DATA SCI-BOT

A FINAL YEAR CAPSTONE DESIGN PROJECT

Submitted by

G. LOK SUNDAR (9917004144)

BABLOO KUMAR(9917004218)

N.YASWANTHI (9917004085)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

COMPUTER SCIENCE AND ENGINEERING

KALASALINGAM ACADEMY OF RESEARCH

AND EDUCATION

KRISHNANKOIL 626 126

Academic Year 2019-2020

DECLARATION

We affirm that the project work titled **" DATA SCI-BOT "** being submitted in partial fulfillment for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** is the original work carried out by us. It has not formed the part of any other project work submitted for award of any degree or diploma, either in this or any other University.

G.LOK SUNDAR

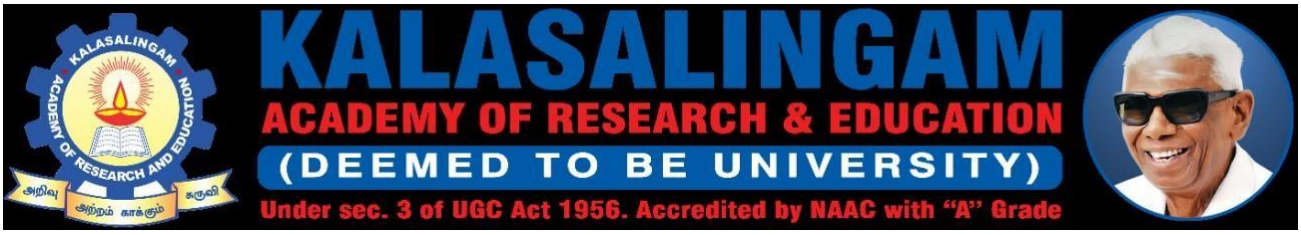
9917004144

BABLOO KUMAR

9917004218

N.YASWANTHI

9917004085



BONAFIDE CERTIFICATE

BONAFIDE CERTIFICATE

Certified that this project report **" DATA SCI-BOT "** is the bonafide work of " **G.LokSundar, Babloo Kumar, N.Yaswanthi "** who carried out the project work under my supervision.

SUPERVISOR

Dr.P. Deepa Lakshmi

Dean of School of Computing

Computer Science and Engineering

Kalasalingam Academy of Research Education
Krishnankoil 626126

HEAD OF THE DEPARTMENT

Dr. A. FRANCIS SAVIOUR DEVARAJ

Professor/Head

Computer Science and Engineering

Kalasalingam Academy of Research and and and Education
Krishnankoil 626126

Submitted for the Project Viva-voce examination held on.....

Internal Examiner **External Examiner**

ACKNOWLEDGEMENT

First and foremost, I wish to thank the **Almighty God** for his grace and benediction to complete this Project work successfully. I would like to convey my special thanks from the bottom of my heart to my dear **Parents** and affectionate **Family members** for their honest support for the completion of this Project work.

I express deep sense of gratitude to "Kalvivallal" Thiru. **T. Kalasalingam**B.com., Founder Chairman, "Ilayavallal" **Dr.K.Sridharan**Ph.D., Chancellor, **Dr.S.Shasi Anand**, Ph.D., Vice President (Academic), **Mr.S.ArjunKalasalingam** M.S., Vice President (Administration) , **Dr.R.Nagaraj**, Vice-Chancellor, **Dr.V.Vasudevan**Ph.D., Registrar , **Dr.P.Deepalakshmi**M.E., Ph.D., Dean (School of Computing) . And also a special thanks to **Dr.A. Francis SaviourDevaraj.**Head, Department of CSE, Kalasalingam Academy of Research and Education for granting the permission and providing necessary facilities to carry out Projectwork.

I would like to express my special appreciation and profound thanks to my enthusiastic Project Supervisor **Dr.P.Deepalakshmi**M.E., Ph.D., Dean (School of Computing) of Kalasalingam Academy of Research and Education [KARE] for her inspiring guidance, constant encouragement with my work during all stages. I am extremely glad that I had a chance to do my Project under my Guide, who truly practices and appreciates deep thinking. I will be forever indebted to my Guide for all the time he has spent with me in discussions. And during the most difficult times when writing this report, he gave me the moral support and the freedom I needed to moveon.

Besides my Project guide, I would like to thank the rest of Class committee members and all faculty members and Non-Teaching staff for their insightful comments and encouragement. Finally, but by no means least, thanks go toall my school and college teachers, well wishers, friends for almost unbelievable support.

KALASALINGAM
ACADEMY OF RESEARCH & EDUCATION
(DEEMED TO BE UNIVERSITY)
Under sec. 3 of UGC Act 1956. Accredited by NAAC with "A" Grade

School of Computing

Department of Computer Science and Engineering

Project Summary

Project Title	DATA SCI-BOT
Project Team Members (Name with Register No)	G.LOK SUNDAR (9917004144) BABLOO KUMAR (9519004218) N.YASWANTHI (9917004085)
Guide Name/Designation	**Dr. P. DEEPA LAKSHMI** Dean of School of Computing
Program Concentration Area	Social, Ethical, Economic
Technical Requirements	Hardware Requirements: 1. Processor Type : Pentium i3 2. Speed : 3.40GHZ 3. RAM : 4GB DD2 RAM 4. Hard disk : 500 GB Software Requirements: 1. Windows Operating System. 2. Anaconda Software 3. Jupyter Notebook 4. Google API

Engineering standards and realistic constraints in these areas.

Area	**Codes & Standards / Realistic Constraints**	**Tick ✓**
Social	This project uses open source software including Java programming. The core part is that the project finds its The focus of our study is essentially warranted because it can be potentially used to nominate intervention targets and help to match students with the prevention and intervention strategies most likely to	✓

	work for them. Hence the project is social.	
Economic	This project helps the organizations and companies to increase their economy, efficiency and productivity.	✓
Ethical	This project is aimed at many ultimately, the goal of grade prediction in similar experiments is to use the constructed models for the design of intervention strategies aimed at helping students at risk of academic failures.	✓

Realistic constraints:

Social :

Economical:

Ethical :

Engineering Standards:

ABSTRACT

In this project, Our goal is to build an AI-based Sci-bot which will do the whole work of a Data Scientist on its own and it will tell the Final Data Analysis Observation to the Stakeholder/client.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full form
ML	Machine Learning
NN	Neural Networks
AI	Artificial Intelligence
API	Application Programming Interface
DT	Decision Tree
XGB	eXtreme Gradient Boosting
RF	Random Forest

CHAPTER I

INTRODUCTION

1.1 Overview

Presently The automation is Limited to Machine Learning it didn't come to Data Science still Researches are going on to automate the Data Scientist Process. At present all the Data analysis part is done by human Physically, and the time required for the project also up to one month for each case study.

1.2 Proposed Approach

The Main Goal is to create a Bot that can use the data and Prepare an Analysis story on it as Data Scientists do. In this project we are particularly doing on Case study on Employee attrition. In this our Sci-bot takes the data in and it does all the Data analysis part by itself and finally it creates a story on its self and explains it to the stakeholders/clients. In the AI algorithm the Sci-bot takes the best algorithm for the problem using Auto-Machine Learning Methods. It will automatically take the best parameters for each algorithm using a grid search and Bayesian optimization Technique. Finally to the client everything is internal but the final output is a story that the client will understand.

CHAPTER II

LITERATURE REVIEW

SL.NO	Title of Paper	Authors	Journal	Description
1.	Learning Feature Engineering for Classification	Fatemeh Nargesian, Horst Samulowitz, Udayan Khurana Elias B. Khalil, Deepak Turaga	Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)	In this paper, they discussed the automated feature engineering techniques. Existing approaches to automating this process rely on either transformed feature space exploration through evaluation-guided search or explicit expansion of datasets with all transformed features followed by feature selection. Such approaches incur high computational costs in runtime and/or memory. For this, they presented a novel technique, called Learning Feature Engineering (LFE).

SL.NO	Title of the paper	Authors	Journal	Description
2	Bayesian Optimization with Unknown Search Space	Huong Ha, Santu Rana, Sunil Gupta, Thanh Nguyen, Hung Tran-The, Svetha Venkatesh	33rd Conference on Neural Information Processing Systems (NeurIPS 2019)	In this paper, they Implemented the new way of optimizing the features space which is unknown to the best fit data can do. To address this problem, they propose a systematic volume expansion strategy for Bayesian optimization. They devise a strategy to guarantee that in iterative expansions of the search space, our method can find a point whose function value within Epsilon of the objective function maximum. Without the need to specify any parameters, our algorithm automatically triggers a minimal expansion required iteratively.

CHAPTER III

PROBLEM DEFINITION

In this our Sci-bot takes the data in and it does all the Data analysis part by itself and finally it creates a story on its self and explains it to the stakeholders/clients. In the AI algorithm the Sci-bot takes the best algorithm for the problem using Auto-Machine Learning Methods. It will automatically take the best parameters for each algorithm using a grid search and Bayesian optimization Technique. Finally to the client everything is internal but the final output is a story that the client will understand.

CHAPTER IV

REQUIREMENTS

4.1 FUNCTIONAL REQUIREMENTS

Data collection

The data collection process involves the selection of quality data for analysis. Here we used dataset from website dataset taken from kaggle for machine learning implementation. The job of a data analyst is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques.

Data visualization

A large amount of information represented in graphic form is easier to understand and analyze. Some companies specify that a data analyst must know how to create slides, diagrams, charts, and templates. In our approach, the histogram plot and features extraction as shown as visualization.

Data preprocessing

The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

Dataset splitting

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

Training set. A data scientist uses a training set to train a model and define its optimal parameters it has to learn from data.

Test set. A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model's ability to identify patterns in new unseen data after having been trained over a training data. It's crucial to use different subsets for training and testing to avoid model overfitting, which is the incapacity for generalization we mentioned above.

Model training

After a data scientist has preprocessed the collected data and split it into train and test can proceed with a model training. This process entails "feeding" the algorithm with training data. An algorithm will process

data and output a model that is able to find a target value (attribute) in new data an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

Model evaluation and testing

The goal of this step is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That's the optimization of model parameters to achieve an algorithm's best performance.

4.2 SOFTWARE REQUIREMENTS :

Windows 10 or above with more than 2.0 clock speed.

Python 3.0 or above

Jupyter notebook in Anaconda platforms

Voice software for telling stories.

4.3 HARDWARE REQUIREMENTS :

Processor	:	Intel(R) Core(TM) i3
Processor Speed	:	3.06 GHz
Ram	:	4 GB
Hard Disk Drive	:	250 GB
Monitor	:	"15.6" inches

CHAPTER V

SYSTEM DESIGN

5.1 DATA FLOW DIAGRAM:

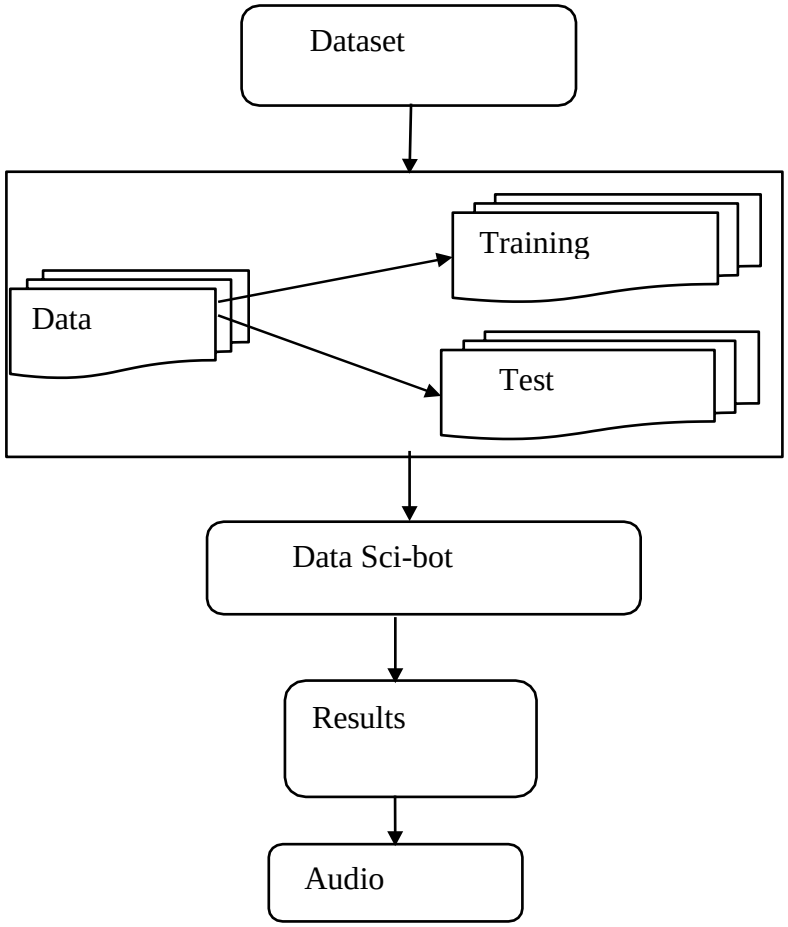
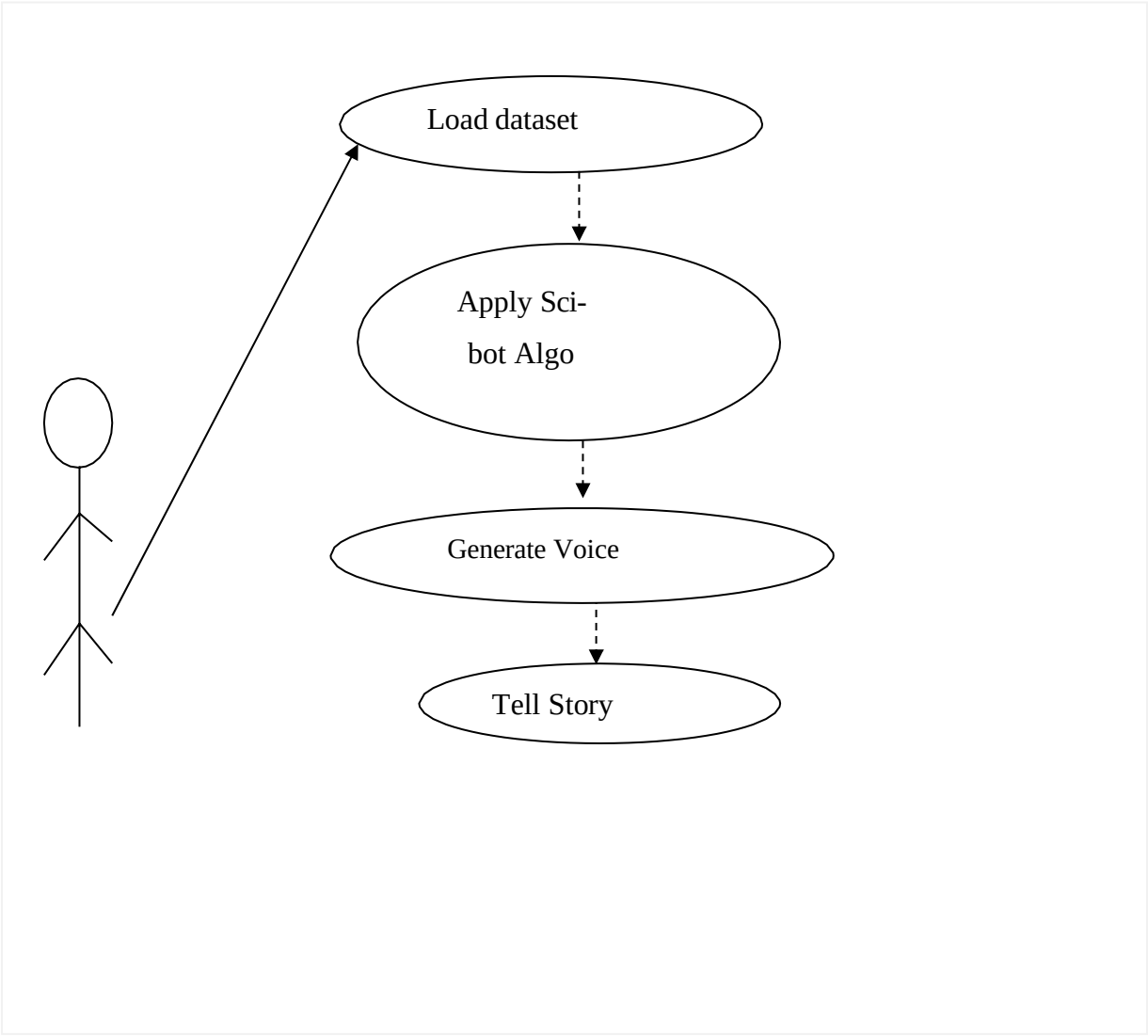


Figure: Data flow Diagram

5.2 USECASE DIAGRAM：



The above figure represents usecase diagram, in which user upload dataset is pre-processed and applied algorithm. They are analyzed for creating the story.

5.3 SEQUENCE DIAGRAM:

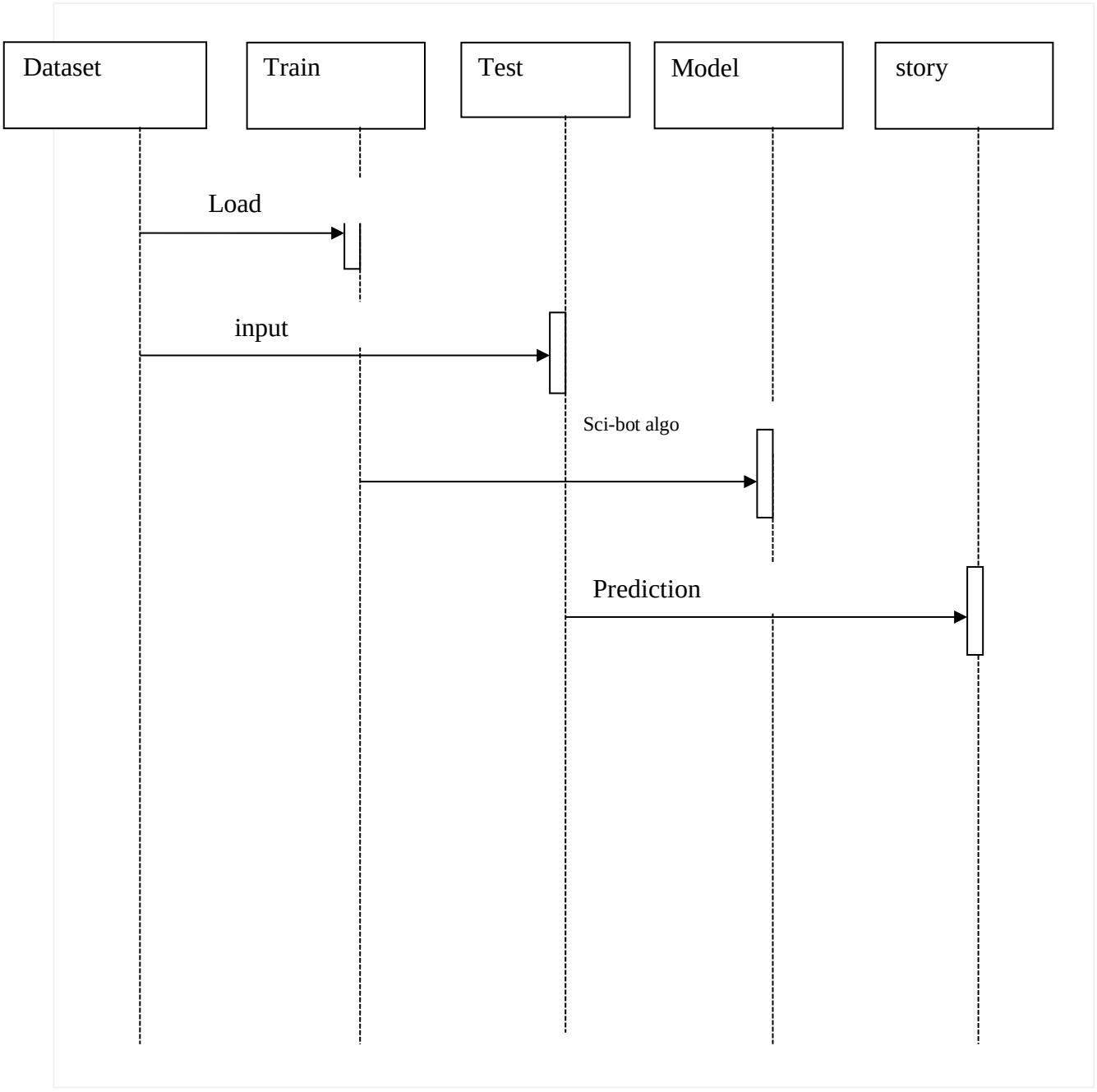


Figure: Sequence Diagram

The above figure represents sequence diagram, the proposed system's sequence of data flow is represented.

5.4 ACTIVITY DIAGRAM:

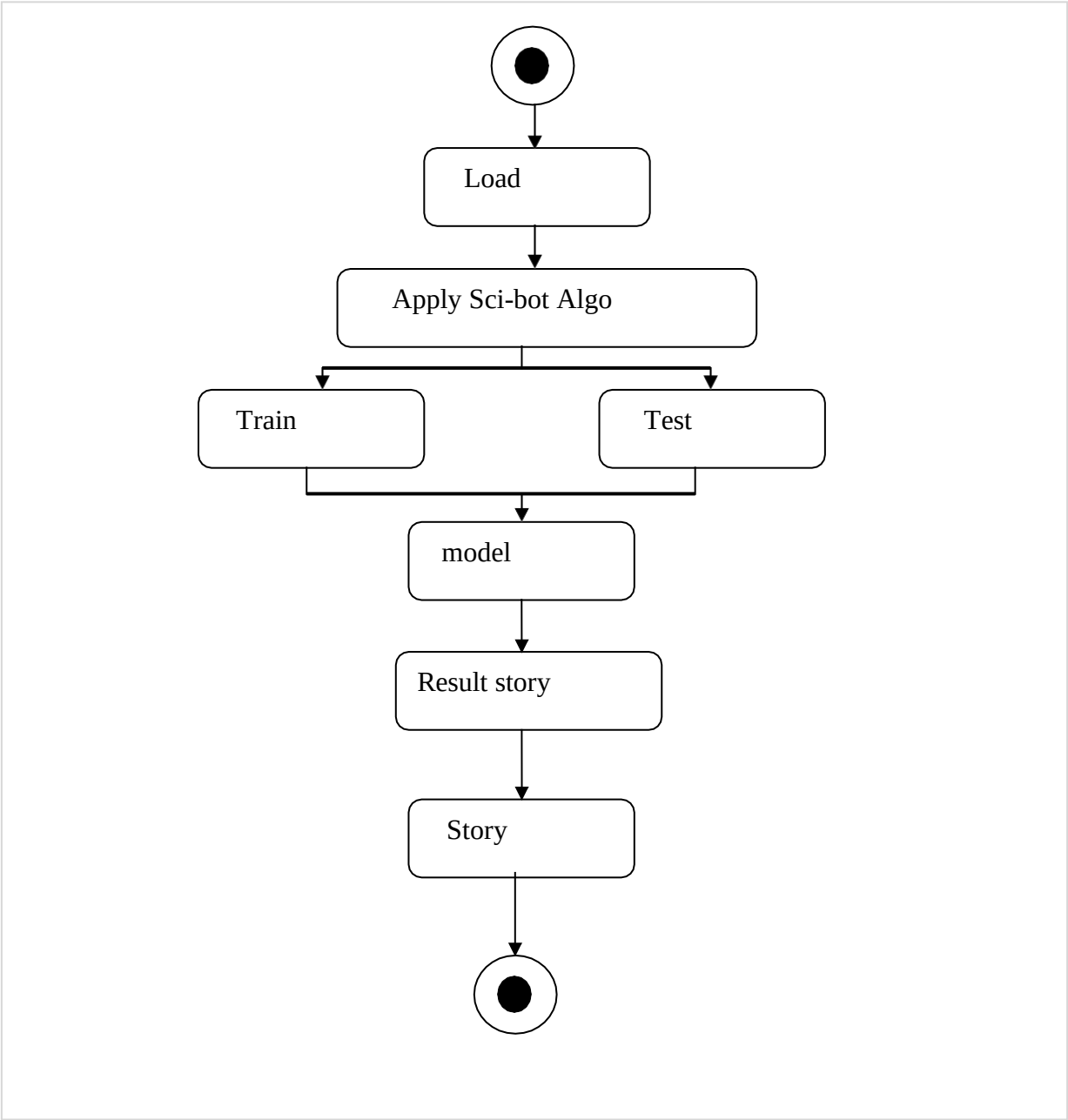


Figure: Activity diagram

The above figure represent activity diagram of proposed system. The figure shows complete flow of activity from dataset loading and all sequence of module.

CHAPTER VI
ALGORITHMS

6.1 Decision Tree:

Decision trees are useful when data possess nonlinear patterns present in it. It is represented as a flow chart in which internal nodes represent the test on a particular-feature and the branches represent output of the test and the child node represents the class in which the input / item is classified. We consider Entropy as the splitting constraint. The decision tree which we get for the employee attrition prediction is shown in Figure 6. The formulae for the Entropy is given by

$$\text{Entropy}(S) = \sum -p_i * \log_2(p_i)$$

6.2 Random Forest :

The working principle is given in Figure. for random forest algorithm It constructs a number of decision trees using CART procedure and the output will be either mean or mode of all trees present in a random forest. This helps the algorithm to avoid overfitting. The diagram representation of the Random forest algorithm is shown below.

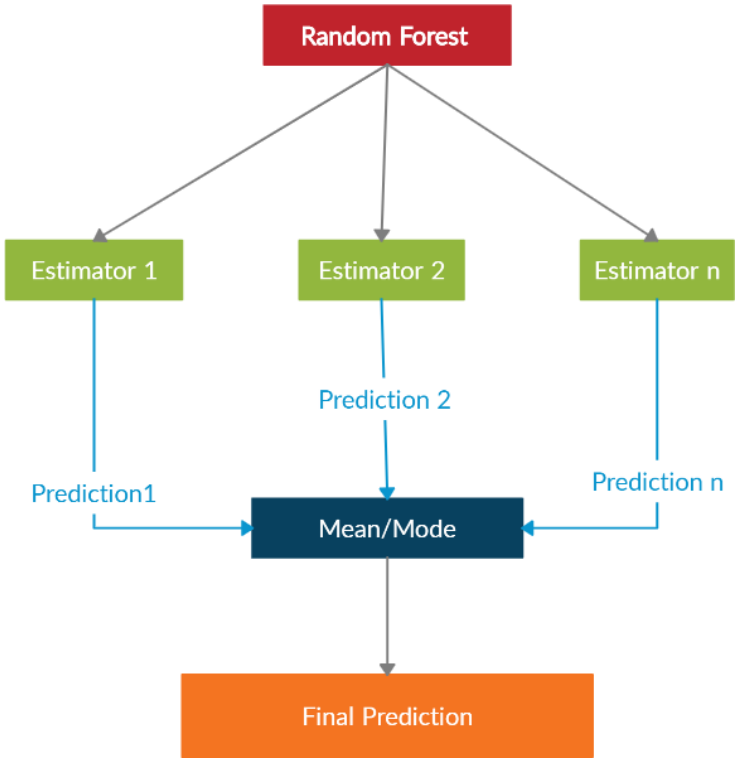


Figure : Illustration of Random Forest Woking Principle

In random forest we use Mean Squared Error to how our data branches from each node.

$$MSE = \frac{1}{N} \sum_i (f_i - y_i)^2$$

yi = Actual data

fi = Predicted data

N = Total number of data points

This formula calculates the distance of each node from the predicted actual value, helping to decide which branch is the better decision for your forest.

While performing random forest we can use either Gini-index or entropy as the criteria for the splitting of

node in decision trees. The formulae for the Gini-index is

$$Gini = 1-\sum(p^2)$$

The formulae for the Entropy is

$$Entropy(S) = \sum -p_i * \log_2(p_i)$$

6.3 Neural Networks : Neural networks are mostly used for image classifications and computer vision to learn patterns on the unstructured data. As we know that Neural Networks works best for unstructured data, it is capable of taking out difficult patterns from data. Since our data is biased, we selected neural networks to find how best this works to our problem.

The input for each individual neuron can be given by $z=w^Tx+b$ where W^T is weight vector and b is bias. The output of the neuron can be given by the function y = g(z) the function g is called sigmoid function . The loss function can be given by

$$J(W,b) = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)})$$
$$L(\hat{y}, y) = -(y \log \hat{y} + (1-y)\log(1-\hat{y}))$$

The sigmoid function is given by

$$G(x) = \frac{1}{1+e^{-x}}$$

6.4 K-Nearest Neighbourhood : KNN algorithm works based on geometrical principles. This algorithm plots all the points in the multi-dimensional plain and then indexes them according to their class . When a new point enters, it gets plotted on plain and then takes the mode of the k nearest points as the class of it. It works best for classification models when there exists a geometrical relationship present among the features of the data.

6.5 eXtreme Gradient Boost (XGB) : XGB algorithm is an advanced version of Random forest. In random forest, each sub tree is independent but in this model the trees are interlinked and error

in the current tree is adjusted through constructing consecutive trees by reducing errors. Every tree has some weight in the final prediction of the model according to its error rate .

The main formulae's we use in the XGB is :

similarity score = ΣΘ / (è +λ)

λ is used for regularization which varies between 0-1

Θ = error residual= difference between the predicted and expected values

è = error samples

6.6 Ada Boost Algorithm: This model is constructed by ensembling the weak learners. Here linear regression is used as the weak leaner for regression models, stumps as the classification models. Stump is a tree with one parent node and two leaf nodes. and work similar to the trees in XGB. Each consecutive stump is created to overcome the errors of previously constructed stump. Adaptive Boosting algorithm performs better on unbalanced data.

CHAPTER VII MODULES

7.1 Data Preprocessing :

We have taken multiple attribute in our case study, dataset 16 features/ attributes are taken fr study. Pre-processing of dataset is done for converting the string attributes to numerals and missing data records are dropped. The pre-processed data is stored in “dataset.csv” file, which is given as input for machine learning models.

7.2 Data Sci-bot Algorithm :

In this our Sci-bot takes the data in and it does all the Data analysis part by itself and finally it creates a story on its self and explains it to the stakeholders/clients. In the AI algorithm the Sci-bot takes the best algorithm for the problem using Auto-Machine Learning Methods. It will automatically take the best parameters for each algorithm using a grid search and Bayesian optimization Technique. Finally to the client everything is internal but the final output is a story that the client will understand.

7.3 Auto Testing :

This is the normal testing procedure in which it calculates the precision and recall and creates a sentence about it automatically . But the key thing here is we are automating the whole process.

7.4 Google/ IBM API:

In this we will send the whole story which was created by our algorithm to any of API and it will return the audio files we will save and store it at some place.

7.5 Local controller :

Its an package which is used to modulate the audio frequency and the phase of the audio . it is used to modulate the algorithm output sound to tell story to the sake holder.

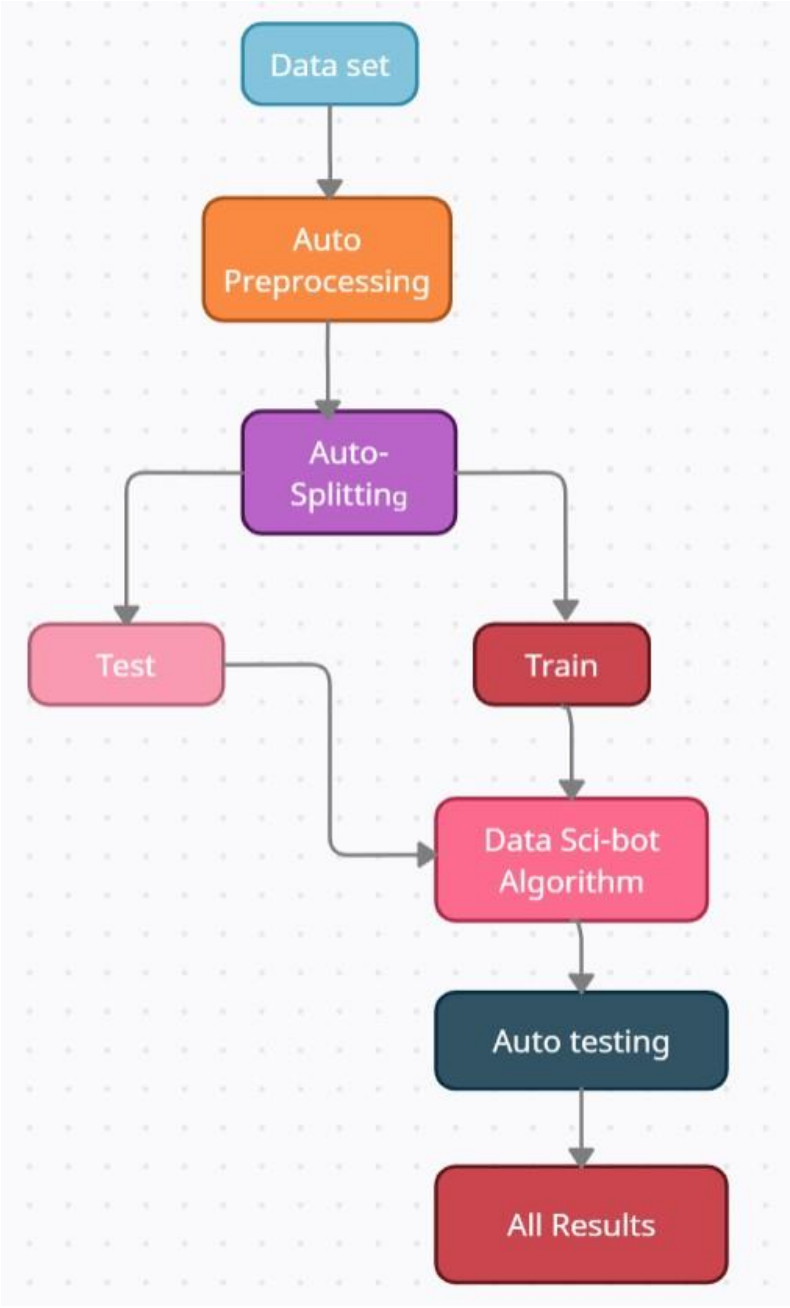
7.6 Scheduler :

Scheduler is an dictionary which is used to store the order of the sentenses that are created during the algorithm worked . It is useful to store the order so that after the text is converted into audio scheduler is used to determine which audio is to be played first.

CHAPTER VIII

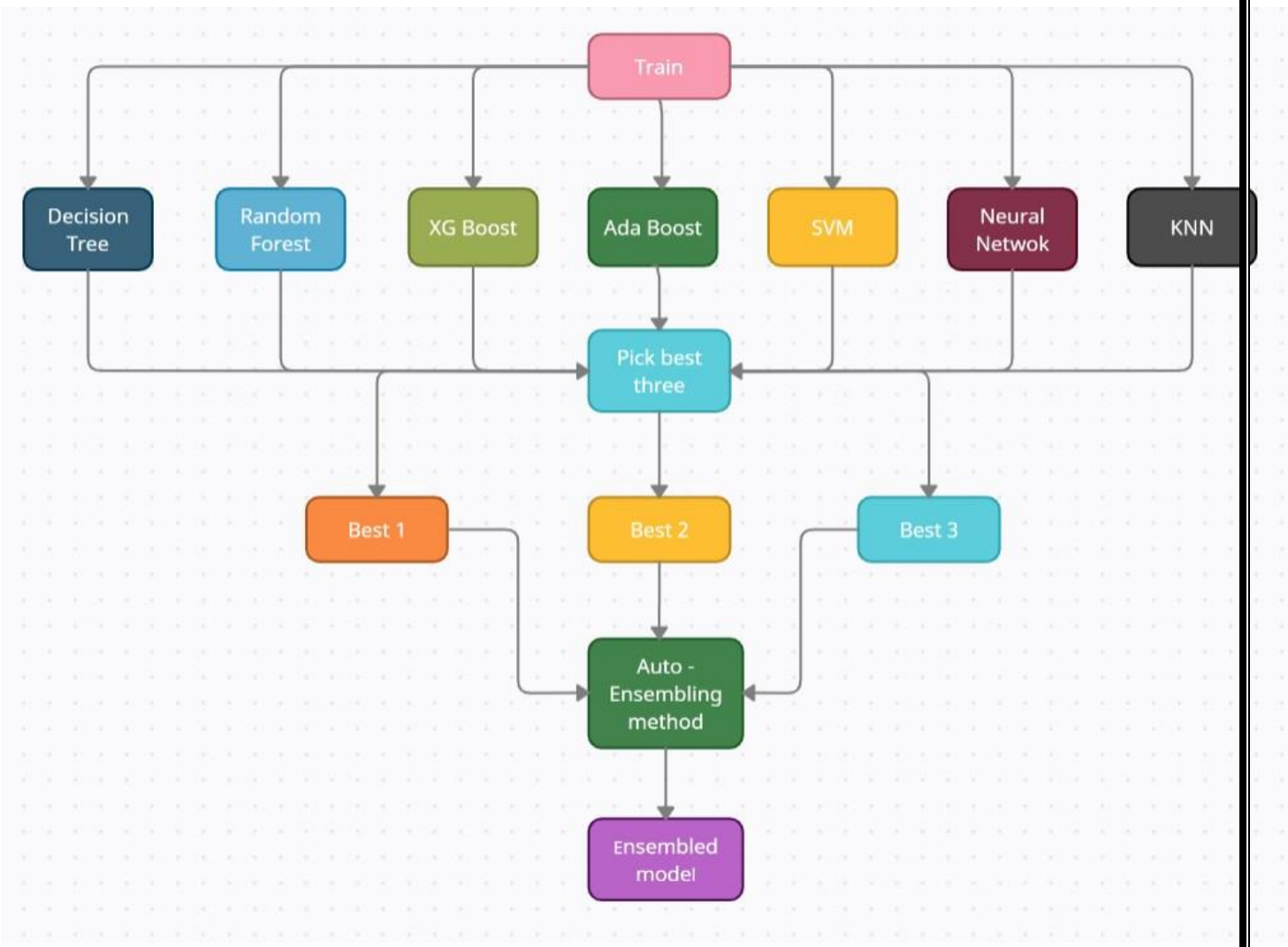
DATA FLOW DIAGRAMS

8.1 Over-all Data Flow Diagram :



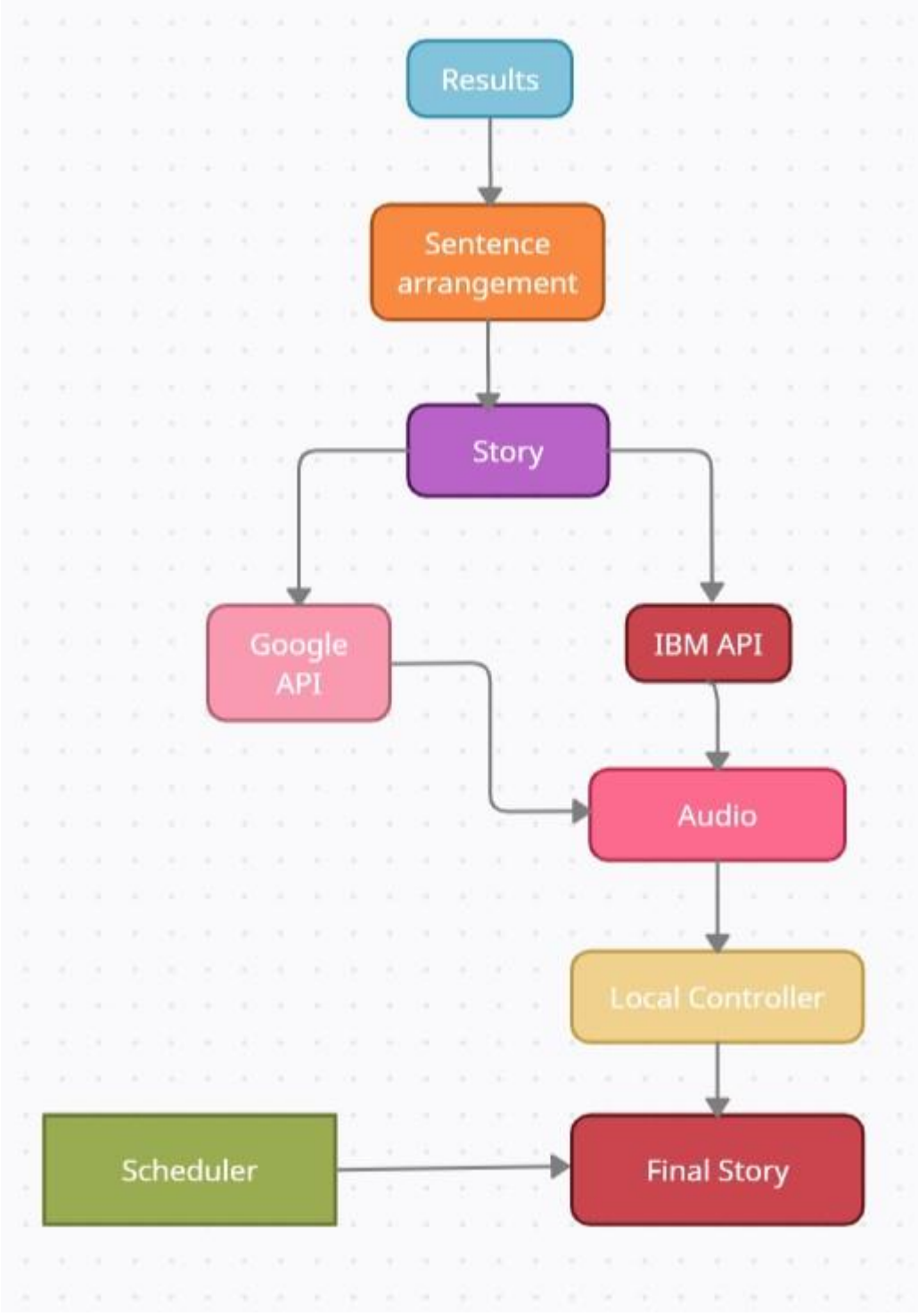
This Diagram shows the overall data flow in the project briefly.

8.2 Sci-Bot Algorithm Data Flow



In the above diagram shows how the data flows in the sci-bot algorithm.

8.3 All Results to Story data flow.



This diagram shows the how the results are converted into story using some api to convert text to audio.

CHAPTER IX

RESULTS AND SCREENSHOTS

CHAPTER X

CONCLUSION

The main objective of the project is to develop an algorithm that will be used to identify answers related to user submitted questions. The need is to develop a database where all the related data will be stored and to develop a web interface. The web interface developed will have two parts, one for simple users and one for the administrator. A background research took place, which included an overview of the conversation procedure and any relevant chat bots available. A database will be developed, which will store information about questions, answers, keywords, logs and feedback messages. A usable system will be designed, developed and deployed to the web server.

REFERENCES

1) Fatemeh Nargesian, Horst Samulowitz, Udayan Khurana Elias B. Khalil, Deepak Turaga. Learning Feature Engineering for Classification, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

2) Huong Ha, Santu Rana, Sunil Gupta, Thanh Nguyen, Hung Tran-The, Svetha Venkatesh, Bayesian Optimization with Unknown Search Space, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).

3) Nagadevara, V., Srinivasan, V. & Valk, R. (2008). Establishing a Link between Employee Turnover and Withdrawal Behaviours: Application of Data Mining Techniques, Research and Practice in Human Resource Management, 16(2), 81-99.

4) Pankaj Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms A case for Extreme Gradient Boosting" International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016.

5) Joao Marcos de Oliveira, Matthäus P. Zylka, Peter A. Gloor, and Tushar Joshi:" Mirror, Mirror on the Wall, Who Is Leaving of Them All: Predictions for Employee Turnover with Gated Recurrent Neural Networks" © Springer Nature Switzerland AG 2019 Y. Song et al. (eds.), Collaborative Innovation Networks, Studies on Entrepreneurship, Structural Change and Industrial Dynamics, https://doi.org/10.1007/978-3-030-17238-1_2 .