Mini Project: Medical Insurance Cost Prediction with ANN

Problem Statement

This mini project involves predicting the medical insurance charges of individuals based on their personal and health-related information.The dataset contains features such as age, BMI, smoking habits, region, and number of children.Students are required to perform EDA (Exploratory Data Analysis), data preprocessing, and build models ranging from a baseline ANN to an optimized ANN. The goal is to understand how deep learning can be applied to regression problems.

Dataset Link

👉 Medical Insurance Dataset (Kaggle)

Guidelines for Students

🔹 Data Understanding

- Explore the dataset structure (`insurance.csv`).
- Identify feature types:
    - **Numerical:** `age`, `bmi`, `children`, `charges`
    - **Categorical:** `sex`, `smoker`, `region`
- Target column: **`charges`** (continuous variable).

🔹 Data Exploration (EDA)

- Plot the distribution of **charges**.
- Analyze the impact of **smoking status** on charges.
- Boxplots of **BMI vs charges**, **age vs charges**.
- Correlation heatmap for numerical features.
- Compare **average charges by region**.

🔹 Preprocessing

- Encode categorical variables (`sex`, `smoker`, `region`) using one-hot encoding.
- Scale numerical features (`age`, `bmi`, `children`) with StandardScaler.
- Split dataset into training and validation sets (80/20).

🔹 Model Building

1. **Baseline ANN**
    - Input → Dense(32, relu) → Dense(1)
    - Compile with MSE loss and Adam optimizer.
    - Evaluate with MAE and R^2 score.
2. **Optimized ANN**
    - Deeper architecture with multiple hidden layers.
    - Dropout layers to prevent overfitting.
    - EarlyStopping & ReduceLROnPlateau callbacks.

🔹 Evaluation

- Metrics: **MAE, MSE, RMSE, R^2 score**.
- Plot **actual vs predicted charges**.
- Visualize **loss curves** (training vs validation loss).

🔹 Optimization & Interpretation

- Show how **dropout** improves generalization.
- Experiment with **learning rates, batch sizes, number of neurons**.
- Interpret feature importance (e.g., smoking status has the strongest impact).

Project Tasks

🔹 Basic Level

1. Display the shape of the dataset and first 5 rows.
2. Check for missing values.
3. Plot the distribution of `charges`.
4. Find the average insurance charges by `sex`.
5. Compare insurance charges for smokers vs non-smokers.

🔹 Intermediate Level

1. Encode categorical variables and scale numerical ones.

2. Train a **baseline ANN** model.
3. Report MAE and R^2 score for baseline model.
4. Plot training vs validation loss for the baseline ANN.

🔹 Advanced Level

1. Build an **optimized ANN** with multiple hidden layers and dropout.
2. Apply **EarlyStopping** and **ReduceLROnPlateau**.
3. Compare MAE and R^2 score of baseline vs optimized model.
4. Plot **actual vs predicted charges**.
5. Discuss which features influence insurance costs the most.

Expected Outcomes

- **Basic:** Students will learn to explore datasets, visualize distributions, and identify relationships.
- **Intermediate:** Students will understand how to preprocess mixed-type data and train a regression ANN.
- **Advanced:** Students will learn how to optimize ANN architectures with dropout and callbacks, and evaluate models using regression metrics.

By the end of this project, students will demonstrate the complete pipeline:
EDA → Preprocessing → Baseline ANN → Optimized ANN → Evaluation → Interpretation.