**HR ANALYTICS PROJECT**

Integrating analytics in the workforce development field is becoming a mandate, and the human resource (HR) analytics approach is a promising technique that allows forming evidence-based decisions and reduce risk associated with investments decisions and focuses on putting HR analytics into practice by connecting the HR analytics approach to machine learning, as one of the most advanced analytical approaches, to solve a real business problem. Machine learning is one of the most advanced analytics techniques that is based on revealing deeper insights in big data by using historical data to predict the future.

**Problem Definition:**

HR analytics is a multidisciplinary approach that focuses on improving the quality of human capital decisions according to the use of data. It focuses on the use of advance analytics techniques as a decision support tool to predict future events) indicated that HR analytics helps organizations understand what happened in the past and what is happening now to give good predictions for future risks and Machine learning is an artificial intelligence approach that enables machines to emulate human’s intelligence. More specifically, machine learning is a set of algorithms that trains data by building a predictive model that learns from the available data, and then assesses the accuracy of the trained model to examine its validity. The unique feature of machine learning models is based on their ability to get smarter by adapting from experience. The model is able to predict outcomes from new data by independently. Machine learning provides new opportunities for organizations to use continuous emerging data to accurately make decisions and direct investments.

**Data Analysis:**

This phase starts by collecting data that is relevant to the machine learning project objectives. Because data collection is a very time-consuming process, an accurate identification to the machine learning project The data understanding phase allows exploring insights in the data by understanding the context in this phase starts by collecting data that is relevant to the machine learning project After collecting the right data that aligns with the predictive modelling goal, it is important to get a deeper understanding of the data. The data understanding phase allows exploring insights in the data by understanding the context in win the machine learning modelling process. It provides a visual representation to the distribution of the data to uncover any outliers or skewness in the data. Additionally, data visualization reveals any correlations between variables which help directing the selection of variables that will be included in the study. Correlation between variables. Using highly correlated variables result in building low performance models. In this study the correlation analysis and the hierarchal analysis techniques will be used to examine the correlation between all attributes, and the correlation between each attribute and the target variable. The goal is to identify variables with high level of correlation as these variables may misguide the machine learning models results.

**Data Preparation:** Another important step in developing a predictive model is preparing the data for the analysis. It is significant to make sure that the data is clean and free of errors, as data could involve various issues that affect the performance of the predictive model. Data mining projects depends on the effort that devoted to preprocessing the data used in building the predictive models. Data preparation involves cleaning data from errors, dealing with missing values, data transformation, detecting outliers, encoding variables, and removing problematic features.

**Modeling:** Selecting machine learning algorithms for modelling. Once the data is prepared and the task of the predicted model is clearly identified, the next step is to select the type of machine learning algorithms that will be used. Determining the type of algorithm that will be used in the modelling is based on the predictive task, as some models are suitable for classifications, regression, or clustering task. This study will focus on predicting employee absenteeism using three classification modelling techniques: logistic regression, decision tree, and neural network model. Although the target variable in this study is a numerical variable that indicates employee absenteeism in hours, a transformation will be applied to create a classification problem by transforming the target variable to a binary variable that indicates whether an employee is absent or not. Training and assessing the model. Before starting the modelling process, data should be randomly partitioned into a training dataset and a test dataset. The training dataset will be used to train the model and the test dataset will be used to evaluate the model using unseen data that has not been part of the training process. Starting with a simple model. After partitioning the data, the modelling process should start with a simple model. The logistic regression model is a good model to start with as it is a simple model that gives information regarding the correlation between variables and the target variable. This information allows improving the modelling process and reduce the model complexity by eliminating variables that don’t contribute to predicting the output variable. Repeating with different models. After training and assessing the first model, the same modelling process will be repeated using different predictive algorithms. To ensure high accuracy in the modelling process, it is important to try out different modelling approaches and choosing the model with the highest performance. In this study, three predictive approaches will be used. The first approach is the logistic regression model as a simple method to create a classification model. For the other two models, the study will use the decision tree and the neural networks approaches as two commonly used predictive models in classification problems.

**Model Evaluation:**

The evaluation process contains two phases. The first phase evaluates the performance of each model individually by comparing the accuracy of the training dataset against the test dataset. The second phase focuses on comparing the performance of all models against each other to choose the model with the highest performance. Both of these evaluation approaches are very effective in evaluating the performance of the machine learning models and presenting the results in an interpretable way. Specifically, the confusion matrix makes a cross-tabulation of the actual and the predicted observations and provides a straightforward interpretation to the model accuracy and the overall error. On the other hand, the ROC chart implements a visual representation to the model accuracy

**Logistic Regression Model:**

The logistic regression model is a simple model to start the modelling process. It gives a foundational understanding of the model’s performance as well as the significance of the relation between the input features and the output variable. Understanding this relation will help to enhance and simplify the model by eliminating all input features that do not seem to contribute to predicting the outcome variable.

**Decision Tree Model:**

The third predictive model will be used in this study is the decision tree model. The decision tree approach is used to solve regression and classification problems. This approach is based on the inductive inference process that focuses on classifying objects based on analysing a set of attributes with known classes. More clearly, it is a decision-making tool that assigns a probability to the possible outcomes. This probability is based on asking a set of questions regarding one of the input features. The decision tree model is based on setting up a series of rules that is structured in a tree shape. The tree-based structure consists of series of split points called the nodes. The first node is called the root node and it starts at the top of the tree. The root node splits into two or more branches. These branches lead to further nodes that might split to form another node called the test node or terminate to form a node called the leaf or decision node. The root node places a condition regarding one of the input features. Basedonthat condition, the root node classifies the attribute to further move to the right side or the left side of the tree to form a new inner-node. This process is repeated until the tree reach the leaf node where a specific value of the output is given There are various approaches to building a decision tree model. As this study focuses on building a classification model, the Classification and Regression Tree (CART) methodology will be used as one of the most recent approaches to building a tree-based model . Building a classification tree model with the CART approach relies on measuring the purity of a node to score nodes based on whether they belong to one of the output classes. To measure the node purity, the analysis will use the Gini index that is frequently used in CART to solve a classification problem. For a binary classification, the Gini index is computed by estimating the probability of every class and multiplying this by the probability of not being in that class.

**Conclusion:**

Machine learning models can be constructed as a project planning process that follows the structure of the CRISP-DM model. The first step for building a machine learning model is clearly understanding the business problem and constructing a data mining problem relevant to the business objective. Another important step that contributes to bringing about the machine learning model’s success is collecting the right data that could answer the business question. Additionally, understanding and effectively preparing the data used in the analysis are very important steps to avoid biased predictions. Building various models is also an important step in machine learning to ensure using the model with the highest accuracy. Finally, explaining the model’s predictions is essential for building trust with the model’s results and ensuring the effectiveness of decisions made based on those predictions.