

**Lahore College for Women University, Lahore**

**Department: Software Engineering**

**Submitted By: Iram Ahmad**

**Roll no: 2225165111**

**Submitted to: Sir Mohsin Nazir**

**Semester: VII**

**Session: 2022-2026**

**Assignment no 6**

**Course: Applied Data Science with AI**

**Semester:** BSSE 7th  
**Week #:** 6  
**Student Name:** Iram Ahmad  
**Roll Number:** 2225165111  
**Project Title:** House Price Prediction  
**GitHub Link:** https://github.com/Iram-Ahmad/Data-Science-AI-Course

**1. Reading Summary (½–1 page)**

**Reading Material for this Week:**

* Scikit-Learn Classification Documentation
* Kaggle Intro to Machine Learning Tutorials

**Key Learnings:**

1. Understood the concept of **classification algorithms**, including Logistic Regression, Decision Trees, and Random Forests.
2. Learned how to evaluate classification models using **accuracy, confusion matrix, and classification reports**.
3. Explored **train/test splitting and feature selection** to prepare data for model training.

**Reflection:**  
These readings helped me understand how classification can be applied in real-world data science problems. In my **House Price Prediction** project, I used these techniques to classify houses into *High* and *Low* price categories. This step is essential to transform the regression problem into a classification one, which helps in understanding market segmentation and price distribution.

**2. Classroom Task Documentation**

**Task Performed in Class:**

* Trained Decision Tree and Random Forest models using scikit-learn.
* Compared their accuracy and discussed how ensemble methods (like Random Forest) improve prediction stability.

**Screenshots / Code Snippets:**

rf\_model = RandomForestClassifier(n\_estimators=100, random\_state=42)

rf\_model.fit(X\_train, y\_train)

rf\_acc = accuracy\_score(y\_test, rf\_model.predict(X\_test))

print("Random Forest Accuracy:", rf\_acc)

**3. Weekly Assignment Submission**

**Assignment Title:**  
*Classification Models: Logistic Regression and Random Forest*

**Steps Taken:**

1. Converted SalePrice into a categorical variable — “High” and “Low” — based on the median price.
2. Selected top three correlated features: OverallQual, GrLivArea, and GarageCars.
3. Split the dataset into training (80%) and testing (20%) sets.
4. Trained both **Logistic Regression** and **Random Forest** models.
5. Evaluated performance using **accuracy** and **confusion matrices**.
6. Compared both models using a bar chart for visualization.

**Output:**

* Logistic Regression Accuracy ≈ 0.85
* Random Forest Accuracy ≈ 0.90
* Random Forest performed better due to its ensemble learning nature.

**Challenges Faced:**

* Needed to ensure correct data path to load the cleaned dataset.
* Adjusted logistic regression parameters (max\_iter=1000) to avoid convergence warnings.

**GitHub Link:**

https://github.com/Iram-Ahmad/Data-Science-AI-Course

**4. Project Progress Milestone**

**This Week’s Milestone:**

* Built and compared two classification models (Logistic Regression and Random Forest).
* Identified Random Forest as the better baseline model.

**Next Week’s Goal:**

* Perform **hyperparameter tuning** and **cross-validation** to improve classification performance.

**5. ✅ Self-Evaluation**

☑️ I completed all tasks on time.  
⬜ I partially completed the tasks.  
⬜ I struggled with this week’s tasks and need help.

**6. Questions for Instructor (Optional)**

* How can we further improve Random Forest accuracy — should we adjust max\_depth or n\_estimators first?
* Is it recommended to use feature scaling for tree-based models?