A. Problem Statement and Motivation

Many UVic Computer Science students are seeking their first job or internship.
However, many of them do not have a clear understanding of the salary they might expect given their current skill sets, or how their potential salary could grow over the next 3 to 5 years if they pursue roles such as data analyst or full-stack developer.

Why is this problem important?

Salary is often considered a sensitive topic. Without real industry experience, students may struggle to estimate their future salary range. Additionally, they may be unsure which skills or types of experience would lead to better pay in the job market.

We propose building a web application that helps students evaluate their potential salary based on their current skills, experience, job type, and location. This tool can guide students in planning their learning path more strategically to maximize their future income.

Example:

Suppose a student has some knowledge of Python, Go, Java, SQL, JavaScript, and machine learning.
They are unsure which skills to learn next. Our website would allow them to:

- Enter their current skill set (e.g., Python, JS, SQL) and other information (e.g., years of experience, education level, location)
- Get an estimated salary for a junior full-stack engineer with those skills

They could also simulate "what-if" scenarios:

- What if they had 2 years of experience in TypeScript, 1 year in Kotlin, and 2 years in React? What would be their expected salary in Toronto?
- What if they had 1 year of PyTorch, 1 year of Spark, plus TensorFlow and Power BI experience? What could they earn in Vancouver?

Why do we need AI techniques?

This is essentially a salary prediction problem, which requires modeling complex relationships between skill sets, experience, location, and salary. Machine learning, especially linear regression, can learn from existing datasets to make salary predictions. Additionally, explainable AI techniques can help users understand why the model made a certain prediction, by showing similar data samples or feature contributions.

B. Problem Formulation

Objective Function:

We aim to predict a student's salary based on their skill set, experience, location, and job type using a linear regression model.
The objective is to minimize the loss function:

$$\min_{w,b} \sum_{i=1}^{n} (y_i - (w^T x_i + b))^2$$

where y_i is the actual salary, x_i is the feature vector (skills, experience, etc.), and w, b are model parameters.

Search Space:

Each student's background (skills, years of experience, education level, city, etc.) is encoded as a feature vector of dimension d.
The model will learn $d + 1$ parameters (w and b). Therefore, the search space is $\mathbb{R}^{d+1}$.

Output Space:

The output is a continuous value representing the predicted salary.

C. Programming Language and Tools

We will use **Python** as the main language.

For machine learning and data processing:

- `scikit-learn`
- `pandas`
- `numpy`
- `matplotlib`
- `seaborn`

For building the web interface:

- `Streamlit`
- `Gradio`
- `NiceGUI`

D. Evaluation Approach

We will evaluate our model using **cross-validation** by splitting the dataset into training and testing sets.
We will use **Mean Squared Error (MSE)** as our main metric to quantify how close our predictions are to the actual salary values.

Dataset:

We will use the publicly available dataset from the Stack Overflow Developer Survey:
🔗 https://survey.stackoverflow.co/

This dataset contains salary, skill set, job type, and location information. Our model will be trained and tested on this data, and we will aim to minimize the difference between predicted and actual salary values.

User Feedback:

We will also implement a feedback feature on our web interface. If a user believes the predicted salary is not accurate, they can submit a correction. This feedback will be used to evaluate and potentially retrain our model to improve performance over time.

E. Milestones and Team Responsibilities

Timeline:

Date	Milestone
2025-07-04	Data analysis, feature selection, begin model training; design frontend layout and interactions
2025-07-11	Complete model training and evaluation; draft main report structure; build frontend website
2025-07-17	Finalize report and complete website deployment
2025-07-23	Review the report; prepare final presentation slides and rehearsals

Team Responsibilities:

Task	Members	name
Data analysis	2–3 members	Chaoran, Charina, Kai(Review)
Model training	2 members	Charina,Chaoran, Archana(Review)

Task	Members	name
Report structure and review	2 members	Chaoran, Archana(Review),Kai(Revie
Final report writing	2 members	Archana,Chaoran(Review)
Presentation slides	1–2 members	Kai,Charina
Final presentation	2–3 members	Kai,Charina
Website implementation, model deployment, progress management,explainable-ai	1 member	Kai