# LINK PREDICTION ON SOCIAL MEDIA

**MAJOR PROJECT SYNOPSIS**

***Of***

**BACHELOR OF TECHNOLOGY**

***In***

**COMPUTER SCIENCE & ENGINEERING**

***By***


| Himani Sheoran | Udit Jain | Umang Tiwari |
|---|---|---|
| 04514802718 | 10214802718 | 10314802718 |


**Guided by**

**Ms. Savita Sharma**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY**
**(AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY, DELHI)**


**(Session 2021-2022)**

# INTRODUCTION

Currently with the rapid development, online social network has been a part of people's life. A lot of sociology, biology, and information systems can use the network to describe, in which nodes represent individual and edges represent the relationships between individuals or the interaction between individuals. Therefore, the study of complex networks has been the important branch of many scientific fields. Link prediction is an important task in link mining. Link prediction is to predict whether there will be links between two nodes based on the attribute information and the observed existing link information. Link prediction not only can be used in the field of social network but can also be applied in other fields. As in bioinformatics, link prediction can be used to discover interactions between proteins; in the field of electronic commerce, link prediction can be used to create the recommendation system; and in the security field, link prediction can help to find the hidden terrorist criminal gangs. Link prediction is closely related to many areas. Therefore, in recent years there is a lot of correlation algorithms proposed to solve the problem of link prediction.

Social networks are a popular way to interpret the interaction among the people. They can be Visualized as graphs, where a vertex corresponds to a person and edges represent the connection between them. Understanding the dynamics that drive the evolution of social networks is a complex problem due to a large number of variable parameters. But, a comparatively easier problem is to understand the association between two specific nodes.

For the given source node and destination node we have to predict whether there is any probability of connecting between them.

Then the question is what will be target data (Labelled data) and what the training data are.

Because in Machine Learning we need training data(X) and target data(y). And here we have Only two columns i.e. source_node and destination node. To handle this problem we need to create new features. So that it will be helpful for the model. Also we need to create a label (0/1) attribute as the target column, 0 = Not connected and 1 = Connected. When all these are done like adding features and labels then we can say that it is converted to a Machine Learning problem.

## TECHNOLOGY USED

### Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding; make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance.

### Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include:

data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.
JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. JupyterLab is extensible and modular: write plugins that add new components and integrate with existing ones.

### Python PIP

Pip is a package manager for Python. That means it's a tool that allows you to install and manage additional libraries and dependencies that are not distributed as part of the standard library.
Package management is so important that pip has been included with the Python installer since versions 3.4 for Python 3 and 2.7.9 for Python 2, and it's used by many Python projects, which makes it an essential tool for every Pythonista.
The concept of a package manager might be familiar to you if you are coming from other languages. JavaScript uses npm for package management; Ruby uses gem, and .NET use NuGet. In Python, pip has become the standard package manager.

## MODELS AND ALGORITHMS USED

### ENSEMBLE MODEL

An ensemble is a machine learning model that combines the predictions from two or more models. The models that contribute to the ensemble, referred to as ensemble members, may be the same type or different types and may or may not be trained on the same training data.

## THE FIELD OF THE PROJECT

### Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.
The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.
But, using the classic algorithms of machine learning, text is considered as a sequence of keywords; instead, an approach based on semantic analysis mimics the human ability to understand the meaning of a text.

# OBJECTIVE

- An introduction to link prediction, how it works, and where you can use it in the real-world
- Learn about the importance of Link Prediction on social media
- Build your first Link Prediction model for a Facebook use case using Python

# FEASIBILITY STUDY

## 1. FINANCIAL FEASIBILTY

Since the system doesn't consist of any multimedia data transfer, bandwidth required for the operation of the application is very low.

## 2. TECHNICAL FEASIBILTY

Project is a complete web based application and main technology and tools associated are:

- Python
- Machine Learning
- Python PIP

Each of the technology are freely available and the technical skills required and manageable. Time limitations of the product development and the ease of the implementing using these technologies are synchronized.

## 3. RESOURCES AND TIME FEASIBILITY

Resources that is required for the project includes:

- Programming device (personal computer)
- Hosting space (freely available)
- Programming tools ( freely available)
- Programming individuals

This states that the project has the required resource feasibility.

## 4. SOCIAL/LEGAL FEASIBILTY

This project is designed for social media organizations and themed networking software's so it will be accessed by authorized social media websites firm's workers. This is not an open source project.

## NEED AND SIGNIFICANCE OF PROJECT

The benefits of social network analytics can be highly rewarding. Here are a few key benefits:

- Helps you understand your audience better
- Used for customer segmentation
- Used to design Recommendation Systems
- Detect fake news, among other things

# METHODOLOGY/ PLANNING OF WORK

## RESEARCH GAP IDENTIFIED

- Most of the work on link prediction is involves unsupervised learning. However later it was found that supervised technique produces a better accuracy for link prediction.
- While this is not the first time supervised technique is being used, many factors guiding classification remain unexplored.
- We intend to achieve a higher testing and training accuracy through our work.

## STEPS TO BE FOLLOWED

1. Choosing the right dataset.
2. Preparing the dataset.
3. Feature engineering
4. Splitting and training the data.
5. Applying the machine learning model
6. Scrutinizing the result.
7. Showing the result through various graphs.

# BIBLIOGRAPHY


- Liyan Dong,1,2 Yongli Li,3 Han Yin,1,2 Huang Le,1,2 and Mao Rui1, The Algorithm of Link Prediction on Social Network, Volume 2013 |Article ID 125123

- Mahdi Jalili, Yasin Orouskhani, Milad Asgari, Nazanin Alipourfard and Matjaž Perc, Link prediction in multiplex online social networks, Published:01 February 2017 https://doi.org/10.1098/rsos.160863

- Journal of the American Society for Information Science and Technology,The link-prediction problem for social networks,David Liben-Nowell,Jon Kleinberg, First published: 26 March 2007, https://doi.org/10.1002/asi.20591

- New perspectives and methods in link prediction,Ryan N. Lichtenwalter, Jake T. Lussier, KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data miningJuly 2010 Pages 243-252 https://doi.org/10.1145/1835804.1835837

- Link Prediction using Supervised Learning ∗ Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki Rensselaer Polytechnic Institute, Troy, New York 12180 {alhasan, chaojv, salems, zaki}@cs.rpi.edu