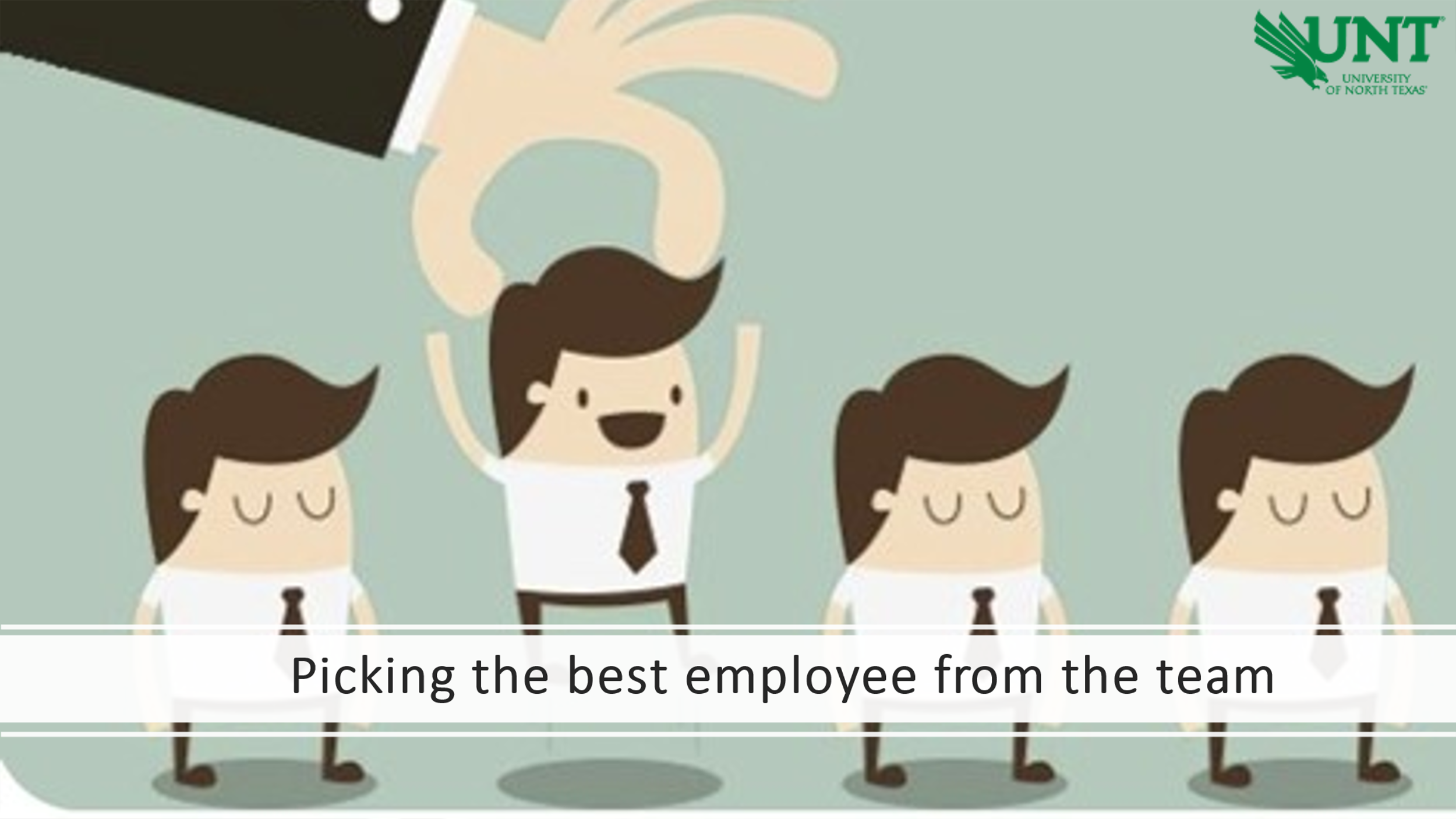
Final Project Report

# Title

**Logo

Description automatically generated**

**HR Analytics – Employee Promotion Prediction AI**



Presented By: **Team 14**

|  |  |
| --- | --- |
| **NAME** | **STUDENT ID** |
| Chinmai Bhavya Kolisetty | 11512962 |
| Ravi Shastri Indasi | 11521492 |
| Vishwambhar Gonguluru | 11616665 |
| Sai Naveen Masadi | 11564630 |

Contents

[1. Title 1](#_Toc121316200)

[2. Executive Summary 3](#_Toc121316201)

[3. Client Organization / Industry Description 4](#_Toc121316202)

[Inspiration 5](#_Toc121316203)

[4. Problem Description 5](#_Toc121316204)

[5. Solution Description 5](#_Toc121316205)

[5.1 Overall description of solution 5](#_Toc121316206)

[5.2 What does solution do? What process does solution serve? 6](#_Toc121316207)

[5.3 Who are the users of the solution? How do they use the solution? 6](#_Toc121316208)

[5.4 What are data inputs into the solution at the runtime (once implemented)? 7](#_Toc121316209)

[5.5 What are the outputs of the solution? Who uses these outputs? Are they used as an input into a different system? 7](#_Toc121316210)

[5.6 What component of the solution can be described as AI and why? 7](#_Toc121316211)

[6. Solution Data Requirements 8](#_Toc121316212)

[6.1 What data will need to be provided for developing the solution? 8](#_Toc121316213)

[6.2 What are the quality requirements on the data vs. existing data quality? 8](#_Toc121316214)

[6.3 Data transformation pipeline requirements? 8](#_Toc121316215)

[7. Acquisition Strategy 9](#_Toc121316216)

[7.1 Will your solution be using AI Cloud APIs? Why is this the best solution? 9](#_Toc121316217)

[7.2 AI Platform/proprietary algorithm? Why is this the best solution? 9](#_Toc121316218)

[7.3 Custom environment/Algorithm? Why is this the best solution? 10](#_Toc121316219)

[8. Underlying ML Model requirements 10](#_Toc121316220)

[8.1 Is your AI ML based? If not, explain why it is AI. Explain what AI components that are non-ML based are present. 10](#_Toc121316221)

[8.2For ML-based AI, what will be the role of the ML model? Describe the model in terms of the target variable and the algorithm used. Describe the algorithm with references 10](#_Toc121316222)

[Bagging 11](#_Toc121316223)

[Boosting 12](#_Toc121316224)

[Unique features of XGBoost 12](#_Toc121316225)

[8.3 Define performance criteria for determining that the model is fit for use. Accuracy, precision/recall, R-sqrd, MAPE? 13](#_Toc121316226)

[9. Solution Architecture 14](#_Toc121316227)

[9.1 Graphical representation accompanied by brief description of the solution components and how they would work together: 14](#_Toc121316228)

[9.2 Triggers/data inputs 15](#_Toc121316229)

[9.3 Pipelines and data transformation 15](#_Toc121316230)

[9.4 Data/model storage 15](#_Toc121316231)

[9.5 Model endpoint 15](#_Toc121316232)

[9.6 Outputs 16](#_Toc121316233)

[10. Which of the components have you prototyped? Describe your prototype and include any screenshots necessary 16](#_Toc121316234)

[11. Code 19](#_Toc121316235)

[12. References 19](#_Toc121316236)

# 2. Executive Summary

Promotion processes are one of the most important processes in terms of human resources. A promotion process organized fairly within the organization is a managerial tool that motivates employees and contributes to business continuity. Promotion is an important extrinsic motivation for many employees. It ensures the employee's engagement and commitment to the organization and contributes to the continuity of his current performance. It is also an important rewarding and performance control mechanism for the organization. Many factors such as seniority, performance level, competencies, age, awards, training score, organizational commitment of the personnel who will be promoted are taken into consideration. In this report, a prediction methodology will be studied based on the criteria evaluated for the employees in the promotion processes by Machine Learning algorithms such as Decision Tres, Random Forest, Naïve Bayes Classifier, Logistic Regression and XGBoost. XGBoost on tuning with the right hyperparameters achieved the highest performance with 97.8 % accuracy, 97.7 % precision, 98 % recall and 97.8 % f1-score values with ROS approach. This study could be used by HR and manager to predict the probability of promotion so that managers can find the right parameters for someone to get promoted (Sahinbas, 2022).

Promotion is the transition of any personnel in the enterprise from their position to another task that will increase their authority and responsibilities even wider and increase their status. The fact that the employees come to a better place than they are in the institution be an increase in motivation for the workers. Because it is of great importance for the employee to know that his/her labor will be rewarded and that he/she will have the opportunity to rise, both in terms of increasing his/her commitment to the job and working more efficiently. When this is the case, both production, efficiency and quality will increase. This will enable the company to continue its production with great profitability. As a result, both the employee and the employer will be happy. If an employee is promoted and given more benefits, they become more satisfied and committed to their work. Satisfied employees work harder and more readily, and employee turnover is minimized (Sahinbas, 2022).

# 3. Client Organization / Industry Description

A large MNC have 9 broad verticals across the organisation. A large MNC have 9 broad verticals across the organization. Finding the appropriate candidates for promotion(only for manager position and below) and getting them ready in time is one of the issues

Since the evaluation comes only after the final promotions, the transition to new responsibilities is delayed. Therefore, in order to speed up the entire promotion cycle, the organization needs assistance in recognizing the qualified candidates at a given checkpoint

With regard to the employee's previous and present performance as well as demographics, multiple attributes have been presented

Features:

* EmployeeId: Employee Unique ID
* Department: Employee of the department
* Region: Region of employment
* Education: Education of the Employee
* Gender: Employee Gender
* Recruitment\_channel: The method used to recruit employees
* Num\_ of\_ trainings: Number of other trainings completed in previous year on soft skills, technical skills etc.
* Age: Employee Age
* Previous\_ year\_ rating: Rating of employees for the previous year
* Length\_ of\_ service: Service period in years
* Awards\_won?: Awards earned during the previous year, 1; otherwise, 0.
* Avg\_ training\_ score: Average score on training evaluations at the moment
* is\_promoted: (Taget Variable) Recommended for promotion

Inspiration

* Predict whether a potential promote at checkpoint in the test set will be promoted or not following the evaluation procedure.

# 4. Problem Description

It’s often a cumbersome process for a large company to manage the employee promotion problem as there is a continuous flow of new employees, resignations, lay-offs all taking place within similar time frame. In a big MNC, is it essential to identify the right candidate for promotion so that the MNC can prepare him/her for the lead position and managerial posts on time. This process usually requires the support of experienced domain expertise along with the time and cost spent in the process. To cop up with stability of the MNC, mutation in roles in several department is required, which needs employees upskilling and training and picking up the right ones comes at resource cost and time. This is something the industry has been dealing with for a long now and looking forward to some breakthrough solution which can be a decision support system designed for Human Resource (HR) departments about eligibility of employees’ promotion.

# 5. Solution Description

## 5.1 Overall description of solution

The solution focuses on building up a decision support system for the Human Resource (HR) department of the MNC which will help them make a precise decision on picking up the right employee for promotion who can further be sent for upskilling/training.

The solution focuses on following a **Data Centric Approach**, this approach uses the past data of employees of the MNC to make smart decisions using AI Algorithms. The data is stored in the company’s cloud server, which is then pulled into a local database. The raw data is preprocessed and used as sheets. This preprocessed data is fed to AI Algorithms which make up smart decisions based on the features on the employee such as average training score, length of service of an employee etc.

## 5.2 What does solution do? What process does solution serve?

The solution is an **HR Analytics AI App** which takes in the given set of features as input as listed below:

* Education of the employee
* No of trainings the employee went through
* Age of employee
* Previous year rating
* Length of service of employee
* Awards status of employee
* Average training score of the employee

and return results for employee promotion prediction as shown in a GIF below:

Graphical user interface, application

Description automatically generated

## 5.3 Who are the users of the solution? How do they use the solution?

The solution is primarily designed to be used by Human Resource (HR) department of the MNC. The HRs sort out the list of potential employees from whom they select to be in a pool of contenders for promotion. The data for these employees is fed to the HR Analytics AI App to find out if the employee is a god fir for promotion or not

## 5.4 What are data inputs into the solution at the runtime (once implemented)?

The solution uses the following input features from the dataset:

* education: Education of the employee
* no\_of\_trainings: No of trainings the employee went through
* age: Age of employee
* previous\_year\_rating: Previous year rating
* length\_of\_service: Length of service of employee
* awards\_won?: Awards status of employee
* avg\_training\_score: Average training score of the employee

## 5.5 What are the outputs of the solution? Who uses these outputs? Are they used as an input into a different system?

The output of the system is a prediction which can be either of the two:

* The employee can be considered for promotion
* The employee cannot be considered for promotion

While these outputs may be used by the HR department to plan further steps, they certainly do not play any further role in the HR Analytics AI App

## 5.6 What component of the solution can be described as AI and why?

The components which can be described as AI in the solution is the smart prediction, which uses AI Algorithms under the hood to make those decisions. More details about these algorithms is discussed further in the report

# 6. Solution Data Requirements

## 6.1 What data will need to be provided for developing the solution?

AI algorithms work better when they are fed more and more data. To make smart and accurate decisions on employee attributes, there needs to have enough data on which the algorithms could be trained on. In this case, as discussed above in report, there are different attributres such as employee\_id, department, region, education, gender, recruitment\_channel, no\_ of\_ trainings, age, previous\_ year\_ rating, length\_ of\_ service, awards\_won?, avg\_training\_score, is\_promoted

Though these features are filtered in the data preprocessing stage and a subset of features is selected when training the models.

## 6.2 What are the quality requirements on the data vs. existing data quality?

As a rule of thumb, the AI solutions follow a rule of garbage-in-garbage-out, which means that the better and clean data is what makes the AI algorithms perform the best. This means that the data provided to the algorithms should be clean/preprocessed. Similarly, to solve the employee prediction problem, there is a requirement to have algorithms perform better, which require quality data. For example, algorithms like XGBoost require the data to not have any outliers, as it is sensitive to outliers, similarly, it is a better choice to standardise/normalize data to get better predictions.

## 6.3 Data transformation pipeline requirements?

The data needs to be filtered to attain the maximum performance. Various steps involved in the data transformation process are:

* Missing values check and imputation
* Outlier detection and removal
* Data imbalance check and fix
* Feature label encoding
* Feature scaling/normalization
* Feature selection

All these steps were performed in the project to attain an accuracy of ~ 98 % from the algorithm.

# 7. Acquisition Strategy

## 7.1 Will your solution be using AI Cloud APIs? Why is this the best solution?

The solution has been developed keeping in mind the minimum dependencies on third party APIs. This makes the solution robust to roadblocks such as API charges, API rate limiting, server down issues, API integration etc.

Therefore, it can be called as the best solution because it is:

1. Platform independent: Can work on multiple OS such as Linux, Mac OS, Windows etc.
2. Resource efficient: The solution is optimized to run on a machine with minimal working configuration
3. Intuitive UI: The UI is user friendly. Anyone with bare minimum knowledge can adapt to working on it in no time

## 7.2 AI Platform/proprietary algorithm? Why is this the best solution?

The proprietary algorithm used in this solution is XGBoost. This is state of the art algorithm which used boosted trees to deliver accurate results. A comparison of algorithms performance is shown below in the figure. It can be observed that, with the right parameters, XGBoost is the best among all

A screenshot of a computer

Description automatically generated with medium confidence

The evaluation metrics show that XGBoost has ~ 98 % accuracy, 0.97 F1 Score, 0.98 Recall & 0.97 precision scores, which are considered to be good and reliable scores for an algorithm to work.

## 7.3 Custom environment/Algorithm? Why is this the best solution?

The algorithm uses Python 3.9 based Anaconda environment with specific libraries versions such a scikit learn etc. Working with a custom environment is best solution because it doesn’t mess up with the main environment of the system. Since, the solution is hosted on a cloud server, an enclosed environment makes sure that during the update of the OS within the server, the dependencies might not break the app. Therefore, this is the best and reliable solution.

# 8. Underlying ML Model requirements

## 8.1 Is your AI ML based? If not, explain why it is AI. Explain what AI components that are non-ML based are present.

Yes, the AI used in the solution is ML based. Since, ML based AI solutions are very light weight as compared to Deep Neural Network Models which require the weights and biases to be stored. These models are heavy and might introduce latency in the app as well as contribute to the load on the server as libraries such as Tensorflow and PyTorch are relatively big in size as compared to that of scikit-learn and XGBoost . Also, in this case, the performance from classical ML algorithms was found out to be good enough to not rely on Deep Learning solutions.

## 8.2For ML-based AI, what will be the role of the ML model? Describe the model in terms of the target variable and the algorithm used. Describe the algorithm with references

A total of five ML based models tried and tested during the approach:

* Decision Trees
* Random Forest
* Naïve Bayes
* Logistic Regression
* XGBoost

XGBoost was found out to be the best performer with the right set of hyperparameters.

The problem statement was framed into a classification problem. The target variable has values 0 or 1 where 0 represents that the employee is not considered for promotion while 1 represents that the employee can be considered for promotion.

Using the employee features, the learning algorithm is trained to predict that given a set of features, is the output is 0 or 1?

**XGBoost Description**

XGBoost is an ensemble method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. (Jake Robinson, 2018)

The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. **Bagging and boosting are two widely used ensemble learners**. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees. (Jake Robinson, 2018)

### Bagging

Even though decision trees are among the models that are easiest to understand, their behavior is incredibly unpredictable. Think about a single training dataset that was randomly divided into two sections. Let's now use each component to train a decision tree to obtain two models (Jake Robinson, 2018).

At the point when we fit both these models, they would yield various outcomes. Decision trees are supposed to be related with high change because of this way of behaving. Bagging or boosting accumulation assists to reduce the variance in any learner. A few decision trees which are generated in parallel, structure the base leaners of bagging strategy. Data sample with substitution is taken care of to these learners for training. The final prediction is the found the averaged value of result from every one of the learners.

### Boosting

In boosting, the trees are constructed consecutively with the end goal that each ensuing tree means to lessen errors the previous tree. Each tree learns from its ancestors and updates the remaining errors. Consequently, the tree that fills next in the sequence will learn from a updated version of the residuals.

The base learners in boosting are feeble learners in which the bias is high, and the predictive power is somewhat better compared to random guessing. Every one of these weak learners contributes some important data for prediction, empowering the boosting method to produce a strong learner by effectively combining these weak learners. The last solid learner cuts down both the bias and the variance (Jake Robinson, 2018).

As opposed to bagging procedures like Random Forest, in which trees are developed to their maximum extent, boosting makes use of trees with fewer splits. Such little trees, which are not exceptionally profound, are profoundly interpretable. Parameters like the no of trees or cycles, the rate at which the gradient boosting learns, and the depth of the tree, could be ideally chosen through validation techniques like validation. a large number of trees might lead to overfitting. In this way, it is important to choose the stopping criteria for boosting (Jake Robinson, 2018).

Unique features of XGBoost

XGBoost is a well known execution of gradient boosting. A few features make XGBoost so fascinating:

* **Regularization:**XGBoost has a choice to penalize complex models through both L1 and L2 regularization. Regularization helps in forestalling overfitting (Jake Robinson, 2018)
* **Handling sparse data:**Missing values or information handling steps like one-hot encoding make data sparse. XGBoost consolidates a sparsity-aware split tracking down calculation to deal with various sorts of sparsity designs in the data(Jake Robinson, 2018)
* **Weighted quantile sketch:** Most existing tree-based calculations can find the split points when the data focuses are of equivalent weights(using quantile sketch algorithm). Nonetheless, they are not prepared to handle with weighted data. XGBoost has a circulated weighted quantile sketch algorithm to really deal with weighted information (Jake Robinson, 2018)
* **Block structure for parallel learning:**For quicker computing, XGBoost can utilize various cores on the central processor unit. This is conceivable due to a block structure in its system design. Data is arranged and put away in-memory units called blocks. Dissimilar to different algorithms, this empowers the data layout to be reused by resulting cycles, rather than computing it again. This feature additionally serves valuable for steps like split finding and column sub-sampling (Jake Robinson, 2018)
* **Cache awareness:** In XGBoost, non-continuous memory access is expected to get the gradient statistics by row index. Subsequently, XGBoost has been intended to utilize hardware. This is finished by designating internal buffers in each thread, where the gradient statistics can be stored (Jake Robinson, 2018)
* **Out-of-core computing:**This feature enhances the available disk space and expands its utilization while taking care of vast datasets that do not fit into memory (Jake Robinson, 2018)

## 8.3 Define performance criteria for determining that the model is fit for use. Accuracy, precision/recall, R-sqrd, MAPE?

Since this is a classification problem, the evaluation metrics used here are:

* **Accuracy:** The XGBoost Model came out to have the best accuracy
* **F1 Score:** The XGBoost Model came out to have the best F1 Score
* **Precision:** The XGBoost Model came out to have the best accuracy
* **Recall:** The XGBoost Model came out to have the best accuracy

# 9. Solution Architecture

## 9.1 Graphical representation accompanied by brief description of the solution components and how they would work together:

The figure below shows the graphical architecture of the solution. The figure depicts the workflow from data collection, data formatting etc. to the final deliverable, which is, the HR Analytics AI Application.

**Diagram

Description automatically generated**

The solution components consist of the following components:

* Inputs from fields such as textbox, select box, sliders etc.
* A Database having stored employee data with which the app is connected
* Smart AI algorithm running in the backend which is heart of the Application

## 9.2 Triggers/data inputs

The solution uses the following input features from the dataset:

* Select Education
  + Master’s & Above
  + Bachelor’s
  + Below Secondary
* No of trainings done by employee
  + A number in range (1, 10)
* Previous year rating of employee
  + A number in range (1,5)
* Length of service of employee
  + A number in range (1, 37)
* Did employee win any awards?
  + Yes
  + No
* Average training score of the employee
  + Integer input

## 9.3 Pipelines and data transformation

The figure below shows the data transformation pipeline used in the project

Icon

Description automatically generated

## 9.4 Data/model storage

The data and model are in a cloud server which is consumed by the Application

## 9.5 Model endpoint

The model end point is a prediction, which shows if the employee is eligible for promotion or not

Using the Secured certified HTTP Method.

## 9.6 Outputs

We are defining the Output with the Flag as O and 1 for Target variable (Is\_Promoted) which is nothing but Employee Promoted as 1 and Employee not Promoted as 0 in the GUI.

# 10. Which of the components have you prototyped? Describe your prototype and include any screenshots necessary

The screenshots below show the prototype, the prototype takes the user input as discussed above in the report

Graphical user interface, application

Description automatically generated

The predictions of the app are shown below in the figures –

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

# 11. Code

Here is the attached PDF document which contains the code that we developed for this solution. Please click on the link

[HR\_Analytics\_Code.pdf](https://myunt-my.sharepoint.com/:b:/g/personal/ravishastriindasi_my_unt_edu/EUS-rHBa44pHvYkmFAHmwMIBwcYAPlcxdMESJd6IpmRb9w?e=zHCFvK)

# 12. References

**Resources and Related Projects:**

<https://www.kaggle.com/arashnic/hr-ana>

[https://books.google.co.in/books?hl=en&lr=&id=EXiJDwAAQBAJ&oi=fnd&pg=PP1&dq=hr+analytics&ots=bHARVWIhOn&sig=Ap\_22xpfVYtIoBxPjcrqNzmChD4&redir\_esc=y#v=onepage&q=hr%20analytics&f=false](https://books.google.co.in/books?hl=en&lr=&id=EXiJDwAAQBAJ&oi=fnd&pg=PP1&dq=hr+analytics&ots=bHARVWIhOn&sig=Ap_22xpfVYtIoBxPjcrqNzmChD4&redir_esc=y)

<https://www.emerald.com/insight/content/doi/10.1108/14754390580000607/full/html>

<http://u.camdemy.com/sysdata/doc/4/4f6fc50c64b48b94/pdf.pdf>

<https://www.youtube.com/watch?v=Jfh15OGmmtw>