# Project Proposal

**1.Team members.**

1) Parikshit Joshi (210895H) - Leader
2) Jia Jun (210897S) - Group Member

**2.Introduction/Background information.**

A startup is a company typically in the early stages of its development. These entrepreneurial ventures are typically started by 1-3 founders who focus on capitalizing upon a perceived market demand by developing a viable product, service, or platform. Startups face high uncertainty and have high rates of failure, but a minority of them do go on to be successful and influential. Startups play a major role in economic growth. They bring new ideas, spur innovation, create employment thereby moving the economy. There has been an exponential growth in startups over the past few years. Predicting the success of a startup allows investors to find companies that have the potential for rapid growth, thereby allowing them to be one step ahead of the competition.

Seed capital can be used by startups to fund research and the development of their business strategies. A Large portion of a company's seed funding may originate from sources close to its creators, such as family, friends, and other contacts with a high risk of failure in the start-up.

Machine learning uses algorithms to create models that reveal patterns from data, allowing businesses to gain understanding and make predictions to enhance their business model. We can use the help of machine learning to predict the outcomes of a startup whether it will be successful or not. There are several algorithms to choose from and our goal is to make the most accurate result possible, which is called predictive modeling. We will be using Random Forest, Decision Tree, Support-Vector Machine (SVM),

Logistic Regression to provide our forecast result on whether a startup will turn into a success or a failure.

**4. Business objectives identified.**

The business objective is to identify success startups with success criteria of us being able to identify by 25%.

**5. Main and sub tasks identified.**

Collection of datasets using techniques that comply with data protection and privacy ethics. - By using knime, remove/encrypt any sensitive information like name, age.

Use visualizations to get better insights about data

1) Check for imbalanced data to see if the dataset is in the 80 and 20 range by using bar graphs
2) Check for missing data - see for any null values - location
3) Identify trends by using scatter plot - eg. higher rounds of valuation result in higher success rate .
4) Using correlation matrix to check for multicollinearity like zip code, latitude and longitude
5) Remove any outliers using RapidMiner
6) See which factors link to high success and which factors lead to lower success using a Sankey diagram
7) Find differences between success and failed business by analyzing the features.
8) Transforming data to make the model understand data better - eg latitude and longitude categorized to state

Cleaning and modifying data & prepare data for modeling

1. Find out the possible causes of missing data - why is the location missing
2. Figure out the possible approaches and the best approach to fix the missing data - Use RapidMiner to input the data using k-nearest neighbor

Build multiple models using various machine learning algorithm
In the decision tree, we make many models by choosing a different column at every iteration for e.g. bootstrap.

1.Use predictive modeling techniques by using misclassification, AOC, ROC on the processed data to compare models
2. Use unsupervised learning method - k means clustering, to understand the data better
3. Assess the performance of trained models - Accuracy, misclassification rate
4. Tune the parameters to achieve the best possible outcome - testing by increasing more features, does the model get better
    Increase/ Decrease Penalty value for better model
    Kernel function change from linear to quadratic
    Change the number of trees and also the bootstrap value

Feature selection

1. Use feature importance analysis to see which feature affects the results the most and which feature make the model worse
2. By using visualizations -parallel coordinates technique, it will give a deeper look and impacts of the features across the columns.

**6.Delegation of tasks/responsibility.**

Collection of datasets:
JiaJun -kaggle datasets
Parikshit - UCL,google,universities

Make visualizations:

Jia Jun

Use bar chart to find out number of null values and input them
For time series data, change the values into dd/mm/yyyy
If data is unbalanced, use RapidMiner to balance the dataset

Using knime to analyze and input the null values. Modify the time series data so that we can use it in the modeling. To balance the dataset, we can use RapidMiner to balance our dataset by having the same number of records from both success and failure. Some columns are unnecessary or irrelevant to our desired result or deciding factor which we have to remove.

When doing the train-test split, it is important to make sure that the distribution of the data between the training set and testing set is similar: The range distribution should be from 80:20 ratio. Must test out different number of trees and Decrease penalty value for better model.

Parikshit

1) Remove any columns that are not useful or empty
2) Build graphs to analyze feature selection
3) using correlation matrix to check for multicollinearity

It is necessary to do some data cleaning such as from the data before visualizing it. Use feature selection to remove those features where the p value is more than 0.05, and which columns are not improving the model. There are columns such as zip code,

latitude and longitude which have high multicollinearity. Do data transformation if needed such as converting and categorising the latitude and longitude into city. For modelling. Changing the Kernel function from linear to quadratic to see the model get better.

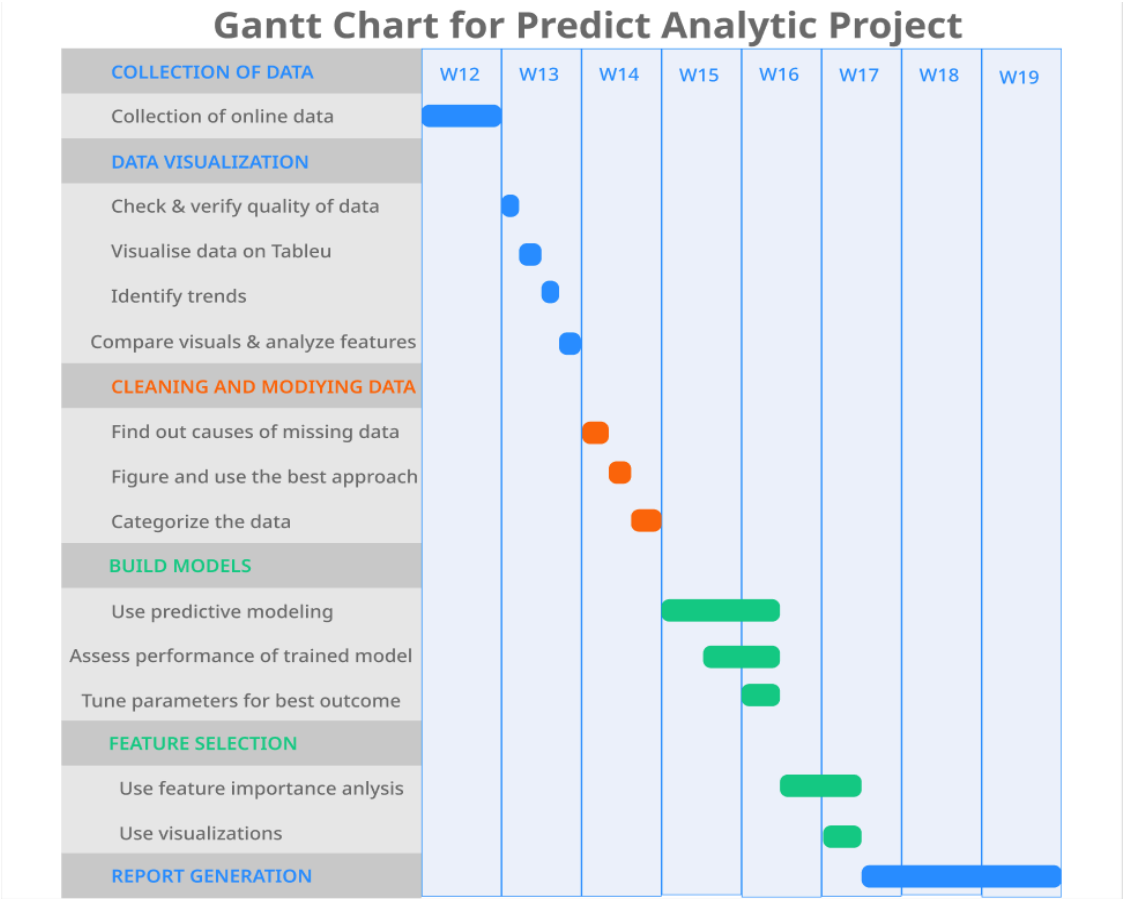
Scoring metrics:

1) Accuracy
2) Precision
3) Recall

Training Model using SAS Viya:

1) SVM,Decision Tree - Parikshit
2) Logistic regression, Random Forest - JiaJun

## 7.Project schedule and task allocation (Gantt chart)



1) **Data Visualisation - Both**

Jia Jun

Use bar chart to find out number of null values

Scatter plot which shows the relationship between the amount spent on acquiring a company to the total funding a company received.

Use map to show which state has the most success and failures using tableau

Use

Parikshit

sankey diagram to show the correlation between the number of company relationships to its success/failure.

Using correlation matrix to check for multicollinearity

2) **Cleaning and Modifying Data**

Jia Jun

For time series data, change the values into dd/mm/yyyy and rounding off time data into whole numbers like 6.2 to 6.

Remove any columns that are not useful or empty

If data is unbalanced, use RapidMiner to balance the dataset

Parikshit

Input values using RapidMiner using k-nearest neighbour

Build graphs to analyze feature selection - which feature is important and needed and which feature does not affect the model by looking at the p-value

3) **Build Models**

Jia Jun

Must test out different number of trees and Decrease penalty value for better model.

Parikshit

Changing the Kernel function from linear to quadratic to see if the model gets better. Check the lift and misclassification rate to see any improvement

Compare models between Different classification methods