# Chapter 1: Project Plan

## Project Title

Comparison of Ensemble Machine Learning algorithms vs Individual Machine Learning algorithms for House Price Prediction

## Research Question

How do ensemble learning methods compare to individual machine learning models in predicting house prices, and what are the effective performance benefits through ensemble algorithms?

## Objectives

* To assess the accuracy of individual and ensemble models on hour price dataset.
* Quantify the performance increase achieved using ensemble models.
* To analyse the importance of features in effecting the house prices.
* To analyse the difference in performance of the algorithms based on the change in volume of the data.
* Summarize the findings through this research and discuss about them in detail in a detailed report.

## Background and Summary

Ensemble models are a special type of machine learning models that use a multiple individual machine learning models to increase the performance. This results in capturing more patterns in the data and generalizes better than individual algorithms. There are multiple variations of these ensemble models, a few most used types are bagging and boosting. Bagging is used in Random Forest algorithm where multiple individual models (Decision Tree) are trained simultaneously and an aggregate of their predictions is used as the final output. Where are boosting trains multiple individual algorithms sequentially where each new individual algorithm reduces the error of the previous algorithm. These ensemble algorithms usually take longer to train than the individual algorithms.

Individual models are easier to understand and easier to implement than ensemble models. The understandability is one major area where the ensemble models lack. The complexity of these models works as a double-edged sword, where on one side it captures complex relationships of the data and on the other side it becomes harder and harder to interpret the result.

In this research these algorithms will be put to test on a ‘House Sales‘ data from King County, USA([link](https://www.kaggle.com/datasets/harlfoxem/housesalesprediction)). Every algorithm needs data to be trained on, this dataset is complex enough with around 20 columns which each represent a feature that helps in predicting the house price. The data also has enough observations ~20k for the model to train on. More information on the dataset can be found in the further sections.

## List of References

Ahtesham, M., Bawany, N. Z. and Fatima, K. (2020) “House price prediction using machine learning algorithm - the case of Karachi city, Pakistan,” in *2020 21st International Arab Conference on Information Technology (ACIT)*. IEEE, pp. 1–5.

https://ieeexplore.ieee.org/document/9300074

Chaurasia, A. and Haq, I. U. (2023) “Housing price prediction model using machine learning,” in *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*. IEEE, pp. 497–500.

https://ieeexplore.ieee.org/document/10303359

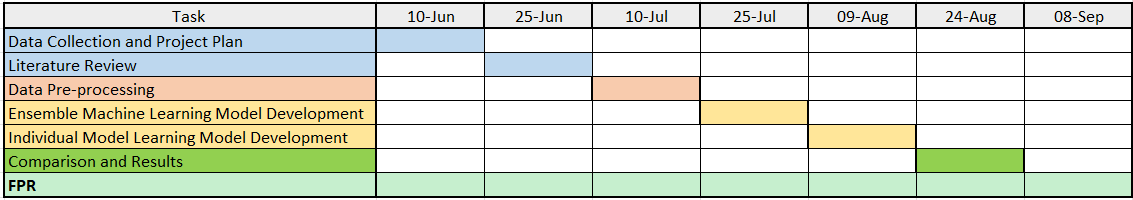
Kiran, A. *et al.* (2023) “Modeling house price prediction model using XG boost and machine learning algorithms,” in *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*. IEEE, pp. 1–9.

https://ieeexplore.ieee.org/document/10526190

The Danh Phan (2018) “Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia,” in *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*. IEEE, pp. 35–42.

<https://ieeexplore.ieee.org/document/8614000>

# Chapter 2: Task List and Project Time Line



- Data Collection and Project Plan: Conduct initial data gathering and outline project steps.

- Literature Review: Review existing research and literature relevant to the project.

- Data Pre-processing: Clean and prepare data for analysis and modelling.

- Ensemble Machine Learning Model Development: Develop and train ensemble machine learning models.

- Individual Machine Learning Model Development: Develop and train individual machine learning models.

- Comparison and Results: Compare the performance of different models and summarize findings.

- FPR (Final Project Report): Compile the final project report with all results and conclusions.

# Chapter 3: Data Management Plan

## Summary of Dataset

It consists of a wide-ranging series of real estate transactions for residential properties (~ 20k observations thereof). Using data on house sales from a specific area and point in time, and detailed data on each individual property. It contains several attributes describing the conditions under which has sold.

This database has the following main characteristics:

1. Selling Date and Price: For every record, there is a selling date and a deal price which is the primary dependent variable for all the analysis.
2. Property Details - Number of Bedrooms and Bathrooms, Size of Living Area, Size of Lot, Number of Stories, Waterfront (Y/N)
3. Additional Features: Additional important items include the condition and rating of the view, whether a basement exists and, if so, the floor size of the basement, the year the property was built and any improved or remodelling.
4. Spatial Context: This describes the street address, city, state, and postal code to which each property belongs.

A collection of data for a variety of cities in the state of Washington, USA the data contains the following attributes City related: Shoreline Seattle Kent Bellevue Redmond Year renovation data allows for a comparison of price per square foot over time for houses in the dataset.

## Document control

GitHub Repository: https://github.com/sp22adn/datascience-project.git

For Version control and Document control a combination of GitHub repository and Git.

## Ethical requirements

* Does the data meet GDPR requirements?
  + Yes, the data is collected is only necessary data, that has been anonymized, and as there is no personal information regarding an individual no consent information is mentioned in the source for the dataset.
* Does the project conform to UH ethical policies?
  + Yes, the data is taken from an open-source data repository ‘Kaggle’ and it is not collected directly from people for the sake of this project.
* Do you have permission to use the data for your proposed research project?
  + Yes, the data is free to use by anyone according to the **CC0 1.0 Universal license** mentioned in the website.
* Are you assured that the data was collected ethical (i.e. by the original people who gathered/collected/ collated/made the data)?
  + As the data comes from a government organization (King County) it is assumed that the data is ethically collected, though this can’t be verified. It is also important to note that there are no notices or legal issues that can be found online relating to the ethical implications for collection of this dataset.

# References:

*Deep learning vs. machine learning* (2020) *Zendesk*. Available at: https://www.zendesk.com/in/blog/machine-learning-and-deep-learning/ (Accessed: June 8, 2024).

Goyal, S. (2021) *Evaluation metrics for regression models - analytics Vidhya - medium*, *Analytics Vidhya*. Available at: https://medium.com/analytics-vidhya/evaluation-metrics-for-regression-models-c91c65d73af (Accessed: June 8, 2024).

Ray, S. (2017) *Top 10 machine learning algorithms to use in 2024*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/ (Accessed: June 8, 2024).

Zhang, Z. (2019) *Understand data normalization in machine learning*, *Towards Data Science*. Available at: https://towardsdatascience.com/understand-data-normalization-in-machine-learning-8ff3062101f0 (Accessed: June 8, 2024).