AI/ML Internship Report

Submitted To :

DevelopersHub Corporation

Submitted By :

Kashaf Zahra

Due Date :

28th July, 2025

Task 1: News Topic Classifier Using BERT

Objective:

To classify news headlines into their correct categories (World, Sports, Business, Sci/Tech) using the AG News dataset and a fine-tuned BERT model.

Dataset:

- AG News Dataset (from Hugging Face Datasets)
- 4 news categories, with ~30,000 training samples

Tools and Methods:

- Hugging Face Transformers (bert-base-uncased)
- Fine-tuning BERT using PyTorch
- Tokenization using BertTokenizer
- Deployment using Streamlit

Key Results:

- Accuracy: 94%+ on test set
- Streamlit app allows user to input news headline and get real-time predictions
- BERT's contextual understanding enabled excellent classification performance

---

Task 2: End-to-End ML Pipeline for Telco Churn

Objective:

To build a complete scikit-learn pipeline for customer churn prediction using the Telco dataset.

Dataset:

- Telco Customer Churn Dataset (Kaggle)
- Features: tenure, contract type, monthly charges, etc.
- Target: Churn (Yes/No)

Methodology / Approach:

- Preprocessing using ColumnTransformer and Pipeline
- Feature encoding: OneHot + StandardScaler
- Models used: Decision Tree + GridSearchCV for tuning
- Final model saved as .pkl using joblib

Key Results or Observations:

- Best Accuracy: **~80%**
- Confusion matrix showed more success on predicting non-churn
- Best parameters found: max_depth=10, n_estimators=100
- Pipeline was clean, modular, and production-ready

Task 3: Multimodal Housing Price Prediction (Image + Tabular)

Objective:

To predict house prices using both tabular data (like bedrooms, bathrooms, sqft) and corresponding images.

Dataset:

- SoCal Housing Dataset from Kaggle
- Tabular file: socal2.csv
- Images folder: socal_pics (15474 images)

Methodology / Approach:

- Image feature extraction using MobileNetV2 (transfer learning)
- Combined with tabular data (bed, bath, sqft)
- Final prediction using Random Forest Regressor

Key Results or Observations:

- Combined features shape: (15474, 1283)
- R^2 Score: 0.44
- MAE: 209,714
- RMSE: 286,386
- Model captures both visual and numeric patterns, though further tuning could improve accuracy

---