Stats 400: Final Project Proposal:

**Dataset:**

Name: Employee

Dimensions: 10

Cases: 15000

Dimensions/columns planning on using: Satisfaction level, Number of projects, last evaluation, Average monthly hours, time spent company, work accidents, departments, salary, left, and promotion last 5 years. These dimensions are subjected to change. We might reduce the dimensions, based on the topic decided.

**About the Dataset:**

The Employee Satisfaction Survey dataset is a comprehensive collection of information regarding employees within a company. It includes essential details such as employee identification numbers, self-reported satisfaction levels, performance evaluations, project involvement, work hours, tenure with the company, work accidents, promotions received in the last 5 years, departmental affiliations, and salary levels. This dataset offers valuable insights into the factors influencing employee satisfaction and can be used to analyze and understand various aspects of the workplace environment.

**Description of each dimension/column**:

Satisfaction level: This is a list of scores, that indicate the level of satisfaction each employee has in the company that is working for different departments.

Number of projects: This is the total number of projects that employees have done for the company.

Average monthly hours: Total hours spent working on the projects for each employee in different departments. This could be within the company or outside of the company.

Time spent at the company: Number of hours spent by the employees within the company.

Work accident: Total number of mistakes that each employee has made during their work hours. This could be the mistakes they might have made on their projects.

Left: Total number of employees that have quit the job from different departments.

Promotion last 5years: This is the number of promotions that employees have gotten in each department for the past 5 years.

Departments: The different subset of people that are working for the company, mainly focusing on different attributes of work.

Salary: The total amount of income that the employee is making on average for different departments.

**Topics:**

We have chosen to analyze the dataset called “Employee”. Some of the data mining and machine learning techniques that we would like to use for the project would be regression, clustering, Decision Trees, or predictive modeling. For example, if we chose to go with regression, these are the following topics that would like to analyze: 1. Predicting satisfaction level based on factors like: Number of projects, evaluation scores, average monthly hours, time spent in the company. The other topic would be predicting the impact of work hours on performance. In this topic we would like to use the following explanatory variables: Average monthly hours, time spent in the company.

If we decided to go with clustering, then these are the following topics that we would like to explore more on: Analyzing work accidence patterns. We will be analyzing the pattern based of the following predictors: Work hours, satisfaction level, and maybe see other factors that could be impactful for the clustering pattern. The other one would be department and salary clustering. For this we would like to explore how different departments and salary levels cluster together based on factors like average hours, and number of projects.

For decision tree, we were thinking about the following topics: Predicting employee departure: We use a decision tree to classify employees based on their chances to leave the company. Some of the factors used to predict this would be Number of projects, satisfaction level, last evaluation, and another predicting variable if we find any.

The other topic we had in mind was to classify employee by their salary. Where we classify them into different categories of salary (low, medium, and high). Factors included would be department, number of projects, average monthly hours, and maybe hours spent in the company.

For predictive modeling (less likely to go with this idea), we would like to analyze the following topics: Forecasting promotion chances for each department, based on their current position, and their contributions (like the number of projects they have done, and how many promotions each department has gotten in the past 5 years). The other one would be predictive analysis of work accident risk. We would base this on the department, salary, and number of hours.

**Research Question:**

* What are some of the factors that affect employee satisfaction?
* Why does one department (Sales) in the whole company have the highest number of employees leaving? What would be some factors contributing to it?