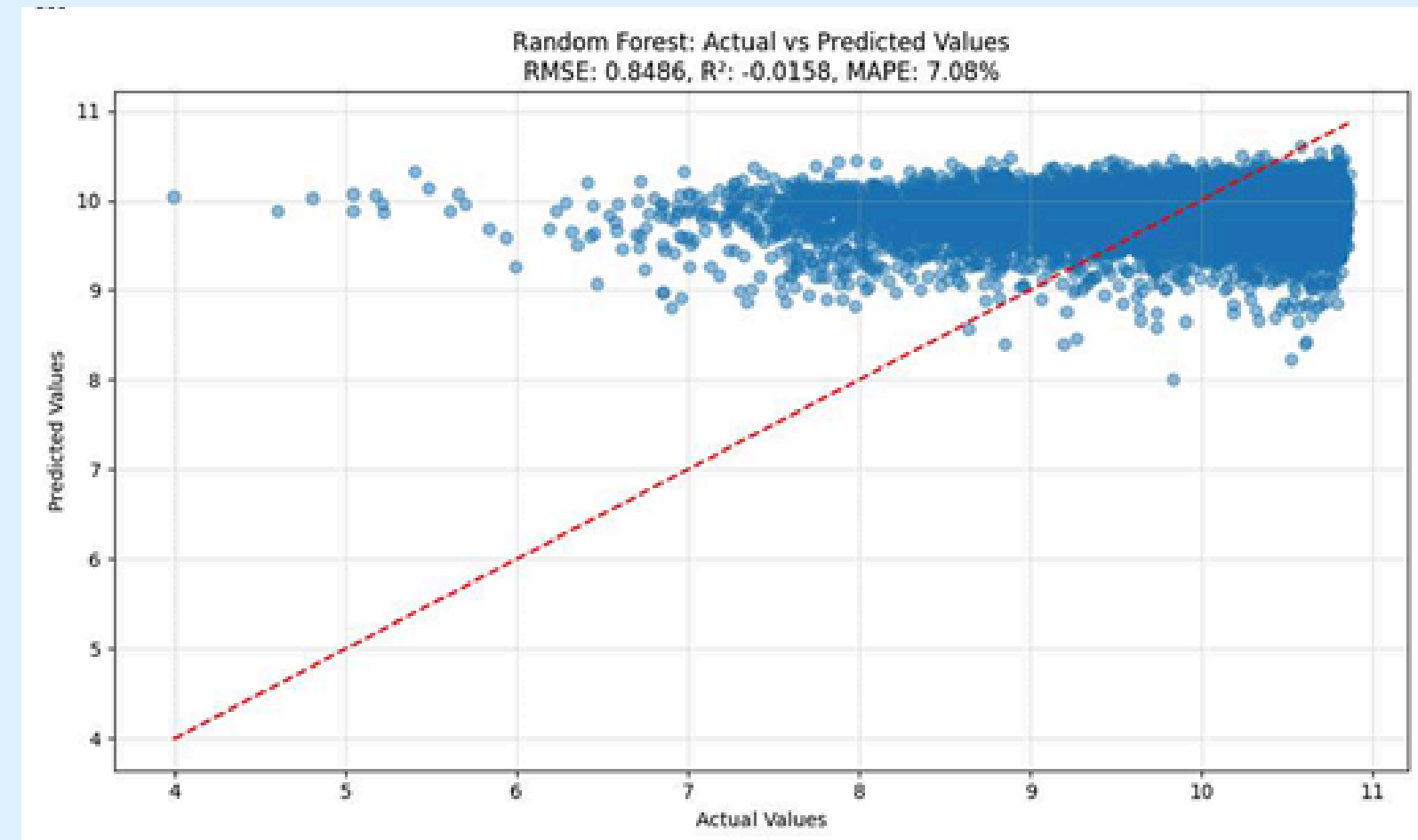
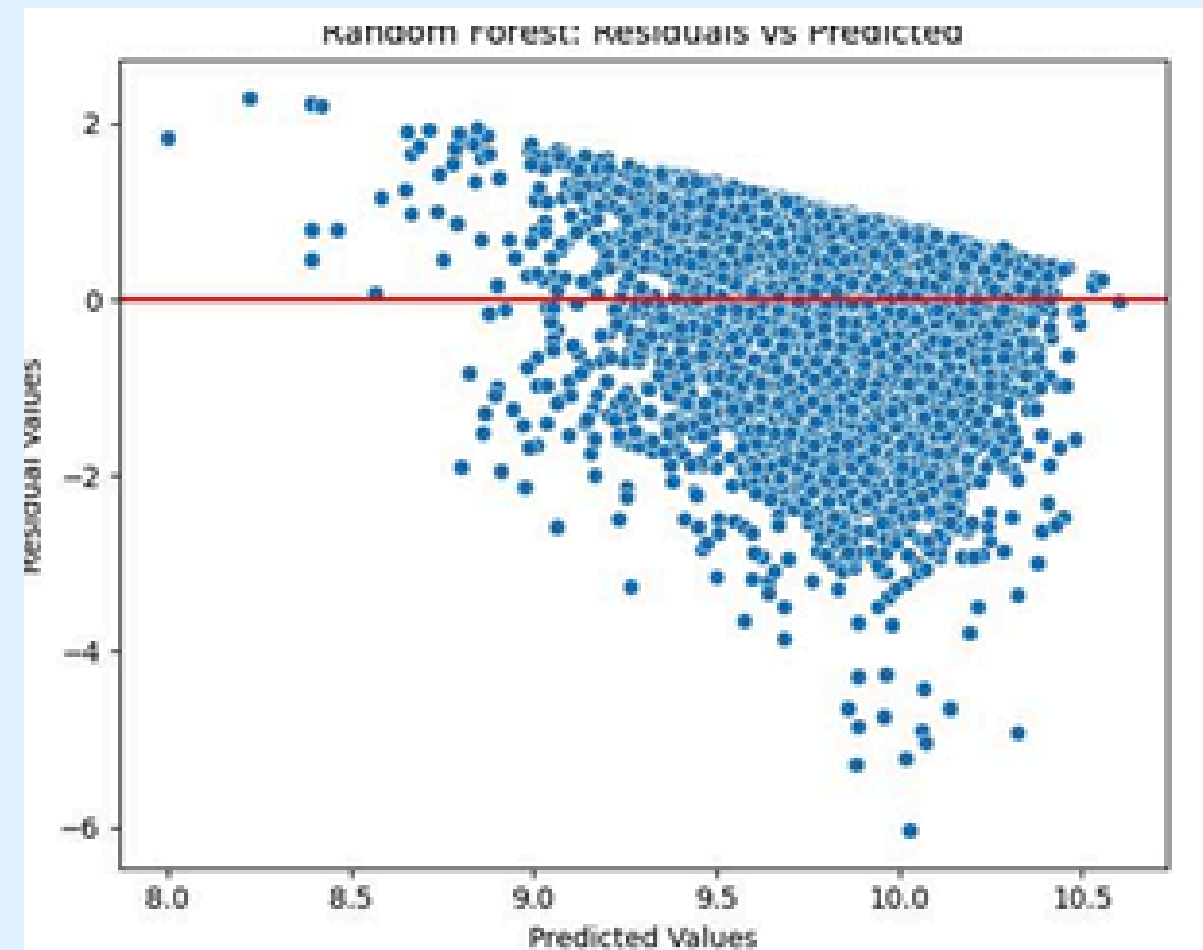
Uncovers critical cost factors (**overlapping diagnoses and procedural variations**)

- **Interpretability for Stakeholder Confidence:**

Provides **clear** feature importance metrics, enabling finance and management teams to understand key cost drivers.

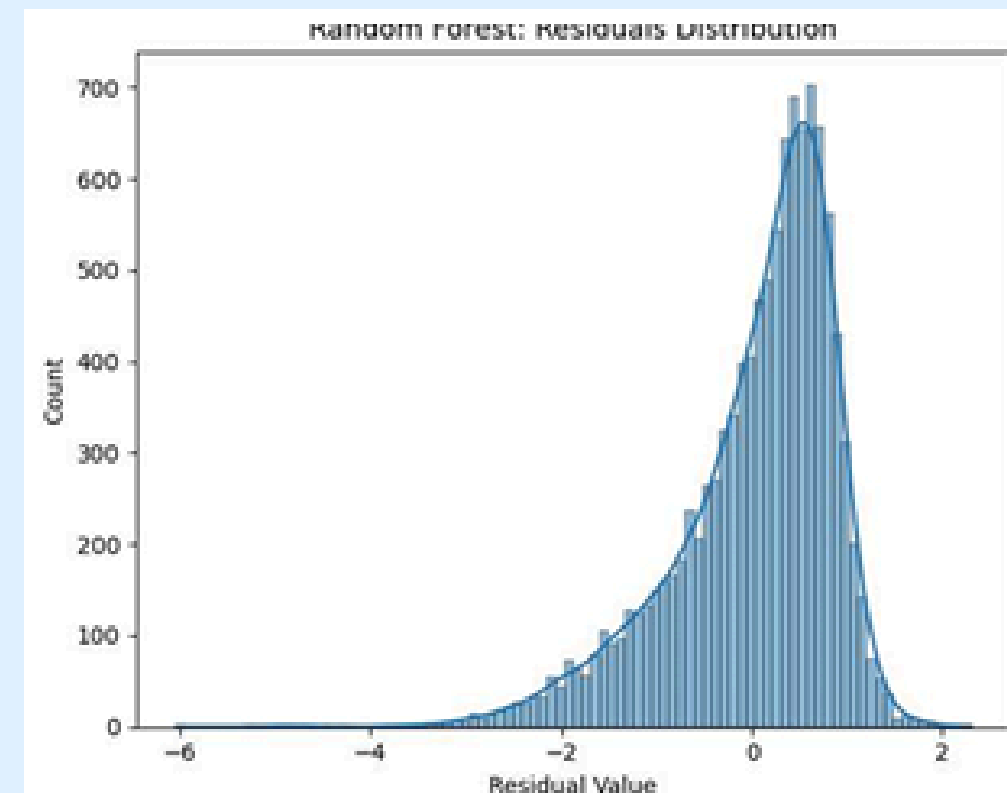
### Model Training

- Objective function: **Minimize MAPE on 5-fold cross-validation**
- Conducted **Bayesian optimization** using Optuna with 100 trials
- n\_estimators: **150**
- max\_depth: **20**
- min\_samples\_split: 5
- min\_samples\_leaf: 2
- criterion: 'mse'



### Performance

- **MAPE: 7.08%**
- **RMSE: 0.8486**
- **$R^2$ : 0.0158**



### Positives

- Provides **clear feature importance metrics**, empowering finance and operations teams to **pinpoint key cost drivers**.
- **Fast training** allows for quick model updates as new billing data arrives.
- Enhanced **robustness** ensuring dependable predictions across **varying patient profiles and treatment scenarios**.

# MODEL IMPLEMENTATION

# MULTI-LAYER PERCEPTRON (MLP) NEURAL NETWORK

A multi-layer perceptron (MLP) is a type of artificial neural network consisting of multiple middle layers of neurons.

## Why?

- **Capturing Non-Linear Cost Patterns:**

The **multi-layer architecture** captures **intricate, non-linear** interactions among billing factors, essential for **multifaceted** healthcare costs.

- **Automated Hierarchical Feature Extraction:**

Automatically **extracts** hierarchical features from diverse inputs, adapting to evolving billing structures.

- **Scalability for Growing Data Volumes:**

Handles **large-scale datasets efficiently**, ensuring **consistent performance** as billing records expand.

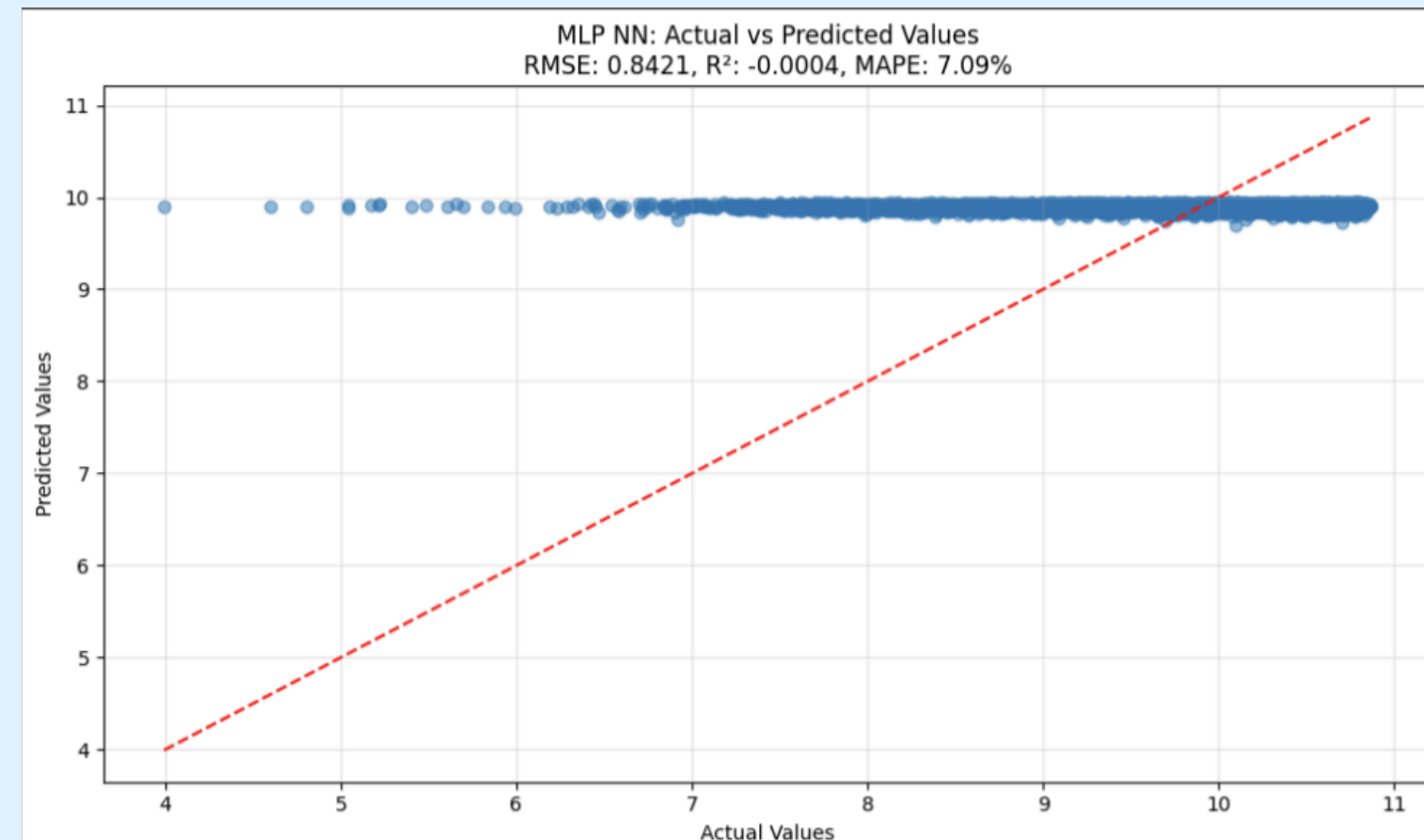
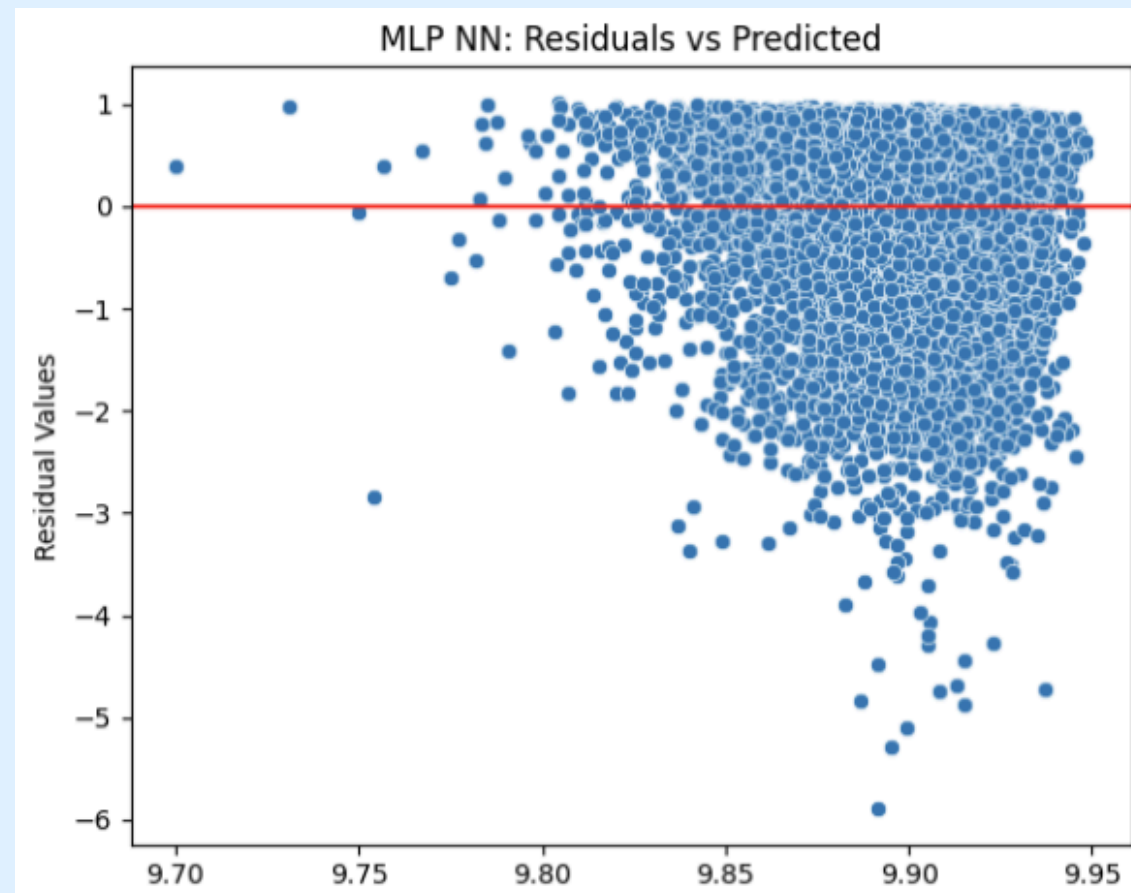
- **Enhance Predictive Accuracy for Financial Optimization:**

Non-linear activations and multi-dimensional processing lead to more **accurate billing forecasts** (financial planning and resource allocation).

## Model Training

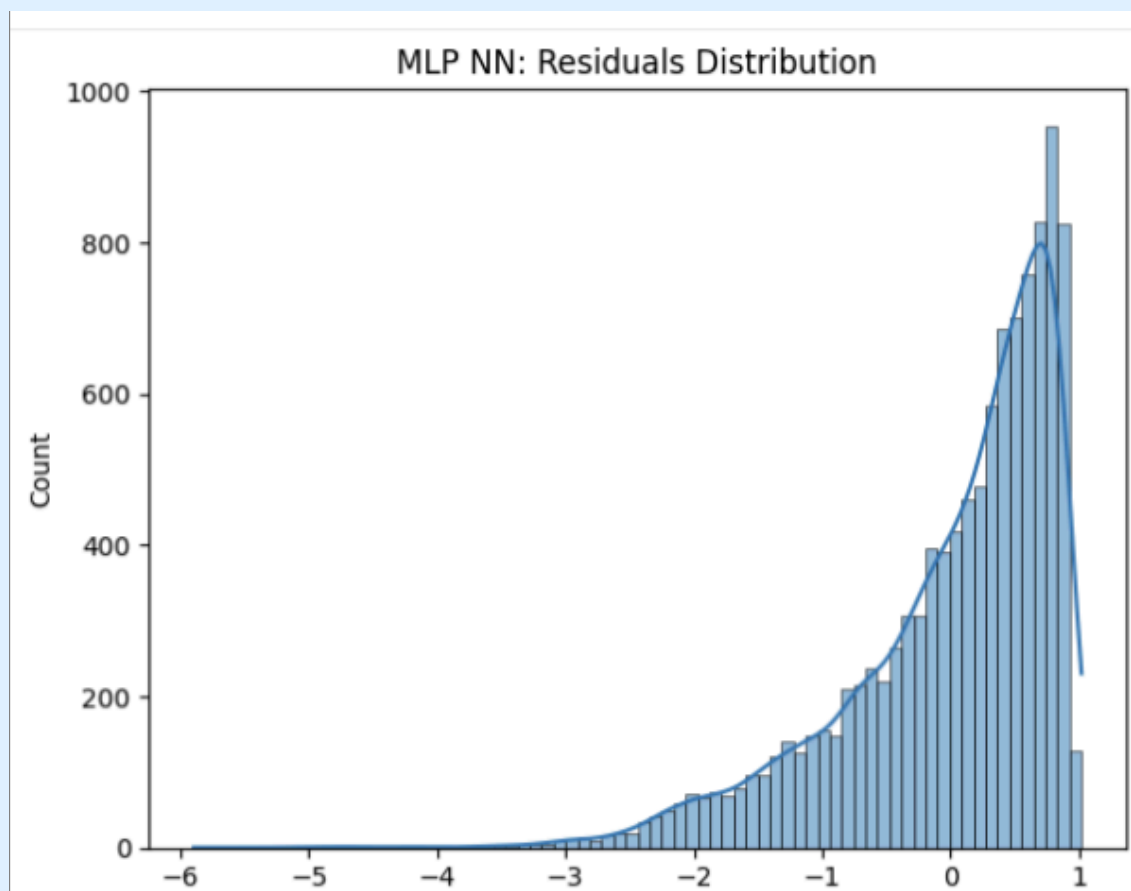
- Input layer: 28 neurons (one per feature)
- Hidden layers: [**128, 64, 32**] neurons, ReLU function
- Output layer: 1 neuron (billing prediction), Linear function
- Loss function: MSLE (Reduce Errors)
- Regularization: L2 with  $\lambda=0.001$
- Adam Optimizer with early stopping (patience=20)





### Performance

- **MAPE: 7.09%**
- **RMSE: 0.8421**
- **$R^2$ : 0.0004**



### Positives

- Improved **accurate** cost forecasts even for **evolving** treatment protocols and patient conditions.
- Flexible structure allows adjusting **seasonal variations and changes** in healthcare practices.
- By automatically extracting key billing-relevant features, support **better financial planning** and targeted interventions.

# MODEL IMPLEMENTATION

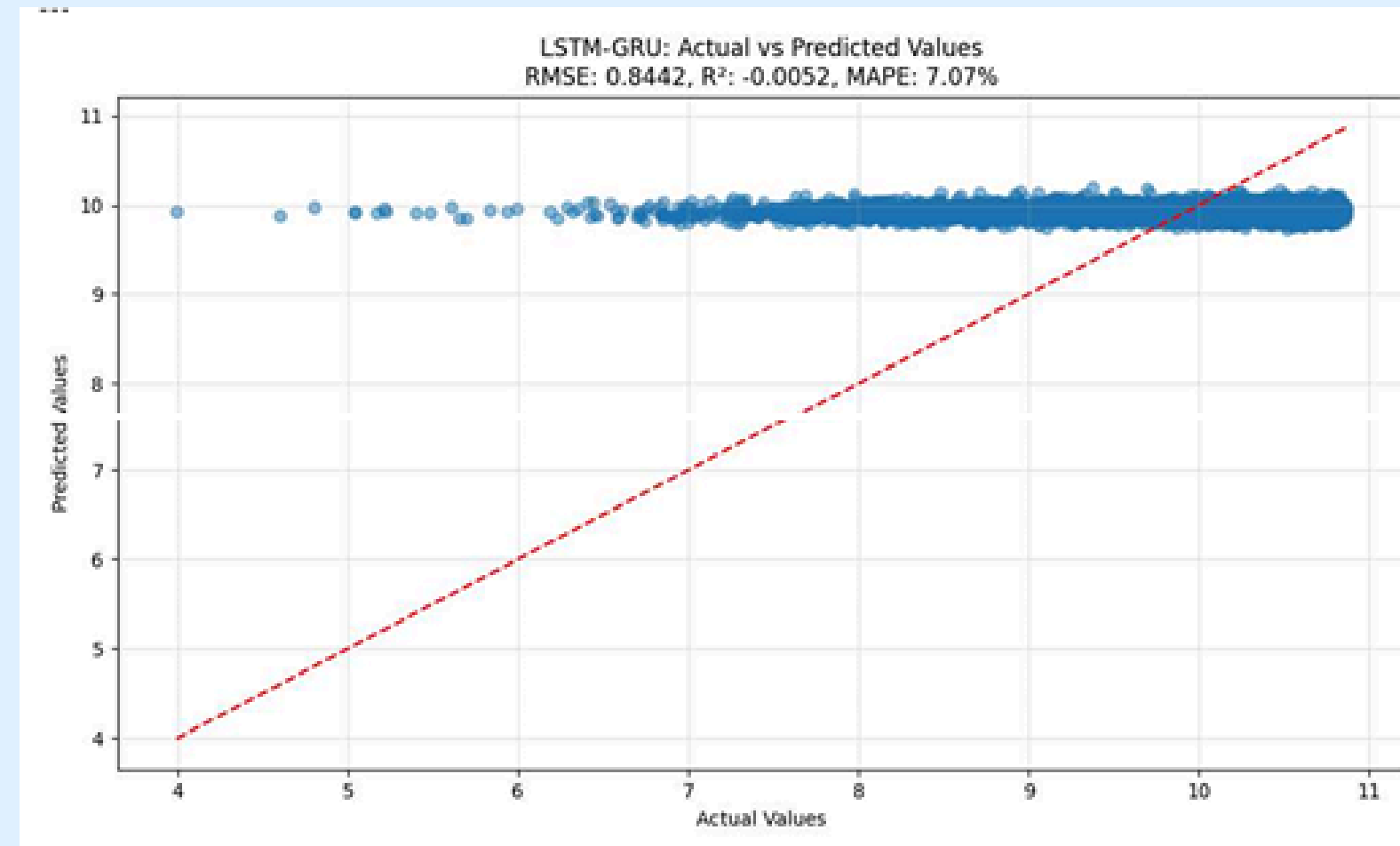
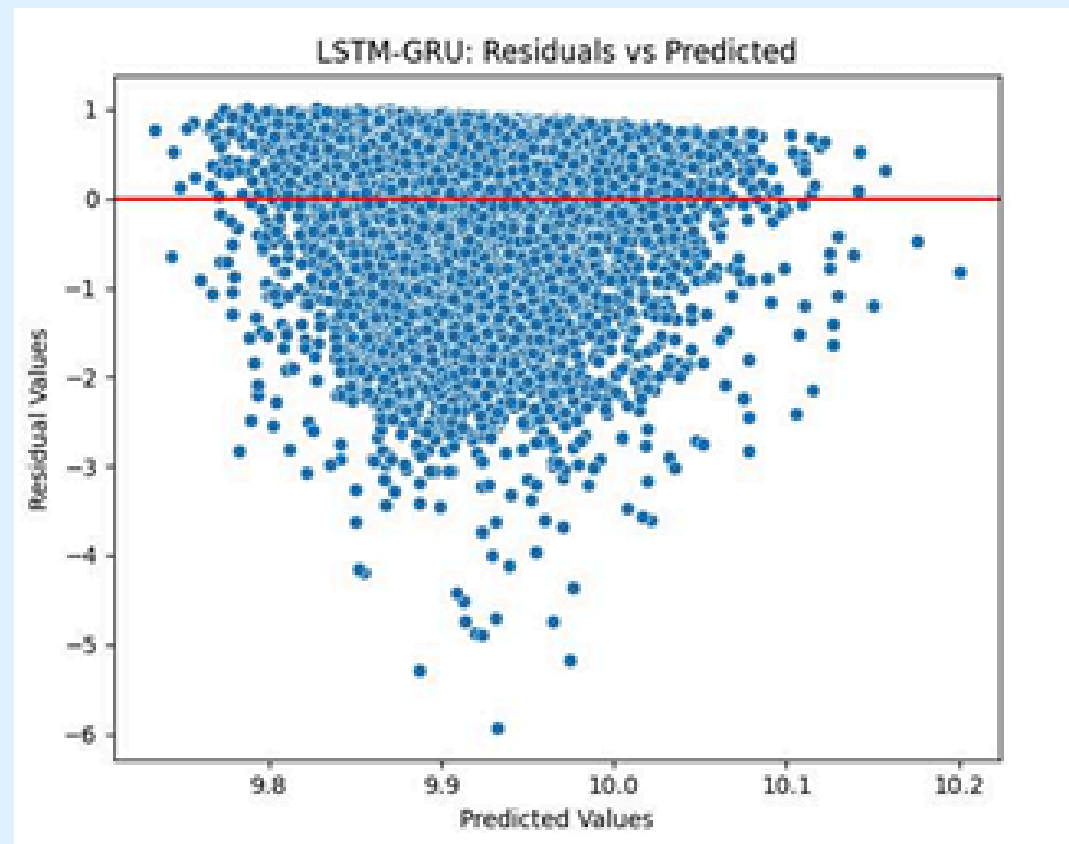
## COMBINED LSTM-GRU NEURAL NETWORK

### Why?

- Bidirectional processing captures the complete patient journey, as healthcare billing depends on both past and future events within a treatment sequence.
- Complementary strengths combine LSTM's long-term memory (for extended stays) with GRU's efficient updates (for recent status changes).
- Enhanced pattern recognition for complex healthcare billing factors, including intersecting diagnoses, procedures, and length of stay.
- Hierarchical feature extraction where LSTM layers capture fundamental patterns and GRU layers refine these into billing-relevant features.

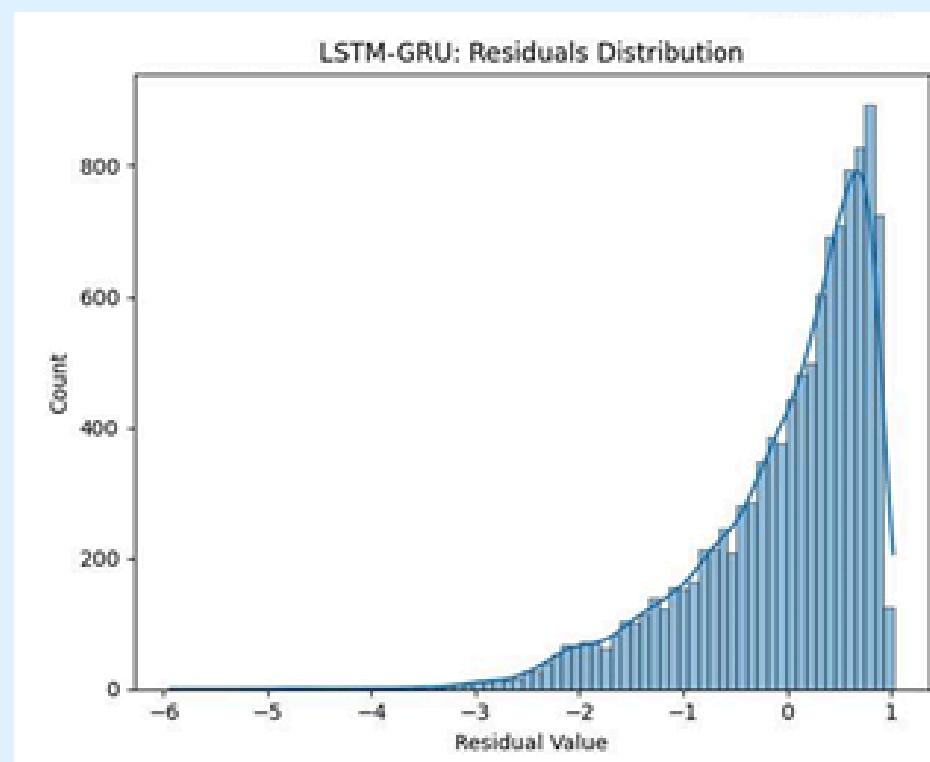
### Model Training

- Input layer: 28 features  $\times$  30 time steps (lookback window)
- Bidirectional LSTM layer (50 units)  $\rightarrow$  Batch normalization  $\rightarrow$  Dropout (0.3)
- Bidirectional GRU layer (50 units)  $\rightarrow$  Batch normalization  $\rightarrow$  Dropout (0.3)
- Dense layer: 32 neurons with ReLU activation
- Output layer: Single neuron with linear activation
- Adam optimizer
- Mean Squared Error loss function
- 50 epochs with early stopping (patience=5)
- Learning rate reduction on plateau
- Batch size of 16
- 10% validation split



#### Performance

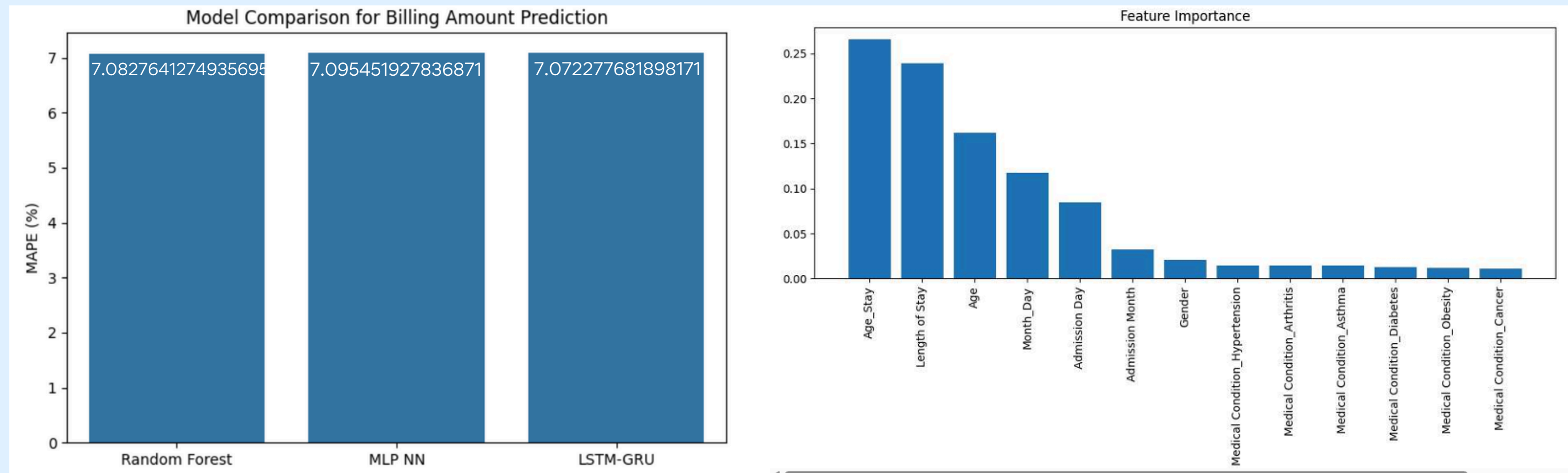
- **MAPE: 7.07%**
- **RMSE: 0.8442**
- **$R^2$ : 0.0052**



#### Positives

- Capturing complex temporal dependencies in treatment patterns
- Predicting billing amounts for extended hospital stays
- Adapting to seasonal variations in healthcare costs
- Handling patients with evolving conditions

# MODEL IMPLEMENTATION EVALUATION



Random Forest: Demonstrated solid performance with explainable predictions and robustness to outliers

MLP Neural Network: Offered competitive results with the capacity to capture non-linear relationships

LSTM-GRU Neural Network: Provided advanced sequence modeling capabilities, potentially capturing temporal patterns in patient data

The final MAPE scores (as shown in the bar plot visualization) indicated the relative accuracy of each model, with lower percentages representing better performance.





# **PART 4:** **INSIGHTS &** **RECOMMENDATION**



Business Problem	Data Analysis	Model Performance
How can we use patient data (from the healthcare dataset) to predict healthcare billing amounts?	Tested models include Random Forest Regressor, Multi-Layer Perceptron (MLP) Neural Network and Combined LSTM-GRU Neural Network.	LSTM-GRU outperformed other models by a significant margin.

Key Insights:

- 1. LSTM-GRU has the lowest error percentage of 7.07%, outperforming other models.
- 2. It shows superior performance in capturing complex temporal dependencies in treatment patterns and predicting billing amounts for extended hospital stays.
- 3. LSTM-GRU performs well in adapting to seasonal variations in healthcare costs and handling patients with evolving conditions.

# BENEFITS OF PREDICTING HEALTHCARE BILLING AMOUNTS

## Financial Benefits

Projected annual savings of around £110 million through:

- 6% reduction in average length of stay (£45M)
- 8% decrease in administrative costs related to billing (£18M)
- 4% improvement in resource allocation efficiency (£25M)
- 3% reduction in emergency readmissions (£22M)

## Operational Improvements

- 15% reduction in billing processing time
- 20% improvement in budget forecast accuracy
- Enhanced ability to identify cost outliers and inefficiencies
- More effective negotiation position with suppliers
- Better alignment of staffing with anticipated patient needs

## Patient Care Enhancements

- Reduction in hospital readmissions
- Reduced waiting times through optimized resource allocation
- Improved continuity of care and targeted preventive care
- Potential for personalized treatment plans optimized for both outcome and cost

## Key Insights:

This healthcare billing prediction solution delivers significant business value:

### Financial Planning

Allowing hospitals to accurately forecast costs and revenue based on current patient population

### Price Transparency

Enabling hospitals to provide patients with more accurate billing estimates prior to procedures

### Resource Optimization

Identifying which patient factors most heavily influence the healthcare billing amounts



# RECOMMENDATION (FOR NHS)

- ✦ · Leverage predictive models to forecast future medical billing, optimize NHS budget allocation, and reduce financial waste
- ✦ · Use predictive models to provide patients with more accurate billing estimates, enhancing trust and reducing payment disputes.
- ✦ · Refine predictive models by incorporating real-world NHS data to enhance accuracy and applicability.
- ✦ · Deploy the best-performing model into NHS billing and decision-making systems to improve operational efficiency.
- ✦ · Utilize predictive analytics for personalized patient care strategies and optimized insurance planning.
- ✦ · Integrate real-time data updates to enhance predictive accuracy and support dynamic decision-making.

# References:

Almunawar, M.N. (2020) 'Predicting outpatient appointment no-shows using data mining techniques', PhD Thesis, University of South Wales. Available at: [https://pure.southwales.ac.uk/files/13650849/PhD\\_EAI\\_Predicting\\_outpatient\\_appointment.pdf](https://pure.southwales.ac.uk/files/13650849/PhD_EAI_Predicting_outpatient_appointment.pdf)

Brilleman, S.L., Gravelle, H., Hollinghurst, S., Purdy, S., Salisbury, C. and Windmeijer, F. (2014) 'Keep it simple? Predicting primary health care costs with clinical morbidity measures', *Journal of Health Economics*, 35, pp. 109–122. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4051993/>

Healthcare Financial Management Association (2016) 'Financial forecasting in the NHS'. Available at: <https://www.hfma.org.uk/system/files/forecasting-briefing-final3.pdf>

Health Foundation. (n.d.). Cost pressures on the NHS will only grow. Retrieved from <https://www.health.org.uk/press-office/news-about-the-health-foundation/cost-pressures-on-the-nhs-will-only-grow>

Kruse, C.S., Mileski, M., Alaytsev, V., Carol, E., & Williams, A. (2018). Adoption factors associated with electronic health record among long-term care facilities: A systematic review. *BMJ Open*, 8(6), e020700. Retrieved from <https://pmc.ncbi.nlm.nih.gov/articles/PMC5977561/>

Miracolo, A., Mills, M. and Kanavos, P. (2021) 'Predictive analytic techniques and big data for improved health outcomes in the context of value-based health care and coverage decisions: a scoping review'. London School of Economics and Political Science. Available at: <https://www.lse.ac.uk/business/consulting/assets/documents/Predictive-Analytic-Techniques-and-Big-Data-for-Improved-Health-Outcomes-Final-Report.pdf>

The Health Informatics Service (2021) 'Predictive analytics: an emerging asset in the healthcare industry'. Available at: <http://www.this.nhs.uk/insights/article/predictive-analytics-an-emerging-asset-in-the-healthcare-industry>

UK Parliament. (n.d.). Written evidence submitted to the House of Commons Health and Social Care Committee. Retrieved from <https://committees.parliament.uk/writtenevidence/104535/html/>

Wyatt, S. (2018) 'Risk and reward sharing for NHS integrated care systems'. Available at: [https://www.strategyunitwm.nhs.uk/sites/default/files/2018-06/Risk%20and%20Reward%20Sharing%20for%20NHS%20Integrated%20Care%20Systems%20-%2020180605\\_0.pdf](https://www.strategyunitwm.nhs.uk/sites/default/files/2018-06/Risk%20and%20Reward%20Sharing%20for%20NHS%20Integrated%20Care%20Systems%20-%2020180605_0.pdf)

Optuna. (n.d.). Optuna - A hyperparameter optimization framework. [online] Available at: <https://optuna.org>.