SOCIAL MEDIA USAGE AND LIFESTYLE FACTORS: A MACHINE LEARNING APPROACH TO HAPPINESS PREDICTION

PREPARED BY:
ESRA YILDIZ
KEVSER KATIRCIOĞLU

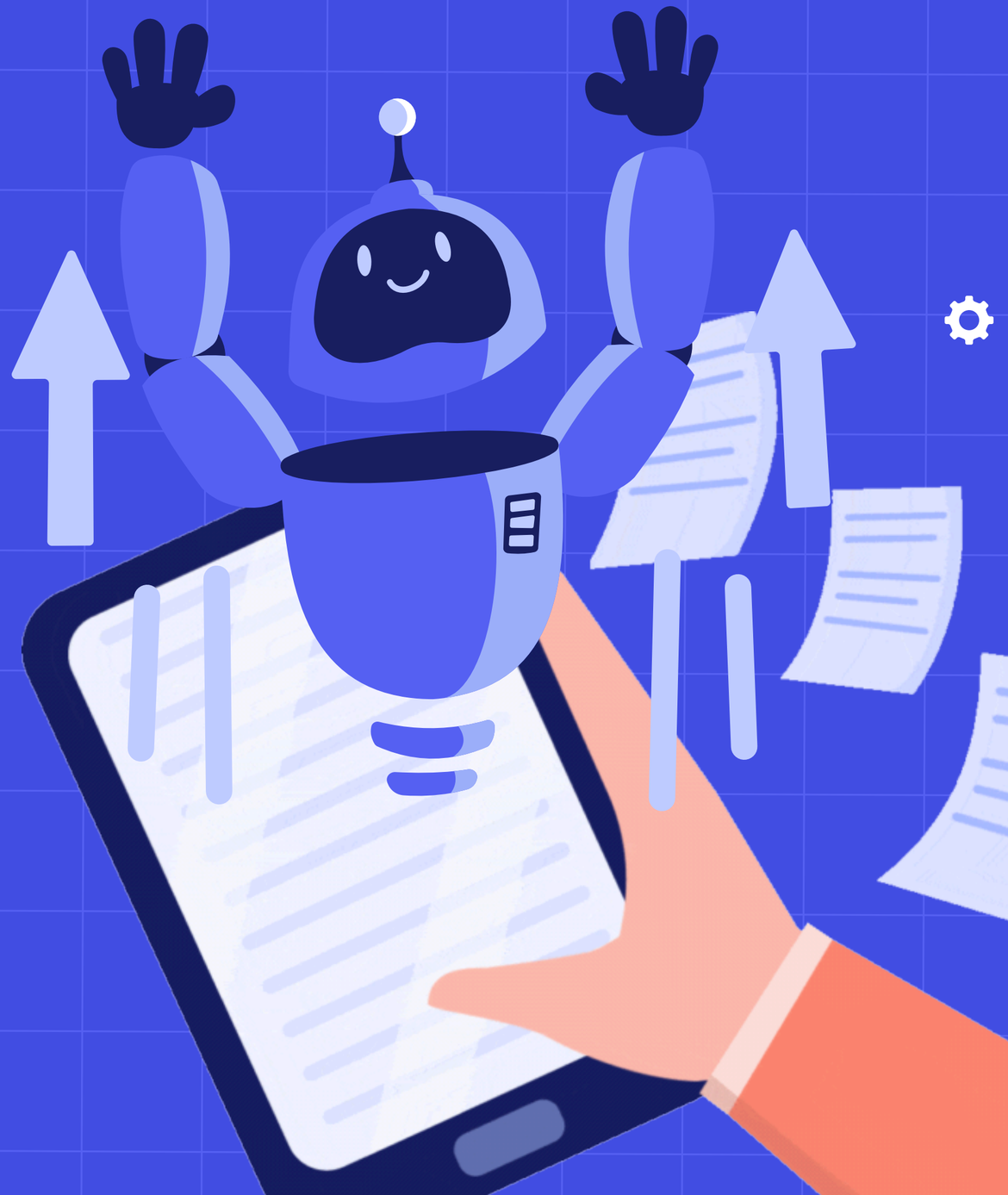
COURSE:
MACHINE LEARNING

PURPOSE OF THIS PROJECT

The main goal of this project is to analyze how lifestyle and digital habits influence the Happiness Index using machine learning techniques.

By applying regression-based models, we aim to:

- Understand the relationship between social media usage and happiness
- Identify the most important factors affecting happiness
- Provide data-driven insights for healthier digital habits



PROBLEM DEFINITION

In the digital age, people spend an increasing amount of time on social media and digital devices.
While social media provides communication and entertainment benefits, excessive screen time may negatively affect mental health factors such as sleep quality, stress levels, and overall happiness.
The main problem is that it is unclear which lifestyle factors have the strongest impact on happiness, and whether screen time alone is a reliable predictor of happiness.

RESEARCH QUESTIONS

- How does daily screen time affect the Happiness Index?
- Do lifestyle factors such as sleep quality, stress level, and exercise frequency significantly influence happiness?
- Is a model using multiple lifestyle factors more accurate than a model using only digital usage variables?
- Which factors are the most important predictors of happiness?

DATASET OVERVIEW

Dataset Name

Social Media and Mental Health Balance Dataset

Data Format

CSV (Comma-Separated Values)

Number of Samples

500 user records

Number of Features

9 input features + 1 target variable

Target Variable

Happiness Index (continuous value)

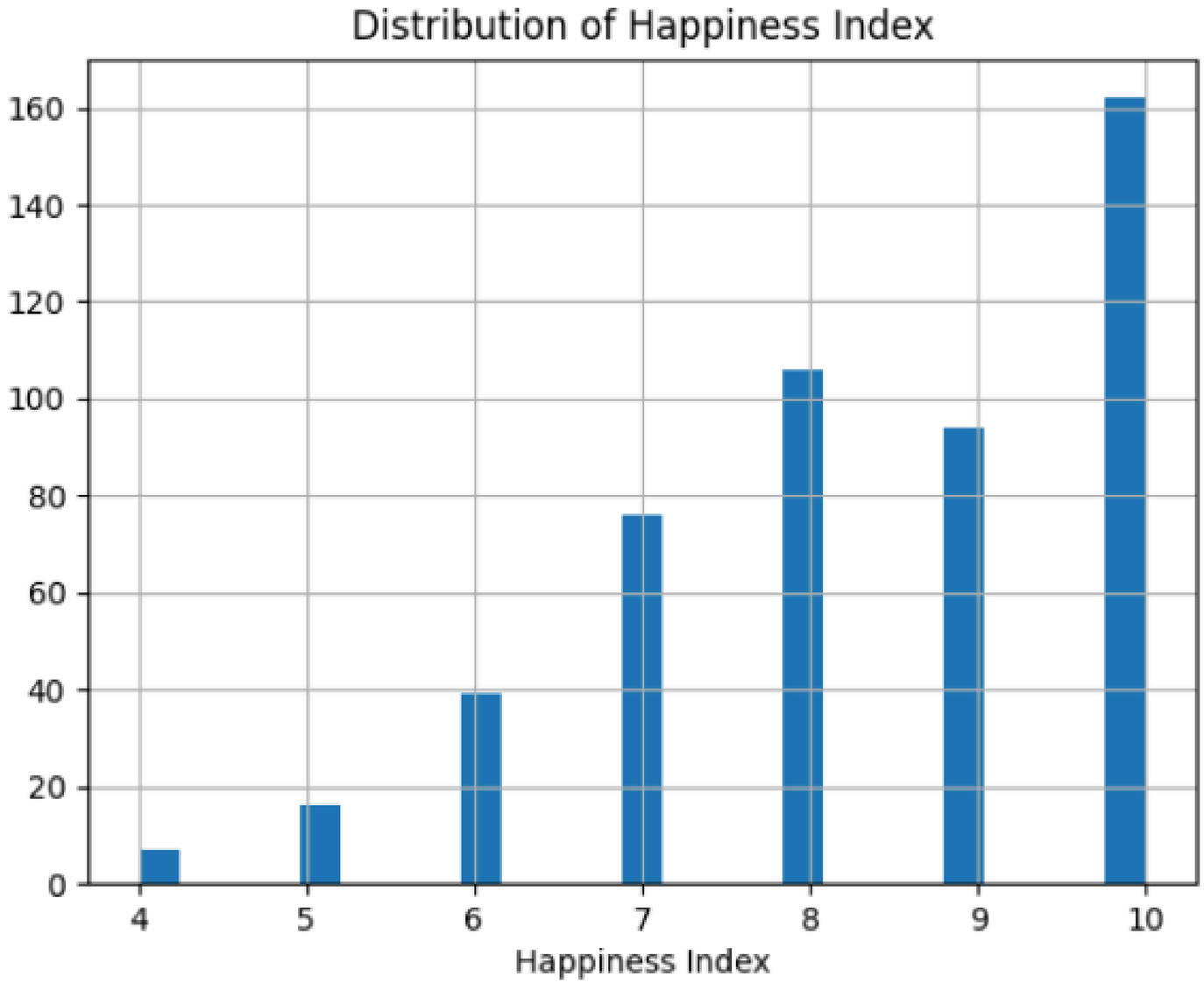
Feature Examples

- Daily Screen Time (hours)
- Sleep Quality (1–10 scale)
- Stress Level (1–10 scale)
- Exercise Frequency (per week)
- Days Without Social Media
- Age, Gender

This step gives a general understanding of the dataset structure before analysis and modeling.

EXPLORATORY DATA ANALYSIS (EDA)

- We explored the dataset to understand its structure and quality.
- Checked data types, missing values, and statistical distributions.
- Visualized the distribution of the Happiness Index.
- Observed general trends before model training.



This histogram shows the distribution of the Happiness Index.
Most users report high happiness scores, indicating a generally positive well-being level in the dataset.

MODELS USED IN THE PROJECT

- **Linear Regression**

Used as a baseline model to capture linear relationships.

- **Decision Tree Regressor**

Used to model non-linear patterns in the data.

- **Random Forest Regressor**

An ensemble method used to improve prediction accuracy and reduce overfitting.

DATA LOADING

- The dataset is loaded directly from a CSV file.
- This step initializes the data analysis pipeline and prepares the raw data for processing.

```python
DATA_PATH = Path(__file__).parents[1] / "data" / "raw" / "Mental_Health_and_Social_Media_Balance_Dataset.csv"

def load_data(path=DATA_PATH):  2 usages
    print("\n--- Loading Dataset ---")
    df = pd.read_csv(path)
```

```python
def train_models(X_train, X_test, y_train, y_test):  4 usages
    print("\n--- Model Training Started ---")

    models = {
        "Linear Regression": LinearRegression(),
        "Decision Tree": DecisionTreeRegressor(random_state=42),
        "Random Forest": RandomForestRegressor(n_estimators=200, random_state=42)
    }

    results = {}
```

MODEL CREATING

- All algorithms are defined here.
- Three models are prepared for training and stored in a dictionary.
- This structure allows all models to be processed within a single loop.

MODEL TRAINING

- Each model is trained using the training dataset.
- Predictions are generated on the test dataset.
- The same steps are automatically repeated for all models.

```
for name, model in models.items():
    print(f"\n► {name} being trained...")
    model.fit(X_train, y_train)
    preds = model.predict(X_test)
```

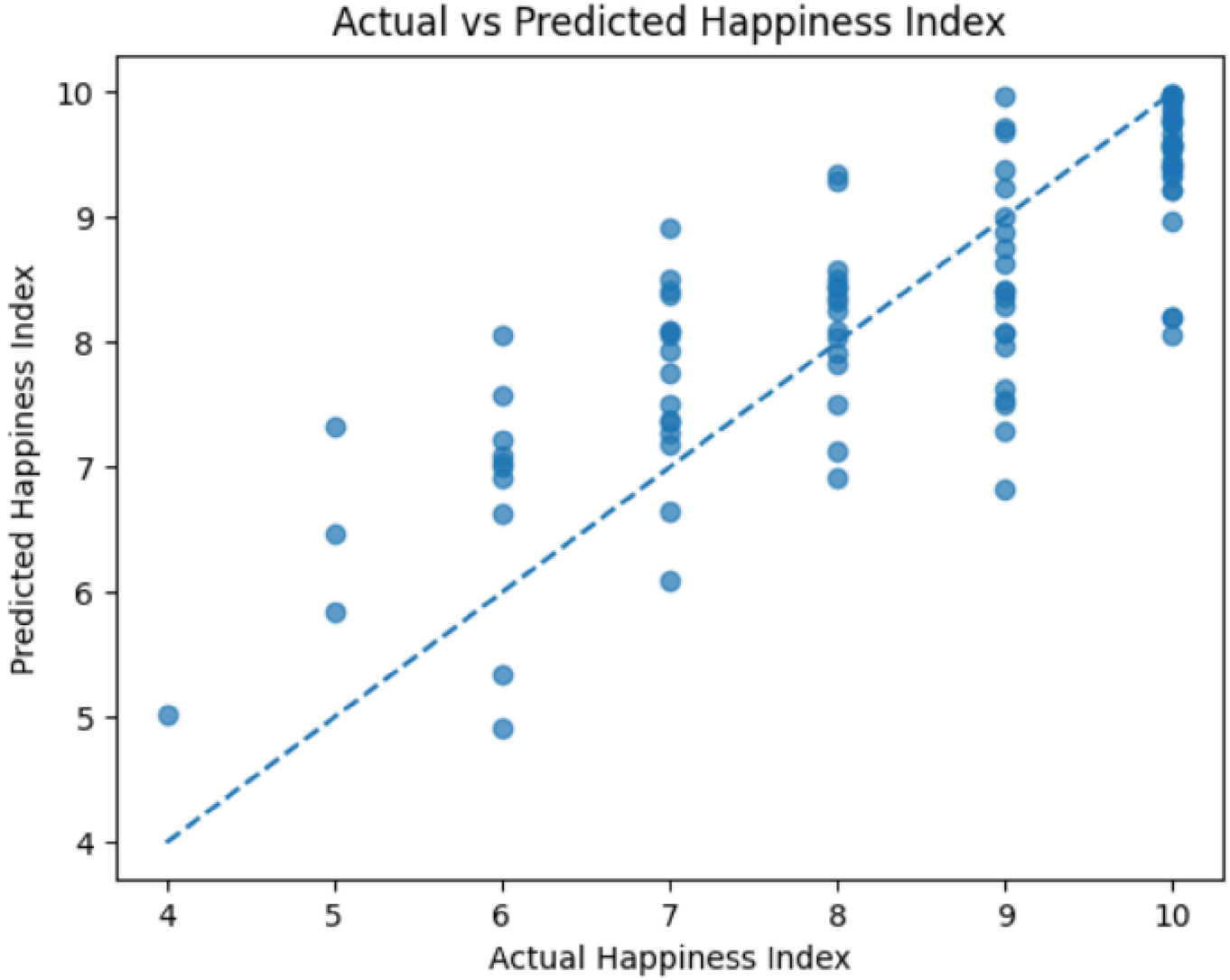
```
mae = mean_absolute_error(y_test, preds)
mse = mean_squared_error(y_test, preds)
r2 = r2_score(y_test, preds)

results[name] = {
    "MAE": mae,
    "MSE": mse,
    "R2": r2
}
```

EVALUATION

- Models are evaluated based on MAE, MSE, and R^2 scores.
- These metrics make it easier to compare the performance of different models.

ACTUAL VS PREDICTED HAPPINESS

- Compares model predictions with the true happiness scores.
- Points close to the diagonal line indicate high prediction accuracy.
- Provides a visual summary of the model’s overall performance.



Displays how close model predictions are to the true values. A good model aligns points near the diagonal line.

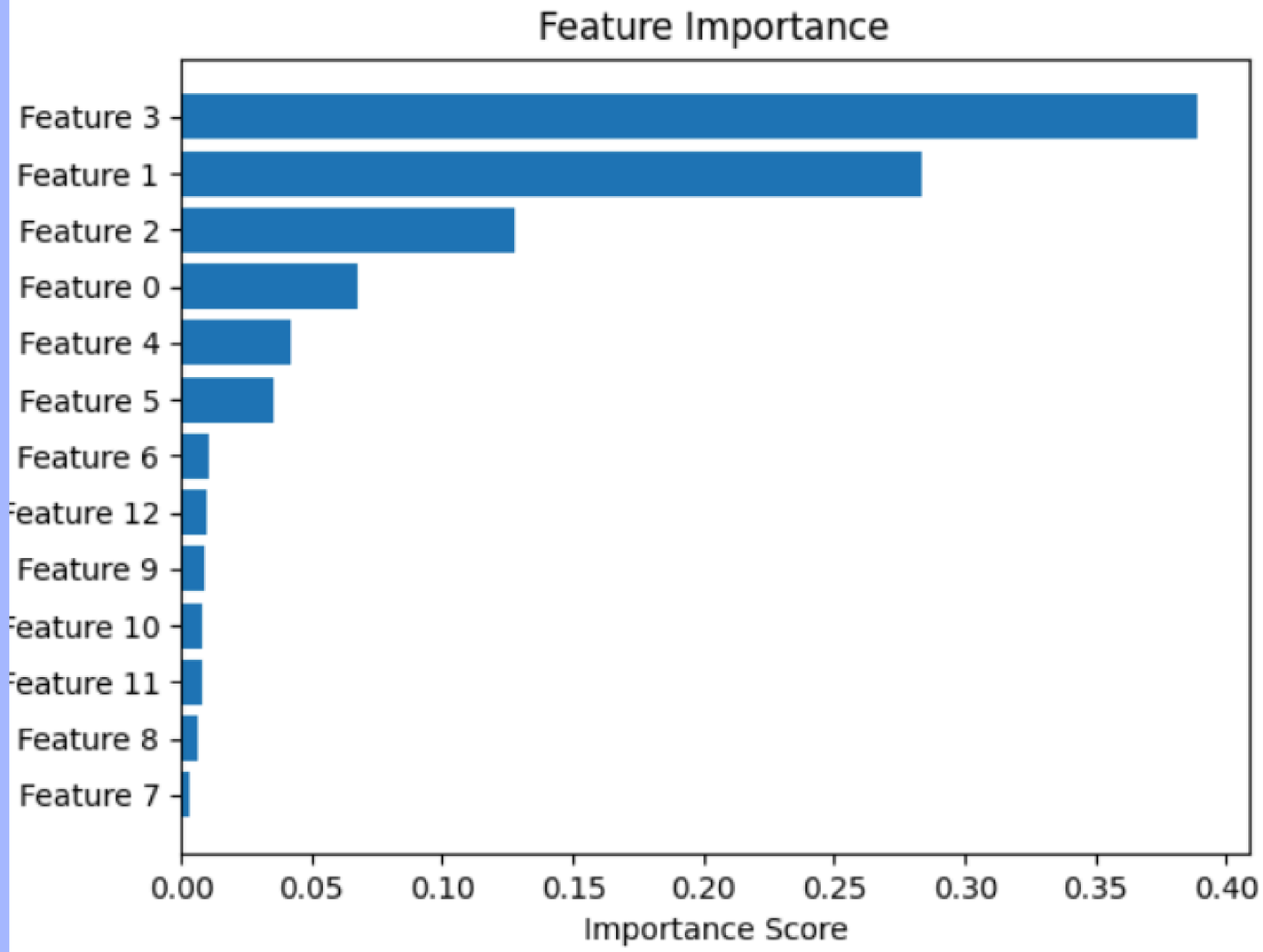
PREDICTION ERROR DISTRIBUTION

- Visualizes the distribution of prediction errors (predicted – actual).
- Errors concentrated near zero indicate strong model accuracy.
- Helps identify outliers or bias in the predictions.

Shows how happiness scores are spread across the dataset.

FEATURE IMPORTANCE

- Highlights which features had the strongest influence on predictions.
- Helps interpret the model’s decision-making process.
- Shows the relative contribution of each variable to the final output.



Shows which features contributed most to the model’s predictions.

CONCLUSION

- Multiple models were trained and compared.
- Random Forest achieved the best overall accuracy.
- Key factors affecting happiness were identified through feature importance analysis.
- The model provides a simple and interpretable approach to understanding social media's impact on well-being.

THANKS

FOR LISTENING