**RANDOM FOREST**

A supervised learning algorithm has been used to predict the employment level. Random forest is an ensemble of decision trees to increase accuracy. Multiple decision trees are a built and merged together to get more accurate and stable predication. This method has been used as it does not have the problem of overfitting attached to it. Hence, a very high level of accuracy can be met. This model has been built in R studio and the “Random Forest” package has been used.

**STEP 1:** In the given dataset, we have predicted the employment status for all the rows where the

Earnings were zero. This includes employed, unemployed and not in labour force.

**STEP 2:** We have further segmented the data according to the education level into 10 categories as the level of income earned is different for each category. As the years of education received increases, there is an increase in the weekly income earned. The categories being

1. 9th Grade
2. 10th Grade
3. 11th Grade
4. 12th Grade
5. High School
6. Some College
7. Associate Degree
8. Bachelors
9. Masters
10. Ph.D or any other Professional Degree

**STEP 3:** For each category, the model has been trained using 90% of the dataset (Train Data) & 10% of the dataset has been used for validation. The cutoff function has been used to create this division.

**STEP 4:** Running the random forest model, the predicted value has been saved in a new column “P1”

**STEP 5:**  The indicator variables are now checked to understand the significance of each. Here a variable importance plot has been used to understand what is the contribution of each variable in building the model. The algorithm shows how the model shall perform if that particular variable has been removed. For each category of education level, the variable importance plot has been built to understand important variables. According to the results, “weekly Hours Earned” and “Weekly Earnings” are highly significant with the accuracy level dropping by ~65% if each of these variables is removed.

**STEP 6:** The robustness of the model has been tested using K-Fold Validation to find the accuracy level if the dataset is randomly split into k subsets as many a times, the model might be overfitting or the variables are biased due to which a high accuracy is achieved but is not correct. Thus, the performance of the model is tested by dividing the dataset into random subsets. Here, we have divided the dataset into 5 subsets at each education level – 4 of these subsets are used to train the model and one is used to validate. Upon performing this K-Fold validation process, an accuracy of 94.42% has been achieved on average with upper limit as 95.57% and lower limit 92.78%.

**STEP 7:** The accuracy of the model has been checked by creating a confusion matrix. This is a table which defines the performance of a classification model by telling us how many rows were accurately predicted. Further, this matrix also gives us more information about how many rows were predicted incorrectly and what was the incorrect prediction which allows us to dive deeper and

maybe fine tune the model for better accuracy. In the model built for this project, an average accuracy level of 94.78% has been witnessed after running the model for each education level.

**Ensemble Model**

Ensemble methods include multiple supervised learning algorithms to obtain better predictive performance by including a concrete finite set of alternative methods and hence a very flexible structure exists amongst the available alternatives.

**STEP 1:**  The initial stage of the process was cleaning the data and preparing it to be used. Variables like age range, education level, and gender were converted to discrete variables to make them usable for our model. Further, we extracted the data where the weekly earnings were zero as that is the level of weekly earnings for which we shall predict the employment status. The dataset was also split in a 9:1 ratio for training and testing respectively.

**STEP 2:** The imbalance level of the data was checked by a simple plotting of the sum of each employment level. This indicated the existence of imbalance in our dataset.

**STEP 3:**  Due to the existence of imbalance, the ensemble was created using 3 learning algorithms, training each model using only one level of employment status. This segmentation was done using the “One Vs All” method for each level of employment.

**STEP 4:** For the first class Unemployed, **Support Vector Machine** was used to train the model which are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a binary linear classifier. Here, it returns the probability for each row of it being Unemployed or otherwise.

**STEP 5:** Next, for the class Employed, **XGBoost** which is an implementation of gradient boosted decision trees designed for speed and performance. Here again, we get the probability of each row being Employed or otherwise.

**STEP 6:** For the last class- Not in Labour Force, a **Random Forest** model was built to determine the probability of row being not in labour force or otherwise.

**STEP 7:** As the probabilities for each class were now available, we combined the three models to create the ensemble. Now, each row had a probability attached to it for belonging to any of the three classes and hence whichever class had the highest probability, the row was assigned to that class. As we had trained each model separately for each level, it helped in being accurate.

**STEP 8:** A confusion matrix was then created to understand the accuracy level of the ensemble. This is a table which defines the performance of a classification model by telling us how many rows were accurately predicted. In the ensemble created, the recall values showed an accuracy level of 91% for the training subset and 91% for the testing subset.

**STEP 9:** A grid search function, a feature extraction method to fine tune the parameters for better results.

**STEP 10:**  The robustness of the ensemble was checked using K-Fold validation to find the accuracy level if the dataset is randomly split into k subsets as many a times, the model might be overfitting or the variables are biased due to which a high accuracy is achieved but is not correct. Thus, the performance of the model is tested by dividing the dataset into random subsets. Here, we have divided the dataset into 5 subsets at each education level – 4 of these subsets are used to train the model and one is used to validate.

**ARTIFICIAL NEURAL NETWORK**

ANN is defined as a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. This machine learning technique is based on the belief that working of human brain by making the right connections, can be imitated using silicon and wires as living neurons and dendrites. Such systems "learn" (i.e. progressively improve performance on) tasks by considering examples, generally without task-specific programming.

This technique was used to train a model to predict the employment level.

**STEP 1:** To begin with, we prepared our dataset by including only those rows where the weekly earnings were zero. Further, we divided the dataset in the ratio of 80:20 as training and testing respectively.

**STEP 2:** For the purpose of K-Fold validation, the dataset has been divided into 5 random subsets.

**STEP 3:** A simple neural network model was built using deep learning estimator with 2 layers. We got 5 models and went ahead with using model 1 as the accuracy score was 93.89%.

**STEP 4:** The accuracy of the model has been checked by creating a confusion matrix. This is a table which defines the performance of a classification model by telling us how many rows were accurately predicted. Further, this matrix also gives us more information about how many rows were predicted incorrectly and what was the incorrect prediction which allows us to dive deeper and maybe fine tune the model for better accuracy. In the model built for this project, an average accuracy level of 99% has been witnessed.

**STEP 5:** The robustness of the ensemble was checked using K-Fold validation to find the accuracy level if the dataset is randomly split into k subsets as many a times, the model might be overfitting or the variables are biased due to which a high accuracy is achieved but is not correct. Thus, the performance of the model is tested by dividing the dataset into random subsets. Here, we have divided the dataset into 5 subsets at each education level – 4 of these subsets are used to train the model and one is used to validate. Upon performing this K-Fold validation process, an average accuracy of 93.5% has been achieved with upper limit as 93.8% and lower limit 93%.