FAKE ACCOUNT DETECTION

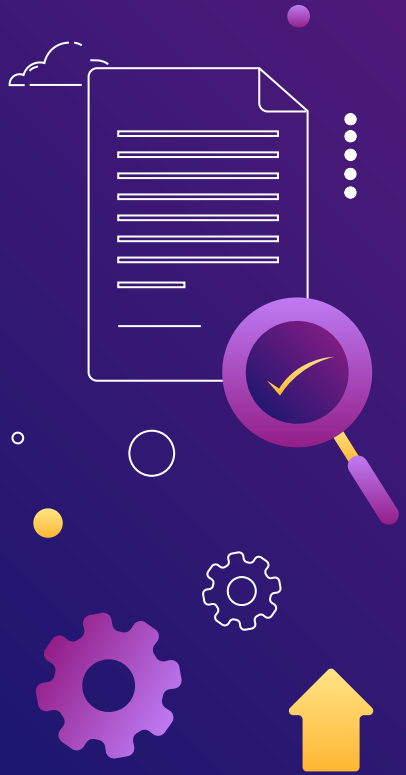
A study using machine learning algorithm

CONTENTS

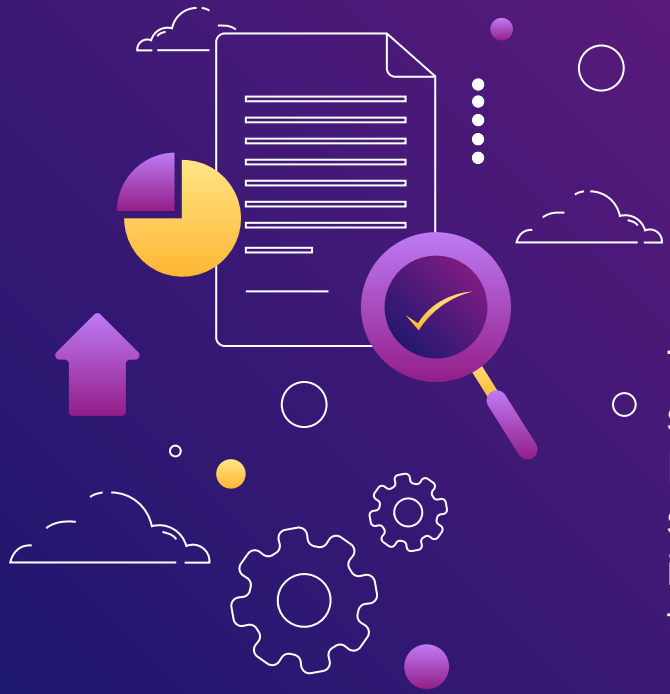
TEAM MEMBERS

S5 CSE - A

- Adil Muhammed Ashraf P. K - AM.EN.U4CSE19003
- Praveena K - AM.EN.U4CSE19043
- Sayooj B Kumar - AM.EN.U4SCE19050
- Taniya Rajesh - AM.EN.U4CSE19057



PROBLEM

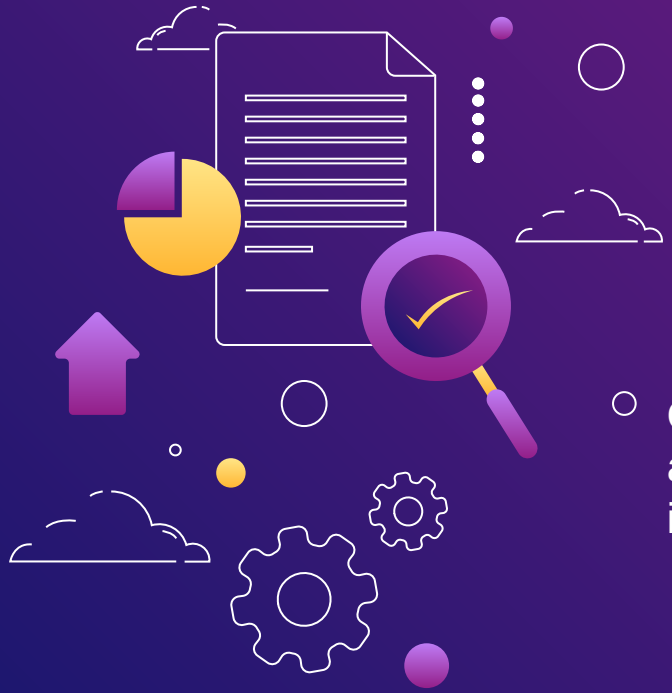


INTRODUCTION

There has been a rise of fake engagements which spread false information which affects users. Detection of these kinds of malicious account is significant, so as to avoid loss of sensitive information, unnecessary trolls and political stir-ups, fraudulent purposes.

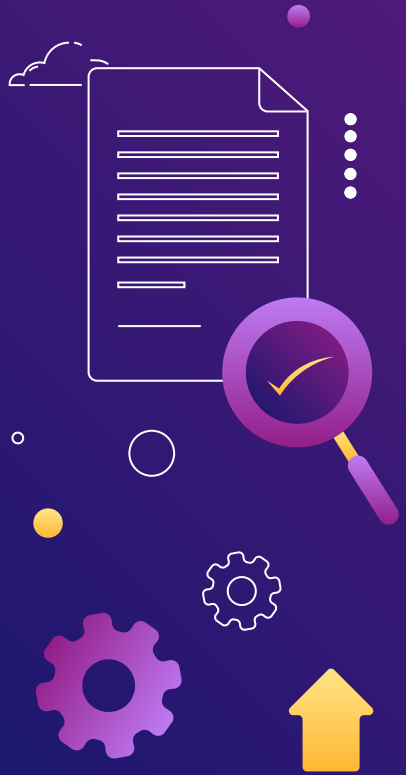
950,000,000

Instagram has an estimated 95 million fake accounts

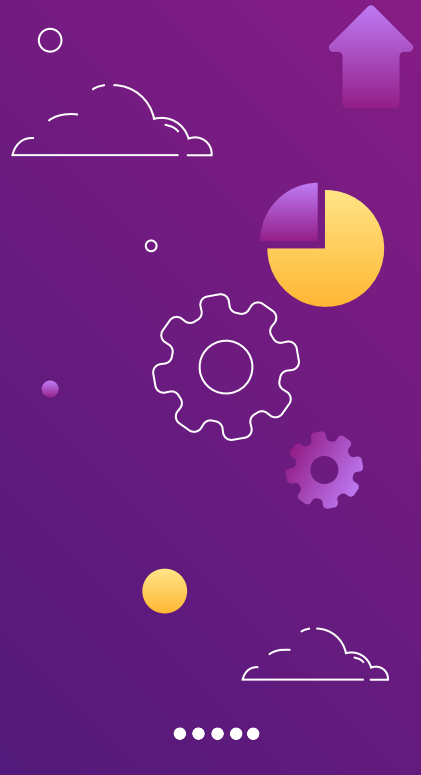


AIM

Our aim is to create a model using machine learning algorithms, which detects the fake accounts present in the Instagram



PACKAGES USED



MODEL

- Numpy : python library for working with arrays.
- Pandas : python library for dataframe manipulation.
- Matplotlib : library for creating interesting visualization.
- Seaborn : library built on top of matplotlib, high-level interface with attractive designs. Used for heatmap.
- Sklearn : library used for predictive data analysis.

DATA COLLECTION

- Request: Python library which send automated http requests

LIVE PREDICTION

- Flask : Flask is a micro web framework written in Python.

SOLUTION TO PROBLEM

01 COLLECTION OF DATA

Collect necessary data needed for models

02 MODELS

Train different models and predict using them

03 ACCURACY

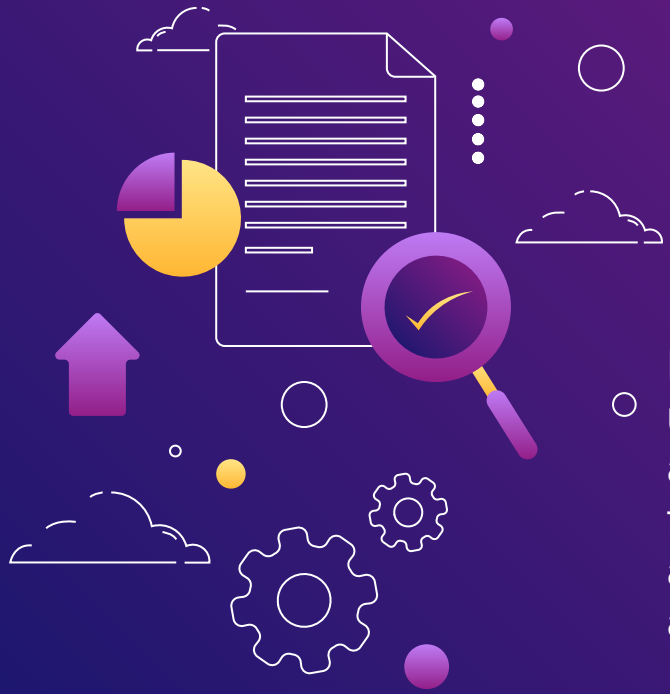
Testing accuracy of the models

04 LIVE PREDICTION

After analyzing, created a live web application which predicts accounts

01

COLLECTION OF DATA



DATA COLLECTION

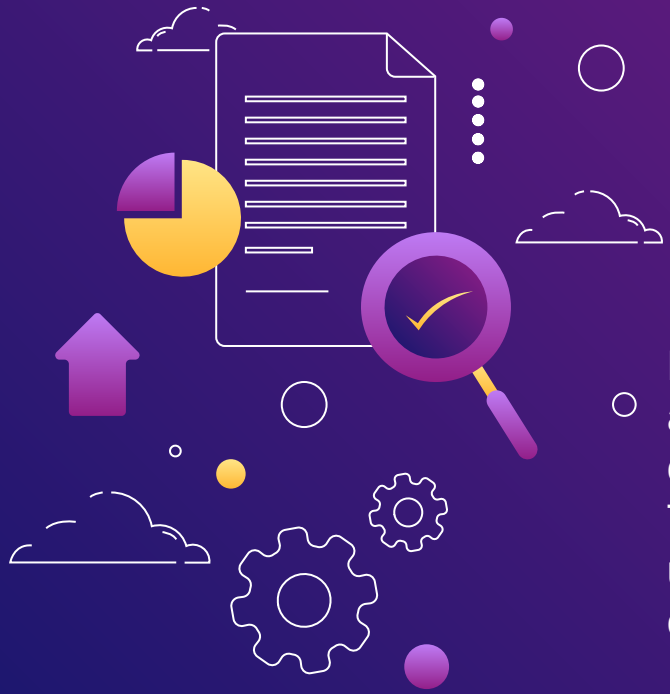
For the dataset, extracted accounts from Instagram using web scraper, which has details on whether the account has profile picture, number of followers and following, if private account enabled, professional account or not, number of posts shared, likes, tags and comments.

Data Collector

DATA FORMATTING

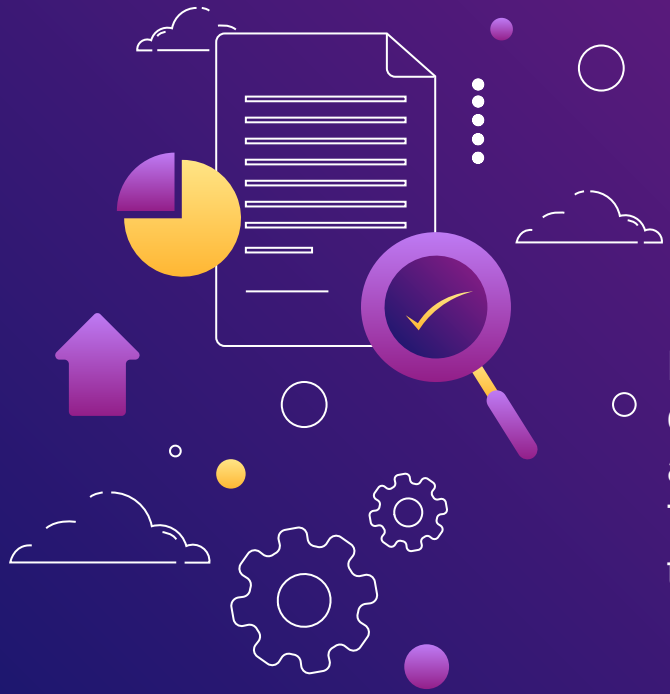
- Converted true to 1 and false to 0
- Null values to 0
- String to length of strings

02 ALGORITHM/MODEL USED



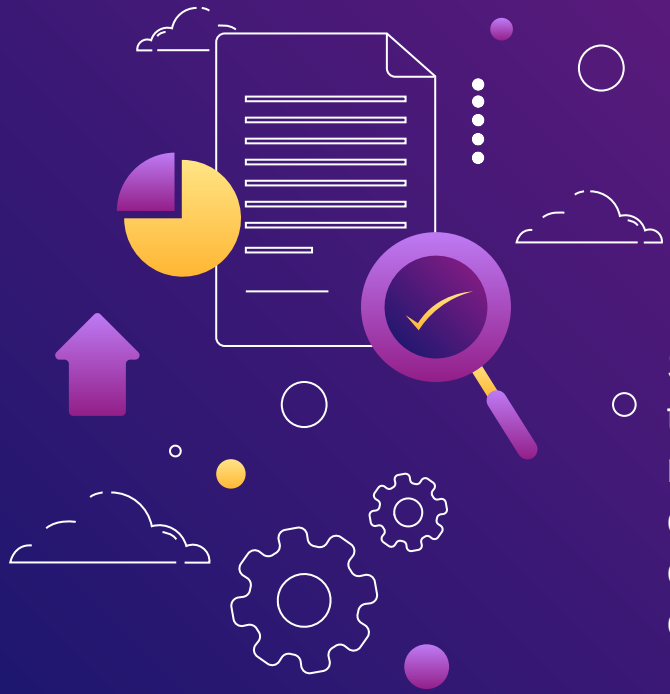
KNN

K-nearest neighbour is a supervised ML algorithm.The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.



LOGISTIC REGRESSION

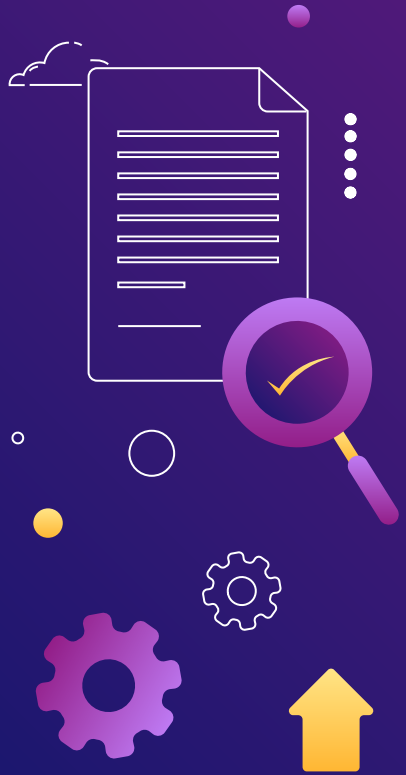
Logistic Regression is commonly used for classification problems. It is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1 .



SVM

Support Vector Machine is a supervised algorithm that can be used for both classification and regression. Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line).

03 ACCURACY OF MODEL



80-85%

04 LIVE PREDICTION SYSTEM

WORKING

TRAIN MODEL

Used previously collected data to train model

USER INPUT

Accepts input from user and fit into model and finally predict them

OUTPUT

Predict is account is fake or not and then predict the output

WEB APP

This Application is developed using flask framework where application accept input from user and collect data then predicts automatically

LIVE PREDICTION

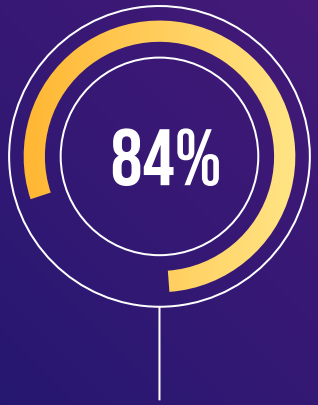
Access the Hosted Application
Source Code

INDIVIDUAL REPORTS

DATASET FEATURES

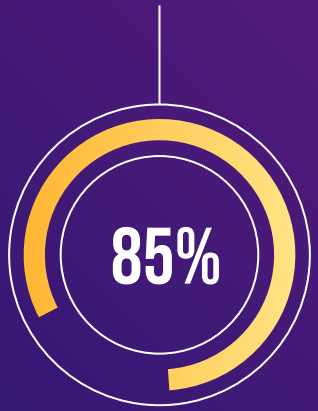
- is_verified
- has_anonymous_profile_picture
- biography_len
- external_url
- followers
- following
- has_clips
- media_count_of_data_collected
- highlight_count
- is_professional_account
- is_private
- media_count
- total_tags
- total_likes
- total_comments
- name_length
- average_likes

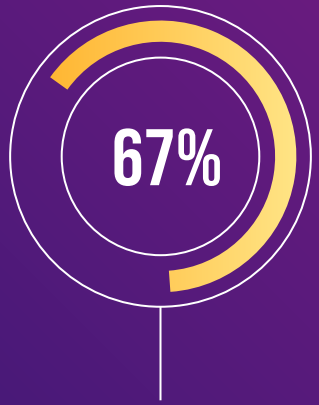
SAYOOJ



KNN

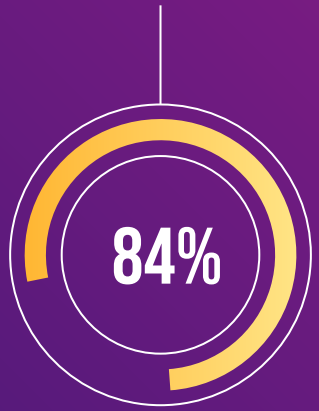
Logistic Regression





SVM

Logistic regression from scratch



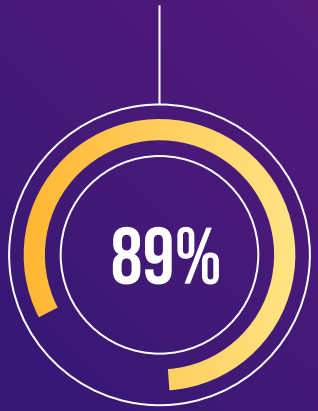
DATASET FEATURES

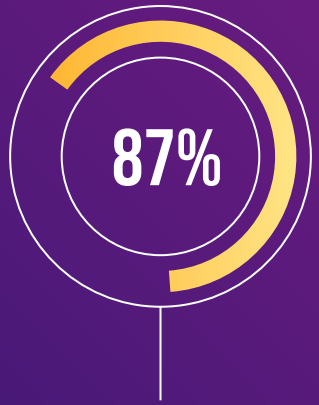
- is_verified
- has_anonymous_profile_picture
- biography_len
- followers
- following
- has_clips
- highlight_count
- is_professional_account
- is_private
- media_count
- total_tags
- total_likes
- total_comments
- is_fake

TANIYA

KNN

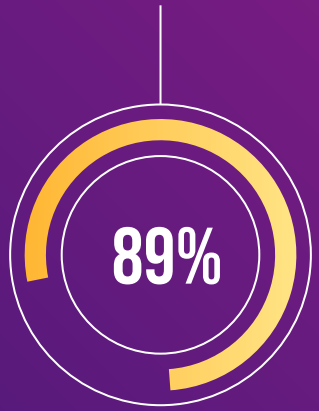
Logistic Regression





SVM

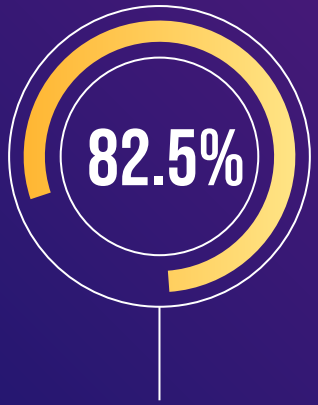
KNN from Scratch



DATASET FEATURES

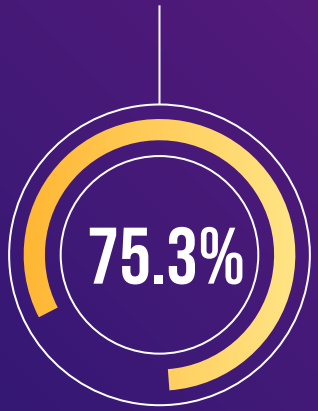
- username
- full_name
- is_verified
- has_anonymous_profile_picture
- biography_len
- external_url
- followers
- following
- has_clips
- media_count_of_data_collected
- highlight_count
- is_professional_account
- is_private
- media_count
- total_tags
- total_likes
- total_comments
- name_length

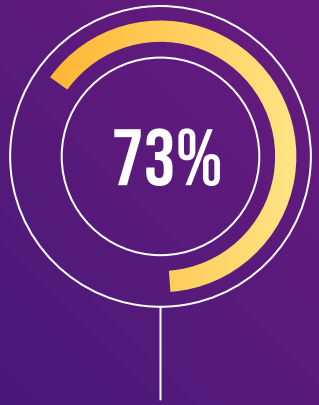
ADIL



KNN

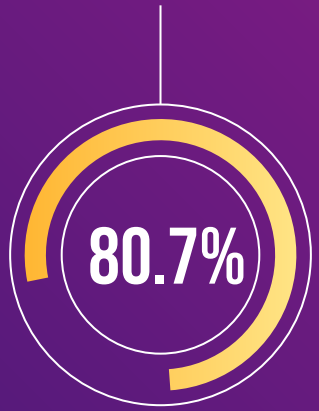
SVM





Logistic Regression

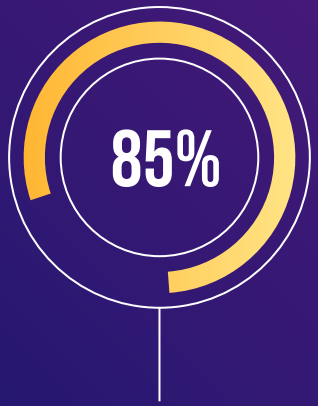
KNN
from scratch



DATASET FEATURES

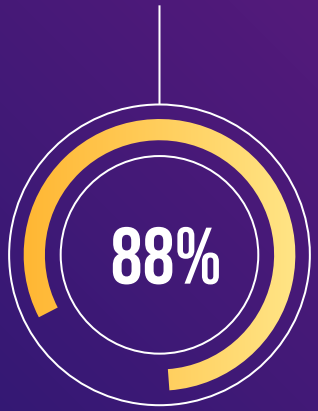
- len_username
- Len_fullname
- Is_verified
- has_anonymous_profile_picture
- biography_len
- followers
- following
- is_professional_account
- is_private
- media_count
- is_fake

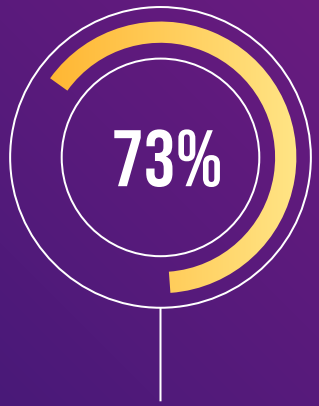
PRAVEENA



KNN

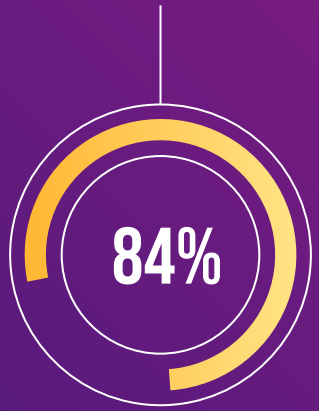
Logistic Regression





SVM

Logistic regression from scratch



THANK YOU!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**