RESEARCH PROBLEM

A research problem, in general, refers to some difficulty which a researcher experiences in the context of either a theoretical or practical situation and wants to obtain a solution for the same. Usually we say that a research problem does exist if the following conditions are met with:

(i) There must be an individual (or a group or an organisation), let us call it 'I,' to whom the problem can be attributed. **The individual or the organisation, as the case may be, occupies an environment, say 'N', which is defined by values of the uncontrolled variables, Yj .**

(ii) There must be at least two courses of action, say C1 and C2 , to be pursued. A course of action is defined by one or more values of the controlled variables. For example, the number of items purchased at a specified time is said to be one course of action.

(iii) There must be at least two possible outcomes, say O1 and O2 , of the course of action, of which one should be preferable to the other. In other words, this means that there must be at least one outcome that the researcher wants, i.e., an objective.

(iv) The courses of action available must provides some chance of obtaining the objective, but they cannot provide the same chance, otherwise the choice would not matter. Thus, if P (Oj | I, Cj , N) represents the probability that an outcome O_j will occur, if I select Cj in N, then $P(O_1|I,C_1,N) \neq P(O_1|I,C_2,N)$

The components of a research problem as under

- There must be an individual or a group which has some difficulty or the problem
- There must be some objective(s) to be attained at. If one wants nothing, one cannot have a problem.
- There must be alternative means (or the courses of action) for obtaining the objective(s) one wishes to attain. This means that there must be at least two means available to a researcher for if he has no choice of means, he cannot have a problem
- There must remain some doubt in the mind of a researcher with regard to the selection of alternatives. This means that research must answer the question concerning the relative efficiency of the possible alternatives
- There must be some environment(s) to which the difficulty pertains.

Selecting the problem..

- Subject which is overdone should not be normally chosen ex:"**The Rise of Netflix and Its Impact on Traditional TV**", for it will be a difficult task to throw any new light in such a case ex:***"How Netflix's AI Algorithms Influence Viewer Retention Compared to Amazon Prime"***
- Controversial subject should not become the choice of an average researcher

  Ex:"**OTT Platforms and the Rise of Explicit Content: Should Governments Regulate?**"

  Alternate that can be done

  "Parental Control Features in OTT Platforms: A Comparative Study of Netflix, Disney+, and Prime Video"

- Too narrow or too vague problems should be avoided

  Example (Too Broad): *"The Future of OTT Platforms"*

  Example (Too Narrow) "Impact of Netflix's 'Stranger Things' Season 3 on Teenagers in New York City" – This is **too specific**, making it hard to collect **diverse data**.

  Better Alternative:

  "Binge-Watching Behavior: A Study on Viewer Engagement Across OTT Platforms" – This is **focused but broad enough** to gather sufficient data.

- The subject selected for research should be **familiar and feasible** so that the related research material or sources of research are within one's reach. Even then it is quite difficult to supply definitive ideas concerning how a researcher should obtain ideas for his research. For this purpose, a researcher should contact an expert or a professor in the University who is already engaged in research. He may as well read articles published in current literature available on the subject and may think how the techniques and ideas discussed therein might be applied to the solution of other problems. He may discuss with others what he has in mind concerning a problem. In this way he should make all possible efforts in selecting a problem

- The importance of the subject, the qualifications and the training of a researcher, the costs involved, the time factor are few other criteria that must also be considered in selecting a problem. In other words, before the final selection of a problem is done, a researcher must ask himself the following questions:
    - **Whether he is well equipped in terms of his background to carry out the research?**
    - **Whether the study falls within the budget he can afford?**
    - **Whether the necessary cooperation can be obtained from those who must participate in research as subjects**
- The selection of a problem must be preceded by a preliminary study. This may not be necessary when the problem requires the conduct of a research closely similar to one that has already been done. But when the field of inquiry is relatively new and does not have available a set of well developed techniques, a brief feasibility study must always be undertaken

NECESSITY OF DEFINING THE PROBLEM

The problem to be investigated must be defined unambiguously for that will help to discriminate relevant data from the irrelevant ones.

- What data are to be collected?
- What characteristics of data are relevant and need to be studied?
- What relations are to be explored?
- What techniques are to be used for the purpose?
- Similar other questions crop up in the mind of the researcher who can well plan his strategy and find answers to all such questions only when the research problem has been well defined.

TECHNIQUE INVOLVED IN DEFINING A PROBLEM

- Statement of the problem in a general way
- Understanding the nature of the problem
- Surveying the available literature
- Developing the ideas through discussions
- Rephrasing the research problem into a working proposition.

Defining the Research Problem

- **Technical terms and words or phrases**, with special meanings used in the statement of the problem, should be clearly defined.
- Basic **assumptions** or postulates (if any) relating to the research problem should be **clearly stated**
- A straightforward statement of the value of the investigation (i.e., the criteria for the selection of the problem) should be provided.
- The suitability of the time-period and the sources of data available must also be considered by the researcher in defining the problem.
- The scope of the investigation or the limits within which the problem is to be studied must be mentioned explicitly in defining a research problem

Example :Reviewing the literature survey

problem:"Predicting house prices using ML"

Searching for the Existing Literature in Your Area of Study

- Previous ML models used for house price prediction.
- Feature selection techniques.
- Datasets used (e.g., Kaggle's housing datasets, Zillow data).
- Evaluation metrics like RMSE, MAE, R^2.

Example: Reviewing papers on regression models (Linear Regression, Random Forest, XGBoost) and deep learning approaches used for predicting real estate prices.

2. Reviewing the Selected Literature

After gathering literature, the next step is to analyze and compare different methodologies. This helps in:

- Understanding the strengths and weaknesses of different models.
- Identifying best practices for data preprocessing and feature engineering.
- Exploring challenges like handling missing data or outliers.

- **Example**: A study might show that Random Forest performs well on structured tabular data, while Neural Networks require more data for better performance. Based on this, the decision might be to experiment with ensemble methods.

3. Developing a Theoretical Framework

A theoretical framework explains the relationships between different variables based on existing theories or principles.

- **Example**: In a house price prediction project, the _____**Pricing Model** can be a theoretical foundation. It suggests that property prices are influenced by:

- Structural attributes (e.g., number of rooms, square footage).
- Location-based factors (e.g., proximity to schools, crime rates).
- Economic factors (e.g., interest rates, inflation).

This theory helps justify why certain features are chosen for the ML model.

4. Developing a Conceptual Framework

A conceptual framework visualizes how the variables interact in your specific study. It defines:

- **Input variables** (independent features like house size, location, number of bedrooms).
- **Process** (ML model selection, training, testing).
- **Output variable** (predicted house price).

🔹 **Example**:

- **Input Features**: Square footage, number of rooms, location, proximity to amenities.
- **Processing**: Data preprocessing (handling missing values, feature scaling), model training (Random Forest, XGBoost).
- **Output**: Predicted house price.

Example :Reviewing the literature survey

Housing Price Prediction Considering Security Factors

1. Searching for the Existing Literature in Your Area of Study

- **Research papers** on how crime rates and security impact real estate prices.
- **ML models** used for house price prediction (Linear Regression, Random Forest, XGBoost, Neural Networks).
- **Datasets** that include security-related variables (e.g., crime rate, proximity to police stations, gated communities).
- **Real estate economic theories** explaining price fluctuations due to security.

🔹 **Example Search:**

- "Impact of Crime Rate on Housing Prices – A Machine Learning Approach"
- "Hedonic Pricing Model in Real Estate Security"
- "Effect of Smart Security Features on Residential Property Values"

2. Reviewing the Selected Literature

After collecting research papers, the next step is to analyze their findings:

- **How security factors (crime rate, police proximity) affect housing prices**.
- **Which ML techniques are most accurate for predicting house prices?**
- **What preprocessing steps and feature engineering are used in previous studies?**

🔹 **Example Insights from Literature:**

- Studies show that **high-crime areas reduce property values by 10-20%**.
- **Properties near police stations sell for higher prices** due to increased safety.
- **Gated communities and CCTV presence increase home prices** by making them more attractive to buyers.
- **Random Forest and XGBoost models perform best** for price prediction due to non-linearity.

🔹 **Decision Based on Review:**

- Include **crime rate, security features, and police station distance** as predictive variables.
- Use **Random Forest/XGBoost** instead of basic Linear Regression.

3. Developing a Theoretical Framework

A **theoretical framework** explains the economic and social relationships between security and housing prices.

Economic Theory: Hedonic Pricing Model

The **Hedonic Pricing Model (HPM)** states that property prices depend on various characteristics:

- **Structural Factors** (size, bedrooms, bathrooms).
- **Location Factors** (crime rate, security, proximity to police).
- **Economic Factors** (demand, interest rates).

🔹 **Applying HPM to Security-Based Pricing:**

Price=f(Structural Features,Security Features,Economic Conditions)Price = f(\text{Structural Features}, \text{Security Features}, \text{Economic Conditions})Price=f(Structural Features,Security Features,Economic Conditions)

Where:

- **High crime rate (⬆️) → Decrease in price (⬇️)**
- **Proximity to police station (⬆️) → Increase in price (⬆️)**
- **Gated community, CCTV, security systems (⬆️) → Higher valuation (⬆️)**

3. Developing a Theoretical Framework

A **theoretical framework** explains the economic and social relationships between security and housing prices.

Economic Theory: Hedonic Pricing Model

The **Hedonic Pricing Model (HPM)** states that property prices depend on various characteristics:

- **Structural Factors** (size, bedrooms, bathrooms).
- **Location Factors** (crime rate, security, proximity to police).
- **Economic Factors** (demand, interest rates).

🔹 **Applying HPM to Security-Based Pricing:**

Price=f(Structural Features,Security Features,Economic Conditions)Price = f(\text{Structural Features}, \text{Security Features}, \text{Economic Conditions})Price=f(Structural Features,Security Features,Economic Conditions)

Where:

- **High crime rate (⬆️) → Decrease in price (⬇️)**
- **Proximity to police station (⬆️) → Increase in price (⬆️)**
- **Gated community, CCTV, security systems (⬆️) → Higher valuation (⬆️)**

4. Developing a Conceptual Framework

A **conceptual framework** visually represents how input features influence house price predictions in our ML model.

- **Inputs (Independent Variables)**:
  - **Structural Attributes:** Sq. ft, bedrooms, bathrooms, age.
  - **Security Factors:** Crime rate, police station distance, presence of CCTV, gated community.
  - **Economic Factors:** Interest rates, market trends.
- **Processing (Machine Learning Pipeline)**:
  - **Data Preprocessing:** Handle missing values, normalize features.
  - **Feature Engineering:** Convert security factors into numeric values.
  - **Model Selection:** Use Random Forest/XGBoost for price prediction.
- **Output (Dependent Variable)**:
  - **Predicted House Price**.

MODULE 1 Modern Programming Quality requirements

Reliability:How often the results of the programme are correct

Definition: The model should consistently provide accurate predictions under similar conditions.

🔹 **Scenario:**

- Suppose a user inputs the same house details multiple times (location, crime rate, security features, etc.), and the model should return the **same or very similar price predictions** each time.
- If restrained with a larger dataset, the model should still **align with past trends** and not produce erratic results.

✅ **Good Reliability:**

- Predictions remain **consistent** with new data updates.
- Minor changes in inputs do not cause **drastic changes** in price estimates.

❌ **Poor Reliability:**

- The same house is predicted at **₹50 lakh** today and **₹80 lakh** tomorrow with no logical reason.

Robustness (Handling Unexpected Inputs & Noisy Data)

how well a programme anticipates problems not due to programmer error.

Definition: The model should perform well even with **incomplete, incorrect, or extreme** input values.

🔹 **Scenario:**

- A user enters a house with **999 bedrooms**, which is unrealistic.
- The dataset has **missing security data** (CCTV presence unknown).
- Sudden **crime surge in an area** drastically changes security perception.

✅ **Good Robustness:**

- The model **flags unrealistic inputs** (e.g., bedrooms > 10) and asks for corrections.
- It **handles missing values** by estimating based on similar houses.
- It **adjusts to sudden changes** in crime trends using real-time updates.

❌ **Poor Robustness:**

- **Crashes** when a field is missing.
- Assigns **extremely high or low prices** to unrealistic inputs without validation.

Usability (Ease of Use for Non-Technical Users)

The ergonomics of a program: the ease with which a person can use the programme for its intended purpose, or in some cases even unanticipated purposes.

Definition: The system should be user-friendly, requiring minimal technical knowledge.

🔹 **Scenario:**

- A real estate agent with **no ML knowledge** wants to use the app.
- A buyer wants a **simple way** to input house details and get a price prediction.

✅ **Good Usability:**

- The interface is **simple** (dropdowns, sliders for features like crime rate, security).
- Shows **why a house is priced higher/lower** (e.g., “Low crime area = ₹5 lakh increase”).
- Allows **voice input or automatic location detection** for ease.

❌ **Poor Usability:**

- Requires the user to **manually enter complex data (e.g., crime rate index)**.
- Produces **black-box predictions** without explanations.

Portability (Running on Different Devices & Platforms)

The range of computer hardware and operating system platforms on which the source code of a programme can be compiled/interpreted and run.

Definition: The model should work across **various devices** and **operating systems**.

🔹 **Scenario:**

- A user accesses the prediction system on a **mobile app, web browser, and desktop software**.
- The model is deployed **on the cloud (AWS, Google Cloud) and local machines**.

✅ **Good Portability:**

- The ML model **runs seamlessly** on Windows, macOS, and mobile (React Native app).
- It is available as a **web app, mobile app, and API** for integration with other platforms.

❌ **Poor Portability:**

- Works only on **a specific OS (e.g., only Windows)**.
- Requires **high-end GPUs** and fails on low-power devices.

Maintainability (Ease of Updating & Improving the Model)

The ease with which a programme can be modified by its present or future developers in order to make improvements or customizations,fix bugs and security holes, or adapt it to new

Definition: The system should be **easy to update**, retrain, and modify over time.

🔹 **Scenario:**

- A new dataset with **2025 housing security trends** needs to be integrated.
- The government changes the way **crime rates are reported**.
- **New features** (smart locks, AI security cameras) need to be added.

✅ **Good Maintainability:**

- The model **can be retrained easily** with new data without complete redesign.
- Uses **modular code** (separating data preprocessing, training, and deployment).
- Updates can be **deployed automatically** without breaking user access.

❌ **Poor Maintainability:**

- Hardcoded parameters make it **difficult to add new features**.
- Requires a **complete overhaul** for every update.

Efficiency/Performance (Speed & Resource Optimization)

The amount of system resources a programme consumes (processor time,memory space, slow devices such as disks, network bandwidth and to some extent even user Interaction): the less, the better.

Definition: The model should run **quickly** without **consuming excessive computational resources**.

🔹 **Scenario:**

- A user requests a **price prediction**, and it should **return results in seconds**.
- The model is **running on mobile devices** with limited processing power.
- The dataset grows **from 10,000 houses to 1 million houses**.

✅ **Good Efficiency:**

- Uses **optimized ML algorithms** (e.g., XGBoost instead of deep learning for speed).
- Predictions are generated in **under 2 seconds**.
- Uses **cloud processing** for heavy tasks while keeping mobile performance smooth.

❌ **Poor Efficiency:**

- Takes **too long** to generate a price (e.g., 10+ seconds).
- Uses **too much RAM and crashes** on mobile phones.