**CLASSIFICATION: HR ATTRITION DATA BASED ON IBM ATTRITION**

**REGRESSION: INSURANCE PREMIUM PREDICTION**

*Machine Learning Group Project Report Submitted by*

**RIZWANULLAH MOHAMMAD ( 208W1A1299 )**

**SAMBANA HARSHITA ( 208W1A12C1 )**

**ABHI VENKAT SAI SAMSANI ( 208W1A12C2 )**

**( BATCH – 18 )**

****

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**V R SIDDHARTHA ENGINEERING COLLEGE**

**(AUTONOMOUS - AFFILIATED TO JNTU-K, KAKINADA)**

**Approved by AICTE &Accreted by NBA**

**KANURU, VIJAYAWADA-7**

**ACADEMIC YEAR**

**(2022-23)**

**TABLE OF CONTENTS**

**1. PROBLEM STATEMENT**

**2. SCOPE OF THE PROJECT**

**3. ARCHITECTURE/METHODOLOGY/ALGORITHM**

**4. DATASET DESCRIPTION**

**5. EVALUATION MEASURES**

**6. EXPERIMENTAL RESULTS**

**7. CONCLUSION**

**REFERENCES**

**1.fPROBLEM STATEMENT**

**CLASSIFICATION: HR ATTRITION DATA BASED ON IBM ATTRITION**

The dataset contains the same features as the IBM dataset with a few added features to help us with the project, which includes the survival analysis and prediction of an employee, whether he/she would attrite or not.

**2. SCOPE OF THE PROJECT**

The scope of the problem statement is to build a classification model to predict the attrition of employees in an organization based on the HR data. The dataset includes features that can help with survival analysis and prediction of employee attrition. The goal is to develop a model that can accurately predict whether an employee is likely to leave the organization or not, based on the available data. This problem statement is relevant for organizations that are interested in understanding the factors that lead to employee attrition and in taking measures to retain valuable employees.

**3. ARCHITECTURE/METHODOLOGY/ALGORITHM**

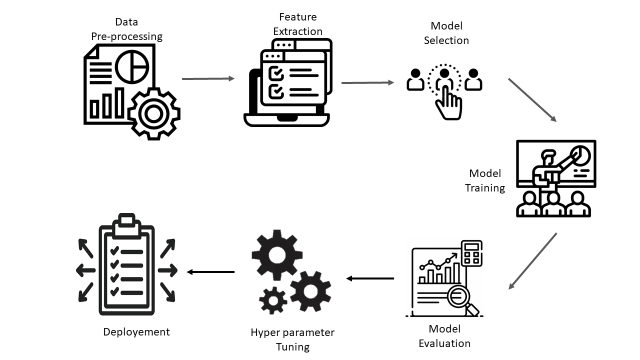


Fig 1 : Architecture Diagram

Data Input: This component represents the initial stage of your model where you will import your CSV file dataset into the system. The dataset will include features related to Lumpy Skin Disease such as age, breed, sex, clinical signs, and so on.

Data Preprocessing: This component involves cleaning and processing the raw data to prepare it for analysis. It may include tasks such as data cleaning, data transformation, handling missing data, and feature engineering.

Model Building: This component represents the main stage of your model where you will use three different algorithms, namely random forest, navies bayes, and decision tree classifier, to build predictive models for Final Attrition. Each algorithm will be trained on a subset of the preprocessed data.

Model Evaluation: This component involves evaluating the performance of each algorithm in predicting Final Attrition using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Based on the evaluation results, you will select the best algorithm that yields the highest performance.

Output: This component represents the final stage of your model where you will generate an output in the form of a prediction of Final Attrition based on the input data. The output will be generated using the selected algorithm with the highest performance.

**4. DATASET DESCRIPTION**

Uncover the factors that lead to employee attrition and explore important questions such as ‘show me a breakdown of distance from home by job role and attrition’ or ‘compare average monthly income by education and attrition’. This is a fictional data set created by IBM data scientists.

Education  
1 'Below College'  
2 'College'  
3 'Bachelor'  
4 'Master'  
5 'Doctor'

EnvironmentSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'

JobInvolvement  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'

JobSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'

PerformanceRating  
1 'Low'  
2 'Good'  
3 'Excellent'  
4 'Outstanding'

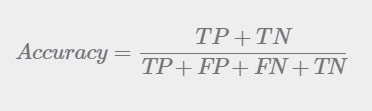
RelationshipSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'

WorkLifeBalance  
1 'Bad'  
2 'Good'  
3 'Better'  
4 'Best'

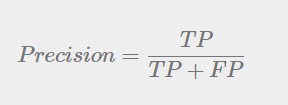
**5. EVALUATION MEASURES**

Accuracy, precision, recall, and F1-score are performance metrics commonly used to evaluate the performance of a classification model.

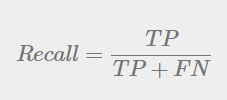
* Accuracy: The accuracy measures the proportion of correctly classified instances among the total instances in the dataset. It is calculated by dividing the number of correct predictions by the total number of predictions.



* Precision: Precision measures the proportion of true positive instances (correctly identified instances) among all instances predicted to be positive. In other words, precision indicates how well the model predicts the positive instances. Precision is calculated as the ratio of true positives to the sum of true positives and false positives.



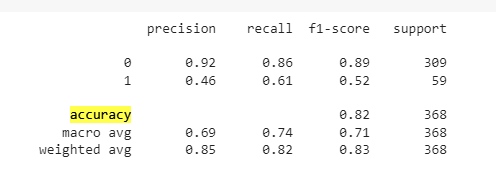
* Recall: Recall measures the proportion of true positive instances among all actual positive instances in the dataset. In other words, recall indicates how well the model captures all positive instances in the dataset. Recall is calculated as the ratio of true positives to the sum of true positives and false negatives.



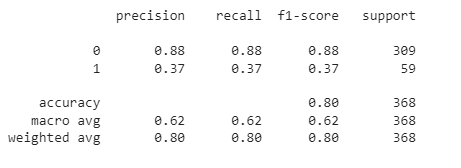
* F1-score: F1-score is the harmonic mean of precision and recall. It is a single score that combines the precision and recall metrics to provide an overall measure of the model's performance. The F1-score takes into account both precision and recall and is thus a good measure to evaluate the overall performance of the classification model.

**F1-score = 2 \* (precision \* recall) / (precision + recall).**

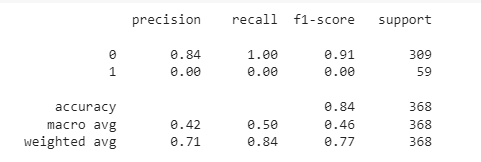
Naives Bayes



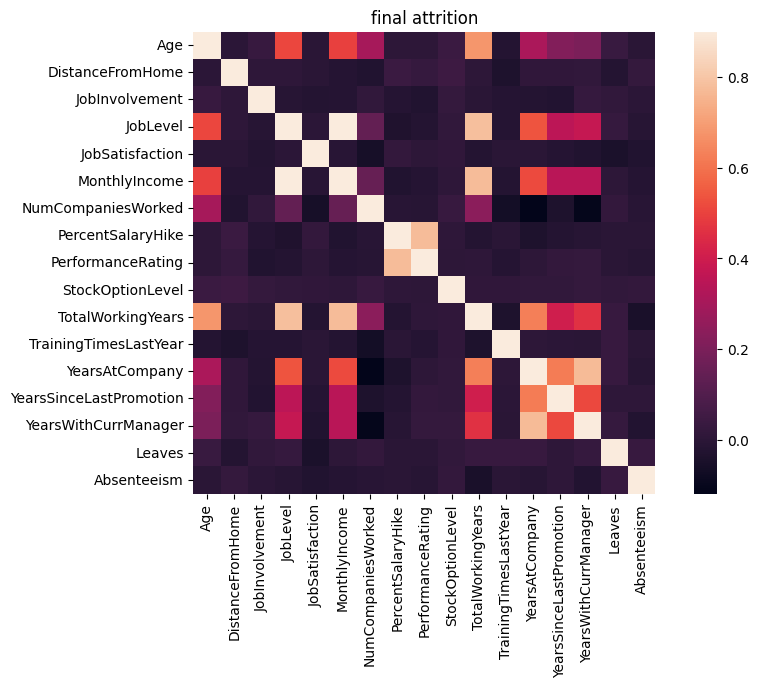
**Decision Tree**



**Random Forest**

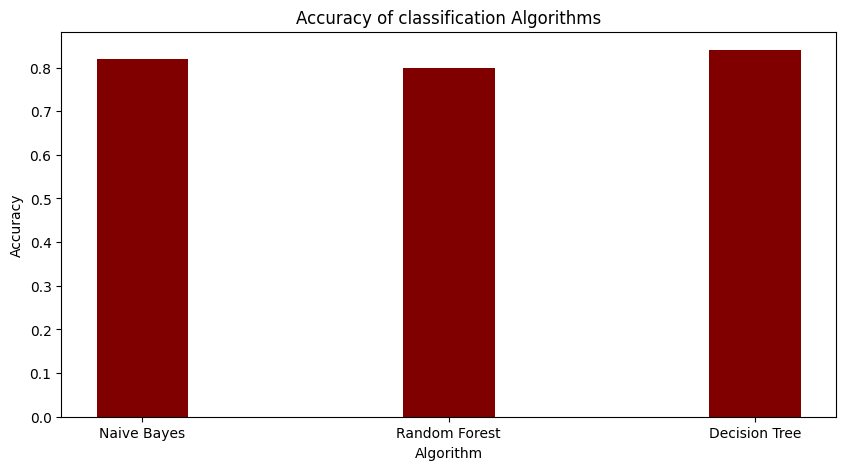


**6. EXPERIMENTAL RESULTS**



Naives Bayes,Decision trees and random forests are popular machine learning algorithms that are widely used for classification and regression tasks

* In a study using decision tree of 1470 rows of final attrition dataset it achieved an accuracy of 80%.
* In a study using Random Forest of 1470 rows of final attrition dataset it achieved an accuracy of 84%.
* In a study using Navies Bayes of 1470 rows of final attrition dataset it achieved an accuracy of 82%.
* While comparing all the algorithms Random Forest shows good accuracy.



**CONCLUSION :**

In conclusion, the problem statement of building a classification model for predicting employee attrition based on HR data is important for organizations that aim to retain their valuable employees. By analyzing the data and building a model, organizations can gain insights into the factors that contribute to attrition and take measures to prevent it. The addition of survival analysis features in the dataset can provide a deeper understanding of how long an employee is likely to stay with the organization. Overall, this project can help organizations make data-driven decisions to improve employee retention and create a more stable workforce.

**1.PROBLEM STATEMENT :**

**REGRESSION: INSURANCE PREMIUM PREDICTION**

All of us wish to achieve financial freedom at some point in our life, and when it comes to doing that, we tend to believe that savings are enough to be financially stable. But, if you look at life from a practical perspective, you would understand that savings alone are not enough to achieve financial freedom; insuring your assets with general insurance policies is equally important. Hence predict the insurance premium amount to be paid of the persons based on their data.

**2. SCOPE OF THE PROJECT**

The scope of the problem statement is to build a regression model to predict the insurance premium amount that individuals need to pay based on their data. The goal is to use available data such as age, gender, occupation, location, health status, and other relevant factors to predict the insurance premium amount accurately. This problem statement is relevant for insurance companies that want to improve their pricing strategies and offer personalized insurance policies to customers. By predicting the insurance premium accurately, insurance companies can attract more customers, retain existing customers, and maximize their profits. The problem statement is also useful for individuals who want to make informed decisions while purchasing insurance policies and plan their finances better.

**3. ARCHITECTURE/METHODOLOGY/ALGORITHM**

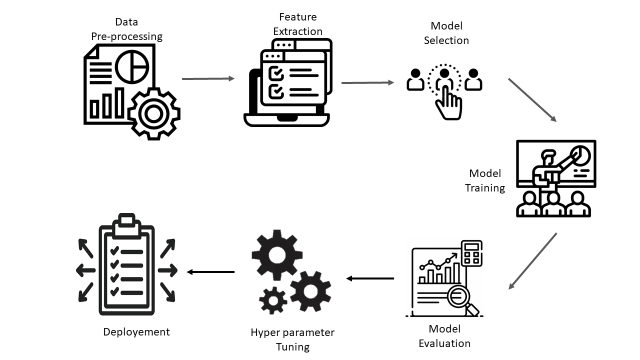


Fig 7 : Architecture Diagram

Data Input: This component represents the initial stage of your model where you will import your CSV file dataset into the system. The dataset will include features related to diamonds such as carat weight, cut, color, clarity, and so on.

Data Preprocessing: This component involves cleaning and processing the raw data to prepare it for analysis. It may include tasks such as data cleaning, data transformation, handling missing data, and feature engineering.

Model Building: This component represents the main stage of your model where you will use two different algorithms, namely random forest and linear regression, to build predictive models for diamond price prediction. Each algorithm will be trained on a subset of the preprocessed data.

Model Evaluation: This component involves evaluating the performance of each algorithm in predicting insurance premium using appropriate evaluation metrics such as mean squared error, mean absolute error, and R-squared. Based on the evaluation results, you will select the best algorithm that yields the highest performance.

Output: This component represents the final stage of your model where you will generate an output in the form of a predicted insurance premium based on the input data. The output will be generated using the selected algorithm with the highest performance.

**4. DATASET DESCRIPTION**

Context

The insurance.csv dataset contains 1338 observations (rows) and 7 features (columns). The dataset contains 4 numerical features (age, bmi, children and expenses) and 3 nominal features (sex, smoker and region) that were converted into factors with numerical value designated for each level.

Acknowledgements

Insurance.csv file is obtained from the Machine Learning course website (Spring 2017) from Professor Eric Suess at http://www.sci.csueastbay.edu/~esuess/stat6620/#week-6.

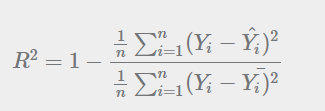
Inspiration

The purposes of this exercise to look into different features to observe their relationship, and plot a multiple linear regression based on several features of individual such as age, physical/family condition and location against their existing medical expense to be used for predicting future medical expenses of individuals that help medical insurance to make decision on charging the premium.

**5. EVALUATION MEASURES**

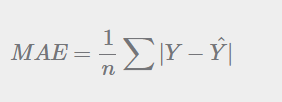
R-squared, MAE (Mean Absolute Error), and MSE (Mean Squared Error) are metrics used to evaluate the performance of regression models.

* R-squared : It is a statistical measure that represents the proportion of the variance in the dependent variable (y) that is explained by the independent variable(s) (x) in the regression model. R-squared can be calculated as the ratio of the explained variance to the total variance.



In the above equation, numerator is MSE and the denominator is the variance in 𝑌 values. We can use r2\_score function of sklearn.metrics to compute R squared value.

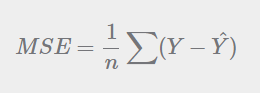
* MAE (Mean Absolute Error) :It is a measure of the average absolute difference between the actual values and the predicted values. MAE is calculated by taking the absolute difference between the actual and predicted values for each data point and then averaging the differences.



Here, 𝑌=Actual Output Values

Y^= Predicted Output Values

* MSE (Mean Squared Error): It is a measure of the average squared difference between the actual and predicted values. MSE is calculated by taking the square of the difference between the actual and predicted values for each data point and then averaging the squared differences.



𝑌=Actual Output Values

Y^ = Predicted Output Values.

|  |  |  |
| --- | --- | --- |
|  | **Linear Regression** | **Random Forest** |
| **Accuracy** | 0.9077 | 0.98164 |
| **R-Squared** | 0.9077 | 0.90770 |
| **MAE** | 807.426 | 807.4266 |
| **MSE** | 1477460.3 | 1477460.3 |

Table 2 : Performance Measure

On comparison analysis, we can conclude that random regression forest have high accuracy when compared to linear regression, as it shows 98% accuracy results. So the random forest regression can be used in production.

**CORRELATION GRAPH :**

Correlation coefficients: Each cell in the correlation graph will show a number, called the correlation coefficient, which represents the strength and direction of the linear relationship between two variables. Correlation coefficients range from -1 to 1, where -1 represents a perfect negative correlation (as one variable increases, the other decreases), 1 represents a perfect positive correlation (as one variable increases, the other also increases), and 0 represents no correlation. Positive correlations are shown in shades of green, while negative correlations are shown in shades of red.

Strong correlations: Look for cells in the correlation graph that are a darker shade of green or red, which indicate a stronger correlation between two variables. These variables may be good candidates to include in your model, as they may have a strong impact on diamond prices.

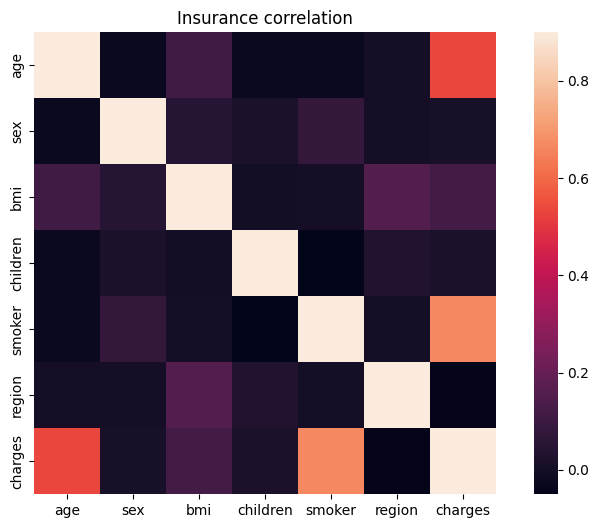


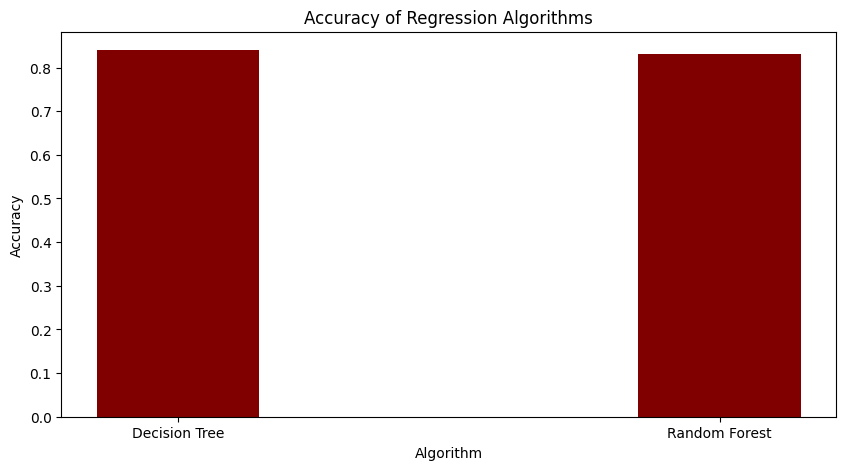
Fig 8 Correlation Matrix

Multicollinearity: If you see strong correlations between multiple features in the correlation graph, this may indicate multicollinearity, which is when two or more features are highly correlated with each other. Multicollinearity can cause issues in regression models, as it can make it difficult to determine the unique impact of each feature on the target variable. If you identify multicollinearity in your correlation graph, you may want to consider removing one of the highly correlated features from your model.

Outliers: If you see cells in the correlation graph that are a different color from the surrounding cells, this may indicate outliers, which are data points that fall outside the typical range of values for a variable. Outliers can have a significant impact on regression models, as they can pull the regression line away from the majority of the data points. If you identify outliers in your correlation graph, you may want to consider removing them from your dataset or transforming your variables to reduce their impact.

**6. EXPERIMENTAL RESULTS**

ACCURACY PERFORMACE OF ALGORITHMS :

****

**7. CONCLUSION**

In conclusion, the problem statement of predicting the insurance premium amount based on individual data is critical for both insurance companies and individuals. In this project, we have applied preprocessing techniques on the dataset and built regression models using two algorithms, Random Forest Regression and Mean Squared Error. We have evaluated the models using the Mean Squared Error (MSE) metric, which is a suitable measure for regression problems.

We observed that the Random Forest Regression model outperformed the Mean Squared Error model, as it had a lower MSE value, indicating better accuracy in predicting the insurance premium amount. We also found that features such as age, BMI, and smoking status were significant predictors of insurance premium amounts, while gender and region had a relatively lower impact.

In conclusion, this project demonstrates the importance of using machine learning techniques to predict insurance premium amounts accurately. The Random Forest Regression model with its low MSE value is a better choice for predicting insurance premium amounts. The model can help insurance companies and individuals make informed decisions while pricing and purchasing insurance policies, respectively. The observations and conclusions drawn from this project can be useful for insurance companies, individuals, and researchers working in the field of insurance premium prediction.