# Chapter 02: Literature Review

# 2.1 Salary Discrimination

Even though over 25 years of time the gender discrimination for salary payment is narrowing down little by little, according to research still their gender gap in the IT industry regarding to the enumeration that an individual receive (Quan, 2020). It is not only the IT industry that is being affected by the pay gap, but many workplaces that has gender discrimination have led to this kind of pay gap and also being the prominent topic worldwide, the remuneration is not being differed by the employee working hours or their qualifications but also by gender (Khatri, 2022). Salary is one of the most important for an individual due to efficient wages being paid is a motivation to optimal production and informative of employee ability, as a result the turnover in the labor market will be reduce if this wage discrimination is turned out.(Meli and Spindler, 2021).

Moreover, salary discrimination is not based on the skills or the education level an individual have, the main other factor that leads to discrimination in remuneration payment is the Region or based on the different location, according to the location that an individual is in the remuneration differs from another location, since the standard of living will be totally different. (Zhou, Bu and Gao, 2021).

According to the study, in between 2018 to 2019 women I the tech career is being offered less than men for the same job, and sometimes in the same company which has led up to 63%. As a result, even though the Tech industry believe that the gender pay gap is narrowing, it false since it is growing throughout.(Perry, 2020). Apart from the gender, location above it is also said that the salary history causes individuals to have discriminated pay (Safstrom, 2019), hence each employees of the same job career of same level experience will earning different remuneration.

Inequality of paying remuneration in this current society is not only based on the gender, it is also decided based on the peoples color, LGTBQ communities and other ethnic minorities (Cziesielski, 2020), hence the author has decided to proposed this thesis in order to overcome some kind of salary discrimination in the IT industry based on Sri Lanka salary information.

## 2.2 Approach taken to solve Salary discrimination.

### 2.2.1 Banning of salary history.

One of the ways in which to make salary pay discrimination is using salary history in setting up the wages. Hence during the process of hiring an employee to the organization the employer should avoid inquiring regarding the individual salary history and adopting “reckless discrimination” legal theory could promote greater pay equity in the industry. (Vandenberg, 2020)

## 2.3 Factors considered whilst considering salary payment.

### 2.3.1 Education

Education is an investment according to the human capital therapist. It has been demonstrated that education increases a person's capacity for production, employability, and discretionary income, which in turn lowers the cost of community social services and creates funds to fund additional education. Additionally, it has been discovered that education raises a person's quality of life by giving them more free time and resources to enjoy it (Owings and Kaplan, 2019). Due to higher education with a better knowledge of the domain that an individual will be earning more than another individual since reducing the risk of loss will be minimal according to the education level an individual holds to (Quan, 2020).

Moreover it clearly indicates that when an individual has a better education qualification, the individual has a higher chance of obtaining a higher salary.(Wang et al., 2022)

### 2.3.2 Experience

Experience is another human capital factor. By having a higher experience it during on-job training will enable an individual to acquire a versatile skills which will be an added advantage since the knowledge in the domain will be very much higher than the new fresh graduate or an individual who is in his intern period (Quan, 2020).

### 2.3.3 Type of work selection

The remuneration that will be paid to an individual will also be considered based on the type of working selection selected. The possible working selection could be either temporary or permanent contract (Marin-Garcia and Martínez-Tomás, 2022), as a result the salary payment for an individual will also be different apart from the gender discrimination.

### 2.3.4 Geographical Location

The remuneration that will be given to an individual will also be affected by the geographical location they are working on, since the economy will be different compared to other countries across which could be either developing, developed or poverty state, as a result the IT industry individual will be given different remuneration even when they working on the same job role across different region (Zhang and Chung, 2018).

### 2.3.5 Environment of the Industry

As per the above factors the remuneration that could be earned by an individual is based on how well the company has been established in the industry (Wang et al., 2022). The higher it has been established the company tends to make higher profit due more projects being involved compared to start up, as a result the salary will be differed.

## 2.4 Recommendation System

Due to increase in information over the internet it is very much difficult for users to obtain data that is most useful for them, therefore recommendation system has been increasingly popular model used to give the best suggestion according to the user preferences (Dhinakaran et al., 2022). The development of smart devices has caused a huge traffic in web, App and SNS platforms, which collects various information related to the users on whatever the user uploads or share among people across the internet, hence the recommendation system should well use these data to make the better preference as well as perform in a better way for the users across the globe (Ko et al., 2022).

## 2.5 Problems in Recommendation system

Even though the recommendation system has it own advantages, it to have different problems when building a recommender system.

### 2.5.1 Cold Start Problem

Cold start problem arises when there is no information found about the user or item in the system, since the user has just started to use the application. An example can be when a user joins Netflix. The second problem that is possible to arise is when there is a new item that has been added to the system the collaborative filter technique uses user-item matrix to give recommendation, since the new item is not given any rating it is unable to give proper recommendation for this item which is known as cold start problem.(Mazumdar, Patra and Babu, 2020)

### 2.5.2 Sparsity problem

Sparsity problem arises when a user gives false information about the rating feedback that user gives to the product that has been recommended to them, since rating seems useless for the users (Mishra et al., 2021). Without realizing the user might give a higher rating considering its best for them or lower rating considered non likely without even noticing the type of the product. These ratings are taken as an input to the recommender system which in turn might display unwanted results to the end user, indicate loss of interest for the user in the platform and lead to non-efficient working (Kitazawa and Yui, 2018). Due to sparsity of rating matrix this kind of problem will arise.

### 2.5.3 Scalability problem

When a system is published there is a possibility the number of users could be growing up day by day, therefore will the system be able to cope up with the same performance level, since when the user of the system increases it will face scalability issue it become slow and make it feel it start to give problem in the recommender system, could e due to hardware and software scalability (Wang and Ke, 2014).

Moreover, the accuracy of the prediction could also be redundant or reduce due to increase in data, some of the algorithm are not being able to take advantage of increased efficiency of the hardware, which creates a problem of scalability (Da Costa et al., 2018).

### 2.5.4 Lack of Data

The efficiency of recommendations algorithm is evaluated by using publicly available dataset from different environment has been a common practice to make a recommender system, but recommender system faces a big issue to make an effective recommendation due to lack of data (Zhang et al., 2010). To make a better recommendation in any environment if there are more consumer/item data it would help to build a better recommender algorithm(Lika, Kolomvatsos and Hadjiefthymiades, 2014).

In addition to all the problems that have been specified above, there are also few problems such as Changing data, changing user preferences, unpredictable items which are also considered as a problem but based on the proposed system the salary recommendation system, this problem will now be suitable.

### 2.5.5 Over Specialization problem

This recommender problem arises when the recommended items are too like each other. When the user keeps on seeing the same item over and over again and not any new unique data, it might make the user loose interest on using that website since, the same result is showing over the time (Adamopoulos and Tuzhilin, 2014).

## 2.6 How will the recommender system help the Users?

### 2.6.1 Personalization

By giving users personalized recommendations based on their tastes and behavior, recommender systems can become more personalized. Collaborative filtering, content-based filtering, or hybrid strategies can all be used to accomplish this. Personalization is a crucial component that raises customer pleasure, loyalty, and engagement (Gorgoglione, Panniello and Tuzhilin, 2019).

### 2.6.2 Time saving

Reducing the time and effort needed to obtain pertinent information or items is one advantage that recommender systems can offer users. Users can quickly and easily identify new products or material that they are likely to appreciate or find useful by automatically producing individualized suggestions, without having to spend a lot of time searching or scrolling through extensive catalogs or websites.

In today's fast-paced, information-rich world, when users are frequently overwhelmed by the sheer volume of available information and content, this time-saving benefit is especially crucial. User-friendly recommender systems can assist users in quickly navigating this information environment and locating what they need or want (Wairkar et al., 2021).

### 2.6.3 Improved decision making

The advantage of offering customers individualized recommendations based on their interests and behavior is referred to as improved decision making in recommender systems. This aids users in avoiding decision fatigue, overcoming information overload, and making better decisions(Forouzandeh, Rostami and Berahmand, 2022).

## 2.7 Similar Works

### 2.7.1 Classification Recommendation system

Based on the research it has been clearly stated that recommendation system can be improved using other machine learning model such as Decision tree, Random Forest to make recommendation than not only depending on content-based filtering and collaborative-filtering (Bhansali and Nagwani, 2021).

**Multi-Domain Recommendation System Using Hybrid Filtering and Support Vector Machine Classification** (Vasanth, PeriyaKaruppan and PoornaKumar, 2020)has proposed a hybrid recommendation system which combines collaborative, content-based as well as support vector in order to improve the recommendation accuracy. The main domain that has been selected by the author is the movie domain. The process of the system first uses the collaborative and content-based filtering to generate recommendation based on the user-item and item features, based on past interactions and characteristics of the user and the recommended products, the system employs a support vector machine classifier to predict whether the user will like or dislike the recommended items. The experimental results indicate that by using a hybrid approach the overall recommendation quality will be improved and overcome the problems the recommender systems.

**Classification Algorithm for Career Recommendation System**(Masika, Rono and Kati, 2022)**.** The study suggests a career recommendation system that matches job searchers with appropriate employment roles using categorization algorithms. The method considers several variables, including the education, experience, talents, and interests of the job seeker as well as the credentials, job type, and geographical criteria.

In terms of recommendation accuracy, the authors evaluate the performance of three different categorization algorithms: decision trees, random forests, and support vector machines (SVM). They make use of a dataset that contains the profiles of job searchers and job specifications, and they assess the algorithms' performance using a variety of performance metrics, including accuracy, precision, recall, and F1-score.

The experimental findings demonstrate that the SVM algorithm, with an accuracy rate of 93.75%, surpasses the decision tree and random forest algorithms in terms of recommendation accuracy. To assess the system's usability and get user input on the suggested changes, the authors also carry out a user study.

Overall, the article offers a potential method for creating classification algorithms-based career recommendation systems. The imbalance between supply and demand in the labor market can be closed by the suggested system by assisting both employers and job seekers in locating appropriate workers.

### 2.7.2 Content based Recommendation system.

**Solving Cold start Problem for recommendation system using Content Based filtering**(Chia and Najafabadi, 2022)**.** To address the cold start issue in recommendation systems, the study suggests a content-based filtering technique. The lack of user-item interaction data, known as the "cold start" problem, makes it challenging to produce precise recommendations utilizing collaborative filtering approaches. The cold start issue can be solved by content-based filtering, which generates suggestions based on the qualities of things.

The authors employ a vector space model to represent the elements in a dataset of movie properties, including title, genre, and storyline synopsis. The similarity between the user's preferences and the item attributes is then calculated using cosine similarity, and the user is then given customized recommendations.

The experimental results demonstrate that the proposed technique performs better in terms of recommendation accuracy and diversity than a few baseline methods. To assess the system's usability and get user input on the suggested changes, the authors also carry out a user study.

Overall, the research offers a viable method for utilizing content-based filtering to address the cold start issue in recommendation systems. By raising the quality of recommendations for new users and items, the suggested method can improve user engagement and experience.

### 2.7.3 Salary prediction system

**Salary prediction using random forest with fundamental features**(Chen, Mao and Yuan, 2022).The purpose of the study article is to use random forests with basic features to estimate individual incomes. Included among the essential characteristics are education, years of experience, job title, and firm size.

The study makes use of a dataset including details on worker incomes, job descriptions, business sizes, educational backgrounds, and years of experience. The data is preprocessed, feature engineering is done, and a random forest model is used to forecast salaries.

The findings demonstrate that the random forest model with core features is highly accurate in predicting salaries. To ascertain which features have the most bearing on wage projections, the authors also conduct a feature important analysis.

The study offers insightful information on how to forecast salaries utilizing fundamental features using machine learning techniques, specifically random forest. Additionally, it emphasizes how crucial feature engineering and feature importance analysis are for enhancing model performance.

## 2.8 Technology Review

Recommendation systems can be built in different ways by using recommendation techniques or by using machine learning algorithms or with the help of deep learning. Even though this method has been proven to perform well it has its own limitations as well, hence according to the business requirement the recommendation techniques and algorithms should be chosen. Building a proper recommendation system with an effective algorithm based on the requirement would help to improve the sales of a business(Jannach and Jugovac, 2019).

### 2.8.1 Recommendation system techniques

|  |  |
| --- | --- |
| Techniques | Description |
| Content based filtering | Content based filtering algorithm is an Information retrieval Technology (Wu, 2022), which recommends or extract information based on the passed or newly added items that matches the users operation behavior.  Steps to do content-based filtering:   1. Extract the information from the people of events that have taken place. 2. Should transform the feature sets into feature vectors. 3. Use the vector space and then make comparison with Cosine Similarity. |
| Collaborative filtering | Collaborative filtering can be divided into 3 different categories.   1. User based collaborative filtering– This algorithm uses user information which will get the user with many similarities from the organization to that makes mutual recommendation between each other. 2. Item-based filtering – This type of algorithm could be used when the number of items is less compared to the number of users in the system, this would recommend items like the target product that the user has selected. 3. Model based collaborative filtering – This approach is based on the user’s behavior and their preference information.   Please check diagram for more clarification about User based and Item based filtering. (PUT IN THE APPENDIX)    Figure User and Item based filtering (Wu, 2022) |
| Knowledge based filtering | Knowledge based recommendation will be using the knowledge about the users as well as the product that the user is interest in, as a result the product that is recommended will be based on the user’s requirement (Burke, 2000). |
| Hybrid recommender system | This type of recommendation system is a combination of two recommendation model, or it could be different kind of machine learning algorithm that is combined with one of the recommendation techniques. |

### 2.8.2 Machine Learning algorithm for recommender systems

|  |  |
| --- | --- |
| Models | Description |
| Random Forest Classifier | The Random Forest algorithm is a supervised learning system that uses decision trees as its foundation. To arrive at a final forecast, it builds several decision trees throughout training. The algorithm grows each decision tree using a random subset of the available features and data samples. This randomness enhances the algorithm's capacity for generalization and guards against overfitting (Shingade, Mudhalwadkar and Masal, 2022). |
| Decision Tree | The Decision Tree algorithm is a supervised learning method that is a member of the classification and regression tree family. Based on the information at hand, it constructs a tree-like model of decisions and their outcomes. To maximize the homogeneity of the generated subsets, the method iteratively divides the dataset into smaller subsets based on the most important attribute. A tree-like structure is produced because of this process, which can be used to anticipate the outcomes of new data (Bansal, Goyal and Choudhary, 2022).  PUT THE BELOW IMAGES TO APPEDIX |
| K nearest neighbor algorithm | The KNN algorithm is a supervised learning method that generates predictions based on how similar new data instances are to training instances. The approach employs the distance between the points to calculate the similarity between all the instances that are stored as points in a multi-dimensional space. When a new instance is introduced, the algorithm locates the k-nearest neighbors in the training set and places the new instance in the class that is most prevalent among its k-nearest neighbors. When a new instance is introduced, the algorithm locates the k-nearest neighbors in the training set and places the new instance in the class that is most prevalent among its k-nearest neighbors. Several fields, including recommender systems, have successfully used the KNN algorithm (Sagdic et al., 2020). |
| Support Vector Machines | The SVM is a supervised learning method that seeks to locate a hyperplane that maximizes the margin between the classes. Using kernel functions, the SVM can handle both linear and nonlinear classification issues. The SVM algorithm is especially well suited for datasets with many characteristics and few data examples. If the regularization value is not properly adjusted, the SVM algorithm may be prone to overfitting (Gil, Freudenthaler and Natschläger, 2018). |

## 2.9 Evaluation Techniques

|  |  |
| --- | --- |
| Metrics | Description |
| Accuracy | A measure of accuracy is the percentage of correctly classified data instances out of all instances. For classification algorithms like the Random Forest Classifier, it is frequently employed as an evaluation metric. High accuracy means that the classifier can accurately categorize a significant fraction of the data instances, whereas low accuracy means that many examples are misclassified. |
| Precision | A measurement of precision is the percentage of suggested things that are relevant to the user's interests. For content-based filtering algorithms, it is frequently employed as a measurement parameter. While low precision suggests that the recommendations are not relevant, high precision suggests that the user will find the suggested items to be highly relevant. Precision is determined by dividing the number of true positives—that is, relevant items that are correctly recommended—by the total number of items recommended. |
| Recall | Recall is a measurement of the percentage of pertinent items that are suggested to the user. It is frequently employed as a statistic for content-based filtering algorithms. Low recall means that many relevant items are missed, but high recall shows that the algorithm can recommend a significant amount of the user's relevant items. Recall is calculated as the proportion of true positives to all relevant items. |
| F1 | A recommendation system is evaluated fairly using the F1-score, a metric that combines precision and recall. Precision and recall are determined as harmonic means, which gives each statistic equal weight. While a low F1-score suggests that the algorithm is deficient in either precision or recall, a high F1-score indicates that the algorithm possesses both. |
| ROC AUC Score | A statistic used to assess the effectiveness of binary classification algorithms is the ROC AUC score. By displaying the True Positive Rate (TPR) versus the False Positive Rate (FPR) at various classification thresholds, it assesses the model's capacity to distinguish between positive and negative classes.  The classification threshold, which determines whether a predicted probability belongs to the positive or negative class, is changed to produce the ROC curve. The proportion of positive events that are correctly categorized as positive is known as the TPR, whereas the proportion of negative events that are wrongly labeled as positive is known as the FPR. Higher numbers denote greater performance. The AUC score is the area under the ROC curve and ranges from 0 to 1. |

## 2.10 Benchmarking

The benchmarking of a system could be done by comparing the performance of the model with other different models in accordance with the requirement of the system. In fact, the dataset used during the process of training to do a better recommender system wither by selection of one algorithm or based on the hybrid approach could be also a part of benchmarking, Since the dataset will be specific for requirement.

## 2.11 Chapter Summary

In this chapter the author has spoken about the problems in the research domain, such as in the Financial as well as the recommendation system. In addition, the author has also explained about the similar products that are already been researched in classification, recommendation techniques as well as prediction. The chapter also clearly explains the possible types of recommendation techniques and the machine learning algorithm which could be used in the recommender system along with the how and how well the build models could be evaluated.